



A glimpse at an early stage of microbe domestication revealed in the variable genome of *Torulaspota delbrueckii*, an emergent industrial yeast

Margarida Silva^{1,2} | Ana Pontes^{1,2} | Ricardo Franco-Duarte^{3,4} | Pedro Soares^{3,4} | José Paulo Sampaio^{1,2}  | Maria João Sousa^{3,4} | Patrícia H. Brito^{1,2} 

¹Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, UCIBIO, Universidade Nova de Lisboa, Caparica, Portugal

²Faculdade de Ciências e Tecnologia, Associate Laboratory i4HB - Institute for Health and Bioeconomy, Universidade Nova de Lisboa, Caparica, Portugal

³Department of Biology, CBMA (Centre of Molecular and Environmental Biology), University of Minho, Braga, Portugal

⁴Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, Braga, Portugal

Correspondence

José Paulo Sampaio and Patrícia H. Brito, UCIBIO, Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal.

Emails: jss@fct.unl.pt (J. P. S.); phbrito@fct.unl.pt (P. H. B.)

Funding information

This research was funded by grants PTDC/BIA-MIC/30785/2017, SFRH/BD/136462/2018, UIDP/04378/2020 and UIDB/04378/2020 (UCIBIO), and LA/P/0140/2020 (i4HB). This work was supported by the strategic programme UID/BIA/04050/2019, and by the project PTDC/BIA-MIC/32059/2017 funded by national funds through FCT, I.P. and by the ERDF through the COMPETE2020 – Programa Operacional Competitividade e Internacionalização (POCI) and Sistema de Apoio à Investigação Científica e Tecnológica (SAICT). This work was carried out with support of INCD funded by FCT and FEDER under the project 22153-01/SAICT/2016.

Handling Editor: Tatiana Giraud

Abstract

Microbe domestication has a major applied relevance but is still poorly understood from an evolutionary perspective. The yeast *Torulaspota delbrueckii* is gaining importance for biotechnology but little is known about its population structure, variation in gene content or possible domestication routes. Here, we show that *T. delbrueckii* is composed of five major clades. Among the three European clades, a lineage associated with the wild arboreal niche is sister to the two other lineages that are linked to anthropic environments, one to wine fermentations and the other to diverse sources including dairy products and bread dough (Mix-Anthropic clade). Using 64 genomes we assembled the pangenome and the variable genome of *T. delbrueckii*. A comparison with *Saccharomyces cerevisiae* indicated that the weight of the variable genome in the pangenome of *T. delbrueckii* is considerably smaller. An association of gene content and ecology supported the hypothesis that the Mix-Anthropic clade has the most specialized genome and indicated that some of the exclusive genes were implicated in galactose and maltose utilization. More detailed analyses traced the acquisition of a cluster of GAL genes in strains associated with dairy products and the expansion and functional diversification of MAL genes in strains isolated from bread dough. In contrast to *S. cerevisiae*, domestication in *T. delbrueckii* is not primarily driven by alcoholic fermentation but rather by adaptation to dairy and bread-production niches. This study expands our views on the processes of microbe domestication and on the trajectories leading to adaptation to anthropic niches.

KEYWORDS

galactose metabolization gene, maltose metabolization gene, microbe domestication, microbial genomics, pangenome and variable genome, *Torulaspota delbrueckii*

1 | INTRODUCTION

In recent decades the study of microbe domestication has uncovered major genomic and phenotypic transformations linked to the adaptation to anthropic niches and to the emergence of traits relevant for the properties of numerous fermented foods and beverages. Evidence for the domestication of fungi and bacteria has been obtained from the study of a wide diversity of microbial transformations. Moreover, emblematic mechanisms of microbe evolution and adaptation, which resulted in substantial human benefit, have been uncovered in the study of *Aspergillus oryzae* (Gibbons et al., 2012) used for sake, miso and soy sauce, *Penicillium* spp. (Cheeseman et al., 2014), employed as starters in soft and blue-veined cheeses, lactic acid bacteria (Makarova et al., 2006) responsible for fermentations of vegetables and milk, and of yeasts of the genus *Saccharomyces*, especially *S. cerevisiae* (Pontes et al., 2020), used for the production of alcoholic beverages and different products based on leavened dough.

Although *S. cerevisiae* is strongly associated with yeast domestication and to wine and bread production, other yeasts also occur spontaneously in human-driven fermentations. Those non-*Saccharomyces* yeasts are also candidates for having undergone processes of artificial selection, but given that their biology is less well understood, their roles in fermentations are more obscure. In such cases, when a clear link between the desired properties of the product and the emergence of novel traits is absent or not evident in the sense of the definition of domestication provided by Diamond (2002), using the concept of “quasi-domesticate” (Pontes et al., 2019) might be justified, even if temporarily, until a better understanding of the contribution to the final product is established.

The yeast *Torulasporea delbrueckii* has long been associated with winemaking (Castelli, 1954) and is frequently found in spontaneous wine fermentations (van Breda et al., 2013). This species has several characteristics that are relevant for winemaking, especially if used in combination with *S. cerevisiae*. The most relevant oenological properties of *T. delbrueckii* are the ability to increase the sensorial complexity of wine while simultaneously contributing to ethanol production (Azzolini et al., 2012), but not to the accumulation of undesirable compounds (Renault et al., 2009). Another area of relevance for *T. delbrueckii* is the bread industry. This yeast appears to also have an ecological preference for colonizing artisanal bread dough (Almeida & Pais, 1996a). It contributes to dough leavening and, because it tolerates freezing and osmotic stresses, it is suitable for commercial applications in frozen dough formulations (Almeida & Pais, 1996; Alves-Araújo et al., 2004; Hernandez-Lopez, Prieto, & Randez-Gil, 2003). Besides these industrial applications, which have not yet been fully explored, *T. delbrueckii* is found in other anthropic environments worldwide related to the artisanal fermentation of foods and beverages. Indeed, this species has been infrequently found in a diverse array of fermented products ranging from cocoa (Papalexandratou et al., 2011), to olives (Kotzekidou, 1997), tequila (Lachance, 1995) and cheese (Welthagen, 1998).

The yeast *Torulasporea delbrueckii* has long been associated with winemaking (Castelli, 1954) and is frequently found in spontaneous wine fermentations (van Breda et al., 2013). This species has several characteristics that are relevant for winemaking, especially if used in combination with *S. cerevisiae*. The most relevant oenological properties of *T. delbrueckii* are the ability to increase the sensorial complexity of wine while simultaneously contributing to ethanol production (Azzolini et al., 2012), but not to the accumulation of undesirable compounds (Renault et al., 2009). Another area of relevance for *T. delbrueckii* is the bread industry. This yeast appears to also have an ecological preference for colonizing artisanal bread dough (Almeida & Pais, 1996a). It contributes to dough leavening and, because it tolerates freezing and osmotic stresses, it is suitable for commercial applications in frozen dough formulations (Almeida & Pais, 1996; Alves-Araújo et al., 2004; Hernandez-Lopez, Prieto, & Randez-Gil, 2003). Besides these industrial applications, which have not yet been fully explored, *T. delbrueckii* is found in other anthropic environments worldwide related to the artisanal fermentation of foods and beverages. Indeed, this species has been infrequently found in a diverse array of fermented products ranging from cocoa (Papalexandratou et al., 2011), to olives (Kotzekidou, 1997), tequila (Lachance, 1995) and cheese (Welthagen, 1998).

Similar to what is known in *S. cerevisiae*, wild isolates of *T. delbrueckii* have been found in arboreal niches in different regions (Carvalho et al., 2018; Limtong & Koowadjanakul, 2012). Yet, unlike in *S. cerevisiae*, an exhaustive survey of the the natural diversity of *T. delbrueckii* and its partition along different geographical locations, niches and technological specializations is still missing. Furthermore, given that both *S. cerevisiae* and *T. delbrueckii* can thrive in natural and artificial environments, this further justifies the need for a detailed population study and the investigation of domestication events in *T. delbrueckii*. Recently, a first approach to access the diversity of this species was carried out using microsatellite markers, and a separation between strains associated with bioprocesses carried out by humans (wine, bread and other fermentations) and strains originating from natural environments was proposed (Albertin et al., 2014). Here we expanded on this study and used genomics to analyse a comprehensive data set of *T. delbrueckii* representatives. By combining various approaches, including population structure analyses, pangenome functional annotation, and gene evolution assessments, we uncovered signs of an early stage of domestication in *T. delbrueckii* with different adaptation trajectories coexisting in the same clade and little differentiation between wild and domesticated lineages.

2 | METHODS

2.1 | Genome sequencing, *de novo* assembly and annotation

Raw sequence reads obtained with paired-end Illumina MiSeq (500 cycles) or NextSeq (300 cycles) were preprocessed with TRIMMOMATIC version 0.39 (Bolger et al., 2014) under different stringency levels to

remove adapter sequences and low-quality bases. BBMERGE version 38.73 (Bushnell et al., 2017) was also used on raw sequences to cut adapter sequences and merge overlapping paired reads into single read sequences. Whole genome de novo assembly was performed on all sets of preprocessed reads using either SPADES version 3.13.1 (Bankevich et al., 2012) or ABYSS version 2.0.2 (Simpson et al., 2009). SPADES was run in “careful” mode with k-mer sizes automatically selected based on read length. ABYSS was run with default parameters for every eighth value of k-mer size ranging from 50 to 90. We used the abyss-fac tool to select the optimal ABYSS assembly by inspecting contiguity statistics. The quality of all resulting genome assemblies for each strain was evaluated using QUAST version 5.0 (Gurevich et al., 2013), and the assembly with the highest genome size and N50 value was considered best. Small contigs (<1 kb) were discarded from final assemblies. For all genomes, ab initio prediction of protein coding genes and annotation was performed with Yeast Genome Annotation Pipeline (YGAP) (Proux-Wéra et al., 2012).

2.2 | Orthology prediction and phylogenetic analyses

Orthology prediction for phylogenetic analyses was performed using all-against-all BLASTP (NCBI Blast-2.2) searches and a Markov cluster algorithm, OrthoMCL version 1.4 (L. Li et al., 2003) with an inflation factor (F) of 1.5, as implemented in the GET_HOMOLOGUES package (Contreras-Moreira & Vinuesa, 2013). Orthologous copies were identified whenever a significant BLASTP hit (E -value cutoff of $1e-5$) was obtained with minimum pairwise sequence alignment coverage of 50% (parameter C50). All gene clusters (GCs) that contained one single orthologous copy per genome were retrieved to build a data set of single-copy core orthologous groups. Cluster sequences were aligned with MAFFT version 7.407 (Katoh & Standley, 2013) using the G-INS-I method and default parameter values, trimmed with BMGE version 1.12 (Crisuolo & Gribaldo, 2010) using the amino acid option, and finally concatenated into a single data set.

Phylogenetic analyses were performed with IQ-TREE version 1.6.12 (Nguyen et al., 2015) using maximum-likelihood inference. MODELFINDER (Kalyaanamoorthy et al., 2017) was used to determine the best model and branch support was estimated using fast bootstrap approximation with NNI optimization (Hoang et al., 2018), both implemented in IQ-TREE. Average nucleotide identity (ANI) was estimated from whole genome assemblies using ORTHOANI (Lee et al., 2016) and USEARCH as applied in Orthologous Average Nucleotide Identity Tool (OAT) (Yoon et al., 2017).

2.3 | Genome diversity, SNP calling, ploidy and segmental duplication

Chromosomal single-nucleotide polymorphisms (SNPs) were extracted following an adapted GATK germline short variant discovery pipeline (Poplin et al., 2017). Sequence reads of each isolate were

mapped to the *Torulaspota delbrueckii* reference genome (CBS 1146, ASM24337v1) using Burrows-Wheeler Aligner (BWA) version 0.7.17 (Li & Durbin, 2009) and duplicated reads were marked with PICARD version 2.22.8 (<http://broadinstitute.github.io/picard/>). SNP and INDEL discovery and genotyping were performed on all samples simultaneously using local re-assembly of haplotypes (GATK HaplotypeCaller, GenomicsDBImport, and GenotypeGVCFs) and standard hard filtering parameters according to GATK best practices recommendations (GATK VariantFiltration with parameter values $QD < 2.0$, $QUAL < 30.0$, $OR > 3.0$, $FS > 60.0$, $MQ < 40.0$, $MQRankSum < -12.5$, $ReadPosRankSum < -8.0$ (Depristo et al., 2011)). Ploidy levels and events of segmental amplification were determined for all genomes with original sequencing reads available. This analysis was performed by combing a systematic survey of coverage depth along 1-kb nonoverlapping windows with the distribution of SNP allele frequencies along the genome. Sequencing coverage was estimated from alignment files that fed the GATK SNP discovery pipeline, and read depth at each position was estimated with samtools depth. The median coverage estimated for each 1-kb nonoverlapping windows was normalized by the genome coverage, estimated as the median of the median coverage for each chromosome. The genomic location of each segmental duplication was compared with the inferred limits of the subtelomeric regions estimated for the reference genome using the definition of subtelomeres as gene-depleted regions (Brown et al., 2010; Winzeler et al., 2003). Plots of genome-wide allele frequencies were constructed from allele balance ratios extracted from all biallelic heterozygous SNPs, as these are expected to fit to specific ratios depending on the chromosomal ploidy level.

2.4 | Population structure

Population structure and individual admixture proportions were estimated from genotype likelihoods of biallelic SNP variants using NGSADMIX (Skotte et al., 2013), which uses a clustering method that takes into account the uncertainty in the genotype calls inherent to next generation sequencing technologies. We varied the number of ancestral populations (K) from 2 to 6 and fit admixture models to estimate individual ancestral proportions for each structure model using default parameter values for stop criteria and SNP filtering. For all runs convergence was achieved when the log likelihood difference for 50 interactions was < 0.1 . Plots with estimated individual ancestries were constructed with R version 3.6 (R Core Team, 2019). Principal component analysis was performed with PCANGSD (Meisner & Albrechtsen, 2018) using default parameters on the previous data set of genotype likelihoods of biallelic SNP variants.

2.5 | Pangenome analyses, pangenome tree and functional annotation

Orthology prediction for pangenome analyses was performed on all inferred proteomes of *T. delbrueckii* following a pipeline similar

to the one used for phylogenetic analyses by applying ORTHOMCL to cluster nodes in a pairwise BLAST graph. Here, gene clusters within paralogues were not excluded, and orthologous copies were identified whenever a significant BLASTP hit (E -value cutoff of $1e-5$) was obtained with query sequence coverage $>60\%$ and identity $>80\%$ (parameters C60, S80). These parameters were chosen as the lower expectations for the intraspecific polymorphism present in the data set. GET_HOMOLOGUES classifies each orthologous cluster into frequency classes (core, softcore, shell and cloud) depending on their presence in a data set. The core genome contains all GCs present in 100% of the genomes, while the cloud genome includes GCs present in $<5\%$ of genomes (Contreras-Moreira & Vinuesa, 2013). This latter class represents all rare GCs in the species pangenome, and hence is more prone to be biased by sequencing and annotation artefacts. We therefore performed a series of verification steps designed to identify false positives, and to determine the taxonomic origin of the true cloud genome. This analysis was performed in three steps. First, we performed a BLASTP search (E -value cutoff of $1e-6$) of each putative cloud gene against a database of all *T. delbrueckii* proteomes, to identify alignment artefacts that were excluded from further analyses. The remaining gene clusters were considered true cloud genes. The second and third steps in this analysis followed a pipeline similar to the one used in Peter et al. (2018) to distinguish between putative introgressions and putative lateral gene transfer (LGT) events. As in Peter et al. (2018), a putative introgression was identified whenever the best BLAST hits were found in closely related species (i.e., other species of the same genus), whereas presumed LGTs corresponded to transfers from more distant taxa (i.e., outside *Torulaspota* spp.). The fact that cloud genes were necessarily at low frequency in the data set (present in just one or two genomes) provided a strong indication that gene transfers occurred towards the *T. delbrueckii* genome and not on the opposite direction. We performed a BLASTP search (E -value $1e-6$) against a curated database of 70 Saccharomycetaceae proteomes. We retrieved all significant hits (E -value $1e-6$) and identified a putative introgression whenever BLASTP best hits included orthologues belonging to *Torulaspota* spp. with both query and subject coverage $\geq 60\%$ and identity coverage $\geq 60\%$ (query coverage $\geq 60\%$ and identity $\geq 60\%$ for "lower-confidence" introgression). Likewise, we considered presumed LGT whenever BLASTP best hits included orthologues belonging to non-*Torulaspota* species with similar statistics as above. Gene clusters without hits in our local Saccharomycetaceae database underwent a third round of BLASTP searches, this time against the NCBI nonredundant database (<ftp://ftp.ncbi.nlm.nih.gov/blast/db/>). Significant results from this last BLASTP analysis were used to identify distant LGT events and the remaining GCs were classified as putative orphan genes that may include transfers from unsequenced genomes, or sequencing/annotation errors. The pangenome tree was inferred using maximum-likelihood on the pangenome matrix previously estimated using the GET_PHYLOMARKERS tool estimate_pangenome_phylogenies.sh (Vinuesa et al., 2018). This script launched 10 independent IQ-TREE searches fitting binary (BIN) models.

We performed functional characterization and enrichment analyses of GCs belonging to all three pangenome classes using the PANTHER Classification System as described in Mi et al. (2019). The full pangenome was scored against the PANTHER HMM library (library of HMMER3 models, Eddy, 2009) using the PANTHER scoring Tool with the HMMSEARCH program to generate the PANTHER Generic Mapping file. Analyses were performed with PANTHER version 16.0, Release December 1, 2020, and PANTHER GO-SLIM (Molecular function, cellular component and biological process), and PANTHER Protein Class annotations. All results with E -value $<10^{-6}$ were included in the analyses. Each GC from each pangenome group was characterized and tested against the full pangenome classification using a PANTHER Overrepresentation Test (Release 20210224) with significance levels estimated with Fisher's Exact test with false discovery rate (FDR) correction. Only the categories with $p < .05$ were analysed. The functional annotation of clade-specific genes, and putative introgressed and lateral transferred cloud genes was further achieved using BlastKOALA (Kanehisa et al., 2016), KofamKOALA (Aramaki et al., 2020), FungiDB (Basenko et al., 2018) and *Saccharomyces* Genome Database (SGD; Cherry et al., 2012).

2.6 | Survey of domestication footprints

To identify domestication signature genes in *T. delbrueckii* we queried protein sequences of known *Saccharomyces cerevisiae* domestication signature genes against *T. delbrueckii* genomes. *S. cerevisiae* protein sequences were obtained from the SGD and UniProtKB (Bateman et al., 2021) database, and used to query the proteome of *T. delbrueckii* CBS 1146 in KEGG (Kyoto Encyclopedia of Genes and Genomes) using BLASTP and the SEARCH tool with best-best and paralogues options from the KEGG Sequence Similarity DataBase (SSDB). In the absence of a significant hit in CBS 1146, the query sequences were searched locally using BLASTP against a database built with all studied *T. delbrueckii* proteomes. Once an orthologous copy was identified in *T. delbrueckii*, we queried the corresponding nucleotide sequences against a local database built with all complete genomes of *T. delbrueckii* using BLASTN. This strategy enabled us to identify orthologous copies of *S. cerevisiae* signature genes in *T. delbrueckii* and to recover pseudogenes when present. Extracted sequences were aligned with MAFFT and analysed using IQ-TREE as previously described.

3 | RESULTS

3.1 | A phylogenomics view of *T. delbrueckii*

We assembled a data set of 96 genomes of *Torulaspota* spp., 67 of which were sequenced for the first time in this study (Table S1). We gathered previously sequenced genomes of good quality and added additional ones in order to provide a broad diversity with respect to geography and substrates of isolation. To better confirm species-level assignments to *Torulaspota delbrueckii* we also included in our

analysis representatives of all *Torulaspota* species currently known. A maximum-likelihood phylogeny of the genus *Torulaspota* based on 2083 concatenated single copy core genes is shown in Figure 1a. Most strains previously identified as *T. delbrueckii* clustered in a major clade that encompassed the type strain of the species, for which we added some redundancy since we analysed three genome copies corresponding to strains maintained in different culture collections (Table S1). The main clade also included two strains misidentified as *T. franciscae* and *T. microellipsoides* (Table S1). Moreover, several strains identified as *T. delbrueckii* based on sequencing of the D1/D2 and ITS regions of the rDNA were found to represent three undescribed *Torulaspota* species (Figure 1a; Table S1).

To confirm the species-level discriminations suggested by the phylogenetic analysis, we evaluated ANI within and between the clades that contained *Torulaspota* genomes. ANI has been recently assessed as a parameter to evaluate species boundaries in yeasts using genome sequence data, and the value of 95% emerged as a good guideline for the delineation of yeast species (Lachance et al., 2020). In the present study, the average ANI value within the *T. delbrueckii* clade was 98.4%, and ANI values within the other highlighted clades with three or more strains also followed the threshold indicated above (Figure 1b; Table S2). Moreover, values between clades were always lower than 90%, suggesting that each clade indeed represents a distinct species.

In conclusion, our phylogenetic analysis based on single copy core genes and the associated ANI measurements provided a high-resolution circumscription of *T. delbrueckii*, an essential step for the rest of our study. Our phylogenetic analysis suggested also six possible new species in the genus.

3.2 | *Torulaspota delbrueckii* is composed of five main lineages

To obtain a higher resolution of the relationships between lineages of *T. delbrueckii*, we ran a second phylogenetic analysis focused on single copy core genes but restricting the genome data set to the confirmed *T. delbrueckii* genomes. Five main clades could be retrieved, within a geographically organized tripartite structure, Europe (three clades), New World and Global (Figure 2a). The first clade was composed mostly of strains isolated from wine must, a well-known substrate for *T. delbrueckii*, and was therefore designated "Wine - Europe," given the geographical origin of most strains. A second clade was also composed of strains associated

with anthropic environments but could not be precisely targeted to a specific technological niche. Therefore, this clade was designated as "Mix Anthropic - Europe." Nevertheless, the absence of wine strains in this group was notable and strains isolated from bread dough and dairy products were solely found in this clade. The third clade (Arboreal - Europe) contained mostly apparently wild strains. The fourth clade contained a limited number of strains (four) and all were found in the New World. It is possible that this clade is mostly associated with natural environments, but more representatives of this lineage have to be analysed before a clear conclusion is reached. The last clade had a global distribution (America, Asia, Europe) although only six strains had geographical metadata. Its ecology is unclear, and its representatives were found both in natural and in anthropic habitats. In summary, the genetic differentiation of *T. delbrueckii* appears to be driven primarily by geography, with most of the analysed cultures associated with Europe, and secondarily by ecology. Two of the three European clades were associated with anthropic environments, whereas the third one, sister to those two clades, was associated with the arboreal niche.

3.3 | Genome diversity, ploidy and segmental duplications

Chromosomal SNP calling resulted in a data set of 698,435 bivariate SNPs, of which 381,485 had a minimum allele frequency >0.05. As expected, the number of SNPs per genome increased with phylogenetic distance to the reference, being on average ~22,000 SNPs among European strains, but 250,000 and 345,000 among strains of the New World and Global clades, respectively (Table S3). An analysis of the genome-wide distribution of SNP allele frequencies highlighted the highly homozygous nature of *T. delbrueckii* genomes, where all but two have allele frequencies close to 1, as is typical of haploid genomes or higher ploidy homozygous genomes (Figure S1, Table S3). The two exceptions, PYCC 8926 and PYCC 8927, are two strains sampled in a dairy environment (Mix-Anthropic clade) that have in all chromosomes an allele balance ratio for heterozygous SNPs of 0.5. This is the expected pattern of allele frequency in diploid genomes, although, in total, these heterozygous SNPs correspond to no more than 1% of all genotyped biallelic positions in these genomes (Table S3). Overall, *T. delbrueckii* genome diversity is rather low and similar across clades. The ANI within clades varies from 99.7% in all European clades and 98.7% in the New World clade, being 99.4% in Global (Table S4). The average proportion

FIGURE 1 Whole-genome-based circumscription of *Torulaspota delbrueckii*. (a) Phylogeny of 96 genomes of the genus *Torulaspota* inferred from a concatenated alignment of 882,364 amino acid sequences corresponding to 23,083 clusters of single copy core genes (*Zygosaccharomyces baillii* was included to root the tree). The JTT+F+I+G4 model of sequence evolution and the maximum-likelihood method were used as implemented in IQ-TREE. Nodes with black dots have ≥95% bootstrap support (fast bootstrap with NNI optimization), and branch lengths correspond to the expected number of substitutions per site. Strain designations in red depict the copies of the type strain (^T) of *T. delbrueckii*. (b) Average nucleotide sequence identity (ANI) matrix of *Torulaspota* spp. depicting intra- (coloured background) and inter- (white background) clade relationships (clades and analysed single copy core genes same as in (a)) [Colour figure can be viewed at wileyonlinelibrary.com]

of heterozygous sites per genome follows this tendency and varies between 0.15% in the New World clade and 0.27% in the Mix-Anthropic group (Table S3). Hence, patterns of diversity do not indicate a reduction in genetic diversity in the Wine and the Mix-Anthropic groups.

Furthermore, an analysis of genome-wide sequencing coverage indicated that all genomes were euploid (see Figure S1 for some examples). As opposed to what it is described for *Saccharomyces cerevisiae*, we found no evidence for the existence of chromosomal aneuploidies. Large segmental amplifications (>10 kb) were also uncommon, ranging from one to five per genome with an average of 1.9, and a maximum size of 110 kb. These amplifications occurred preferentially at subtelomeric regions and were more frequent at chromosomes IV, V and VIII (the last contains the rRNA genes, and is present in all genomes) (Tables S3 and S5). The genomes with higher number and larger segmental amplifications belonged to the Mix-Anthropic clade. Of note is an ~56-kb amplification at the beginning of chromosome V only present in the genomes of strains sampled in dairy environments, which will be discussed below.

3.4 | Population structure

We analysed population structure and admixture patterns using a data set of 66 genomes (the two redundant genomes of the type strain were excluded) and that contained 698,435 bivariate SNPs, 381,485 of which had a minimum allele frequency >0.05. The NGSADMIX analysis of individual admixture patterns recapitulated the five-cluster structure of the phylogenetic analysis and suggested minor admixture events in the data set (Figure 2b; Figure S3). Assuming models of three and four clusters ($K = 3$ and $K = 4$, Figure S2), the first genomes to segregate were those associated with anthropic environments, whereas Arboreal and New World samples remained undifferentiated. These results suggest a higher divergence of the Wine and Mix-Anthropic populations from the ancestral genetic stock. Moreover, with $K = 5$, admixed (mosaic) genomes in European samples were only found among those two populations and did not occur in the European arboreal population, which can be viewed as a possible wild reservoir. Interestingly, the single genome from the New World population with a clear association with a human-made environment (NRRL Y-50541, mezcal fermentation), had a mixed ancestry that encompassed the two anthropic populations. Next, we performed a principal component analysis in which the first two components (Figure 2c) represented ~87% of the total genetic variance in the data set and recovered the main geographical

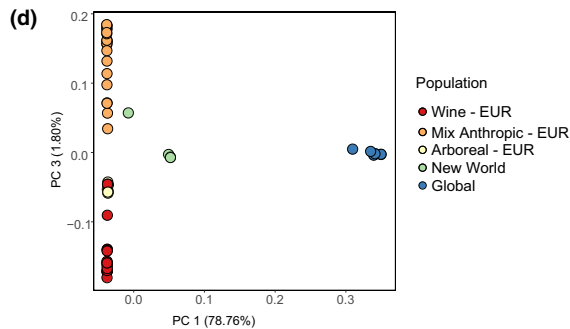
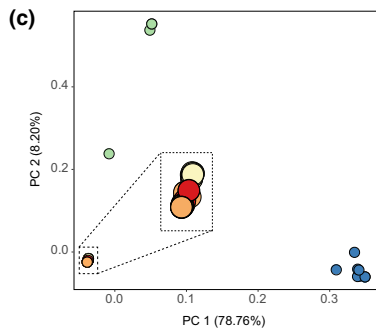
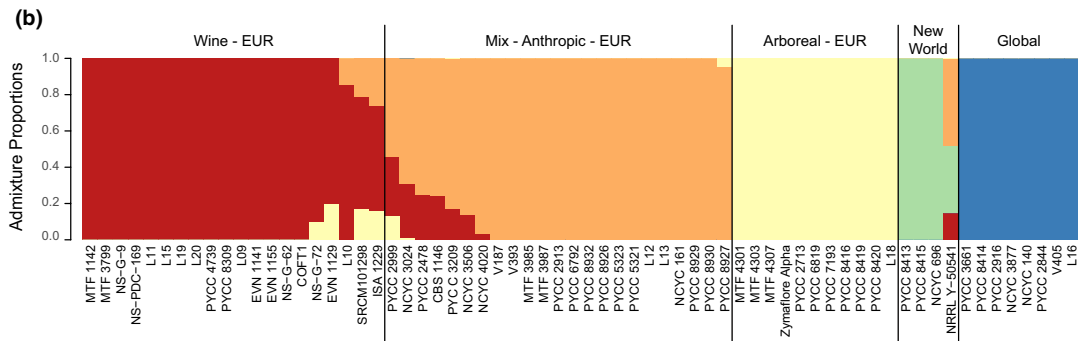
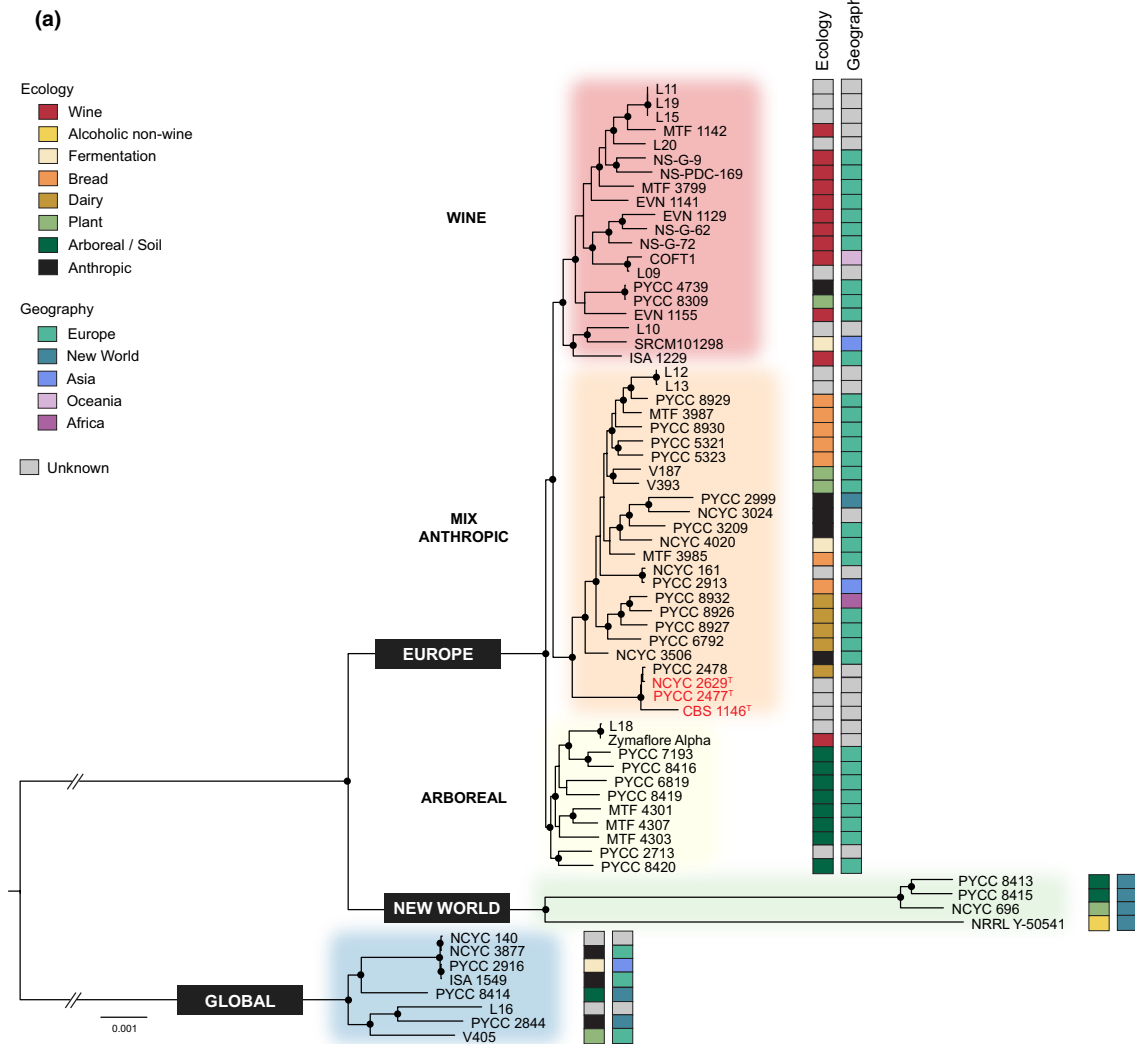
segregation already observed in the phylogenetic analysis. The third principal component discriminated the European samples into the three groups already mentioned (Figure 2d). Moreover, the overlap between the Wine and Arboreal representatives might be indicative of shared polymorphisms that have not yet segregated between the two populations.

3.5 | Variable genome

Although studies of microbial pangenomes were originally designed and implemented in prokaryotes (Tettelin et al., 2005), similar analyses in microbial eukaryotes such as *S. cerevisiae* (Peter et al., 2018) and the opportunistic pathogen *Candida glabrata* (Carreté et al., 2018) have recently been performed. Here, we used 64 genomes to assemble and investigate the pangenome of *T. delbrueckii*. The resulting gene clusters were controlled for the presence of artefacts (Supporting Results, Tables S6 and S7) and classified based on their frequency in the data set using the classification scheme of GET_HOMOLOGUES (Contreras-Moreira & Vinuesa, 2013). Two classes of high-frequency genes, the core and softcore genome, corresponded to genes present in 100% and 95% of the genomes, respectively. The softcore genome included all core GCs but is less sensitive to assembly and annotation artefacts than the strict core. We also considered a class of intermediate- to low-frequency genes, the shell genome, corresponding to genes present in 5%–95% of the genomes. Finally, the cloud genome included rare genes present in only one or two genomes (i.e., <5% of the studied genomes). These gene frequency classes are useful to characterize the intraspecific diversity in genome composition. Genes in the core and softcore classes are typically conserved and ancestrally segregating, while genes in the intermediate–low and rare frequency classes often result from phylogeographical structure, adaptation to specific environments or recent gene acquisitions from other taxa (Brito et al., 2018; Gordon et al., 2017).

The curated *T. delbrueckii* pangenome diversity totalled 5617 gene clusters, where 4289 and 4683 were core and softcore genes, respectively and 934 corresponded to the variable accessory genome (shell and cloud) (Table S6, Supporting Data set). Analysis of the different gene frequency classes across all genomes indicated a constant proportion of the softcore and shell components of the pangenome that averaged 4680 and 231 genes, respectively (Tables S8 and S9). The cloud genome, on the other hand, showed a marked difference across the five clades, representing, on average, 21 and 10 genes per genome among the representatives of the New World and Global clades, respectively, while in the other clades it varied between two and four genes per genome (Table S8).

FIGURE 2 Phylogeny and population structure of *Torulaspora delbrueckii*. (a) Phylogeny inferred from a concatenated alignment of 1,186,140 amino acid sequences corresponding to 2819 clusters of single copy core genes from 68 genomes using the JTT+F+I+G4 model of sequence evolution and the maximum-likelihood method as implemented in IQ-TREE. Nodes with black dots have ≥95% bootstrap support (fast bootstrap with NNI optimization), and branch lengths correspond to the expected number of substitutions per site. The main clades are colour-coded together with strain geographical origin and ecology. Strain designations in red depict the copies of the type strain (T^1). (b) Population structure inferred with NGSADMIX for $K = 5$. (c, d) Principal component analysis plots depicting the two principal components (c) and first and third principal components (d) [Colour figure can be viewed at wileyonlinelibrary.com]



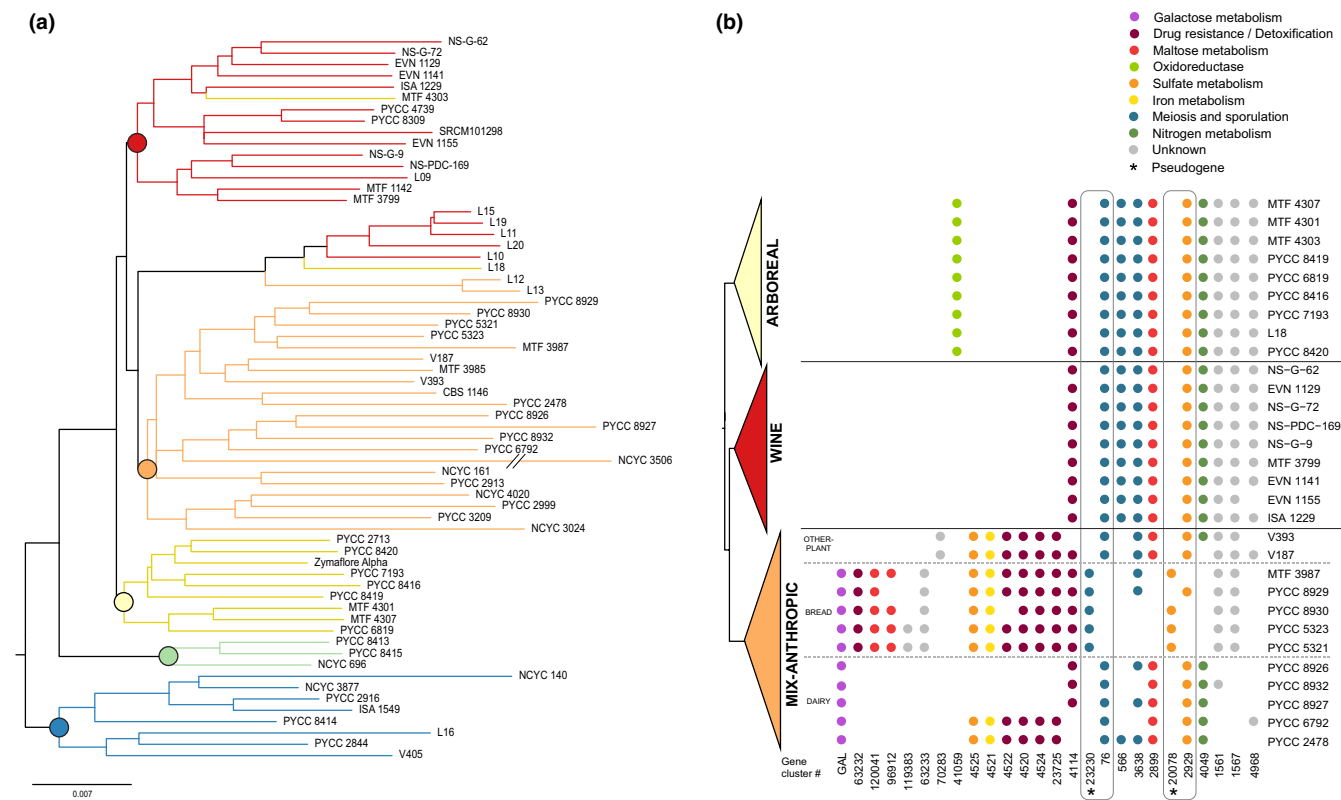


FIGURE 3 *Torulaspora delbrueckii* pangenome diversity. (a) Maximum-likelihood tree of the pangenome matrix of 64 genomes. The five main clusters are highlighted and colour-coded as in the phylogenetic analysis. (b) Lifestyle and gene content among European strains. Three ecological classes are defined for the Mix-Anthropropic clade. Dots represent gene presence and are colour-coded according to function. The GAL gene cluster comprises several genes (two to nine) implicated in galactose metabolism as shown in Table S16. Rounded rectangles indicate pairs of native/truncated (*) genes [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

Functional enrichment analyses performed with all classes of pangenome genes (softcore, shell and cloud) indicated for the shell genome (GCs with intermediate frequencies in the data set) a statistical overrepresentation of GCs associated with carbohydrate, ion and amino acid transmembrane transporter activities, symporter activities and catalytic processes (Table S10). Conversely, the cloud genome (rare GCs) had a strong statistical overrepresentation of GCs involved in aerobic respiration and ATP metabolic processes. This is probably related to the acquisition of mitochondrial DNA, as seen below in the detailed analysis of these cloud genes. None of the analyses performed with the softcore genome retrieved significantly enriched functional classes, as the softcore compartment comprises a substantial proportion of the pangenome.

We traced the origin of the cloud genome by evaluating the pattern of similarity between each gene and its significant BLASTP hits across a database of species of the Saccharomycetaceae (Table S11). We identified a total of 97 putatively introgressed genes, likely to have been acquired from closely related species (other species in the genus *Torulaspora*) and 21 putative LGTs acquired from more distant taxa (outside the genus *Torulaspora*) (Table S12). This classification of rare genes was adopted earlier (Peter et al., 2018), and should be interpreted as a broad and preliminary characterization of the non-vertically inherited genetic material as it is based on the phylogenetic

distance between donor and recipient species. The remaining 148 GCs corresponded to sequences with no significant BLASTP hits besides *T. delbrueckii*. These are putative orphan genes, acquired genes from species for which whole-genome data are not available, or errors. Again, we observed that the amount of gene gains was highly variable across genomes, being higher among genomes from the New World and Global clades with an average of nine and four introgressed genes per genome, respectively. This contrasts with the overall median of one introgressed gene per genome (Tables S9 and S12). Functional annotation indicated that several introgressed genes were associated with galactose and maltose metabolism, which justified a more detailed investigation (see below). It also revealed putative introgressions associated with other metabolic processes such as glycolysis (enolase, EC:4.2.1.11), cysteine and methionine metabolism (1-amino cyclopropane-1-carboxylate deaminase, EC:3.5.99.7), biosynthesis of inositol (1-phosphatase INM, EC:3.1.3.25), and flavone and flavonol biosynthesis (arylsulfate sulfotransferase, EC:2.8.2.22) (Table S13). This suggests that gene acquisition from other *Torulaspora* species can play a significant role in the expansion and reconfiguration of the core metabolism of *T. delbrueckii*.

Putative LGT events were unevenly distributed among our data set, being more frequent (~70% of all cases) in the Mix-Anthropropic clade (Table S14). We detected chromosomal gene acquisitions from

Lachancea, *Kazachstania*, *Pichia* and *Zygorulasporea*. We also detected mitochondrial genes apparently acquired from *Rhodotorula*, *Saccharomyces* and *Zygorulasporea*. A third category corresponded to genes associated with plasmids. In one case the 2- μ m-type plasmid was implicated. Such plasmids had already been found in a strain of *T. delbrueckii* not included in our study (Blaisonneau et al., 1997) and therefore they appear to be present but infrequently. Other cases pointed to the pGKI-1 plasmid, a *Kluyveromyces* killer plasmid. In *T. delbrueckii* the killer protein TdkT shares with the *Kluyveromyces* pGKI-1 protein the same chitin binding and chitinase activity (Villalba et al., 2016). It thus appears that TdkT and/or pGKI-1 might have been acquired by LGT.

3.6 | An incipient domestication signal in the pangenome of *T. delbrueckii*

Pangenome trees, as opposed to phylogenetic trees, are not intended to represent the evolutionary diversification process of an organism but rather to depict similarity in genome content. These trees are built from a gene presence/absence matrix, and tree search algorithms tend to cluster in close relationship strains with more similar genome content. From the comparison of pangenome and phylogenetic trees it is possible to investigate patterns of evolution of genome content, where congruent results suggest gradual changes accumulating through time. Incongruence can be viewed as an indication that other processes such as convergent evolution, introgression or LGT are of relevance. We performed a maximum-likelihood analysis of the pangenome matrix (Figure 3a) and recovered clusters that matched, for the most part, the clades shown in Figure 2a. The most relevant discrepancy concerned strains from the commercial yeast-producing company Lallemand (Table S1) that phylogenetically were placed in the three European clades (Figure 2a) but in the pangenome tree were grouped together in a separate cluster, intermediate between the Wine and the Mix-Anthropoc pangenome clusters (Figure 3a). Since information about the history of these strains is lacking, the reasons behind this apparent higher similarity with respect to genome content are not evident.

To further explore possible genome-level differences between the three groups of European strains while taking into account the environments from which they were isolated, we analysed patterns of gene presence/absence in a subset of strains for which detailed ecological information was available (Table S1). We selected appropriate strains from the Wine and Arboreal clades and, given the diverse sources of the strains in the Mix-Anthropoc clade, we considered the Plant, Dairy and Bread ecological classes in this clade. Whereas the dairy and bread strains are obvious candidates to investigate domestication-related changes, the strains isolated from processed plant materials (green beans and artichoke) have an unclear ecology and appear to be associated with human-made artificial niches. This analysis highlighted the Mix-Anthropoc clade as the one with the most specialized genome content, with more than 10 genes found exclusively in this clade (Figure 3b; Tables S15 and S16). These

genes participate in galactose and maltose assimilation, in iron and sulphur metabolism, and in drug resistance/detoxification. This contrasts with the absence of exclusive genes in the Wine clade and with a single exclusive gene, an oxidoreductase, in the Arboreal clade. This analysis suggested that wine genomes are more similar to wild European genomes than to genomes of the Mix-Anthropoc clade. It also suggested that the Mix-Anthropoc clade represents the most differentiated clade, with several cases of gene gain and of gene loss relative to the other two clades that appear strikingly similar. Interestingly, the genes exclusive of the Mix-Anthropoc clade and present in the three ecological classes encoded mostly enzymes or transcriptional activators associated with resistance to drugs or to detoxification processes, thus supporting the specialization hypothesis for this clade. Also, genes that tend to be absent in this clade (albeit not universally), and consistently present in the other two clades, are those involved in meiosis and sporulation, thus suggesting that sexual reproduction might be impaired in this group. In one case, the mating-type switching endonuclease gene had a considerably shorter sequence and therefore our searching algorithm placed it in a distinct GC (Figure 3b; Table S15). The shorter sequence was detected only in the bread strains and had a deletion (A842del) that caused a stop codon. Therefore, these strains have a nonfunctional HO gene and ability to reproduce sexually is probably impaired. Interestingly, in other yeasts, loss of sexual reproduction is viewed as a case of genome decay resulting from domestication (Steensels et al., 2019). A similar case of pseudogene formation was observed for *MMP1*, a gene coding for an S-methylmethionine permease. A pseudogene with a shorter sequence was detected only in bread strains. The exclusive genes of each of the three ecological classes suggested particular specialization trajectories. The dairy and bread strains had several genes related to galactose metabolism. The bread strains had the highest number of exclusive genes, some related to maltose metabolism.

3.7 | Comparison with *S. cerevisiae* domestication footprints

Given that *T. delbrueckii* and *S. cerevisiae* can co-occur in some fermentations (e.g., bread dough, wine must), we performed a guided search of genome-level changes known to be relevant for the adaptation of *S. cerevisiae* to anthropic niches. In this species, resistance to copper sulphate, an antifungal commonly used in vineyards, is mediated by *CUP1*. This trait is selected for in wine strains that tend to exhibit an expansion of the numbers of copies of *CUP1* (Legras et al., 2018). However, for *T. delbrueckii* we could not detect a *CUP1* homologue of this gene. Next, we considered the gene *SSU1* that codes for a sulphite efflux pump and that has been implicated in sulphite resistance in *S. cerevisiae* wine strains because chromosomal translocations of *SSU1* can lead to increased expression of this gene (Steensels et al., 2019). We investigated the chromosomal context of the corresponding homologue detected in *T. delbrueckii* and found it exclusively located in chromosome I, even in the strains of the Wine clade.

Another domestication-related gene in *S. cerevisiae* is *AQY* that codes for aquaporins and that is frequently nonfunctional in wine strains of this species. Aquaporin loss increases fitness in sugar-rich and high-osmolarity environments, whereas in the arboreal environment functional aquaporins are selected for (Gonçalves et al., 2016; Will et al., 2010). We surveyed a *T. delbrueckii* homologue of the *S. cerevisiae* paralogous aquaporin genes *AQY1* and *AQY2*, and detected a nonfunctional allele with a T756 deletion in most wine strains (Figure S3). Interestingly, although the dairy-related strains of the Mix-Anthropoc clade had a functional aquaporin gene, most of the remaining strains in this clade, notably the bread-related strains, had a nonfunctional *AQY1* orthologue (T756 deletion or TG48 insertion). The wild lineage from Europe and the New World lineage had a functional aquaporin and most representatives were associated with the arboreal niche. Finally, the Global clade exhibited a mixed situation with functional and nonfunctional aquaporins and a distinct mutation (T387del). It thus appears that normal aquaporin function was not selected for during adaptation to wine and bread fermentations, similar to what has been reported for *S. cerevisiae*. By contrast, wild populations and an association with dairy products maintained the functional version of the gene. The *AQY* inactivation observed in the Global population, which we confirmed through PCR, appears puzzling at this stage.

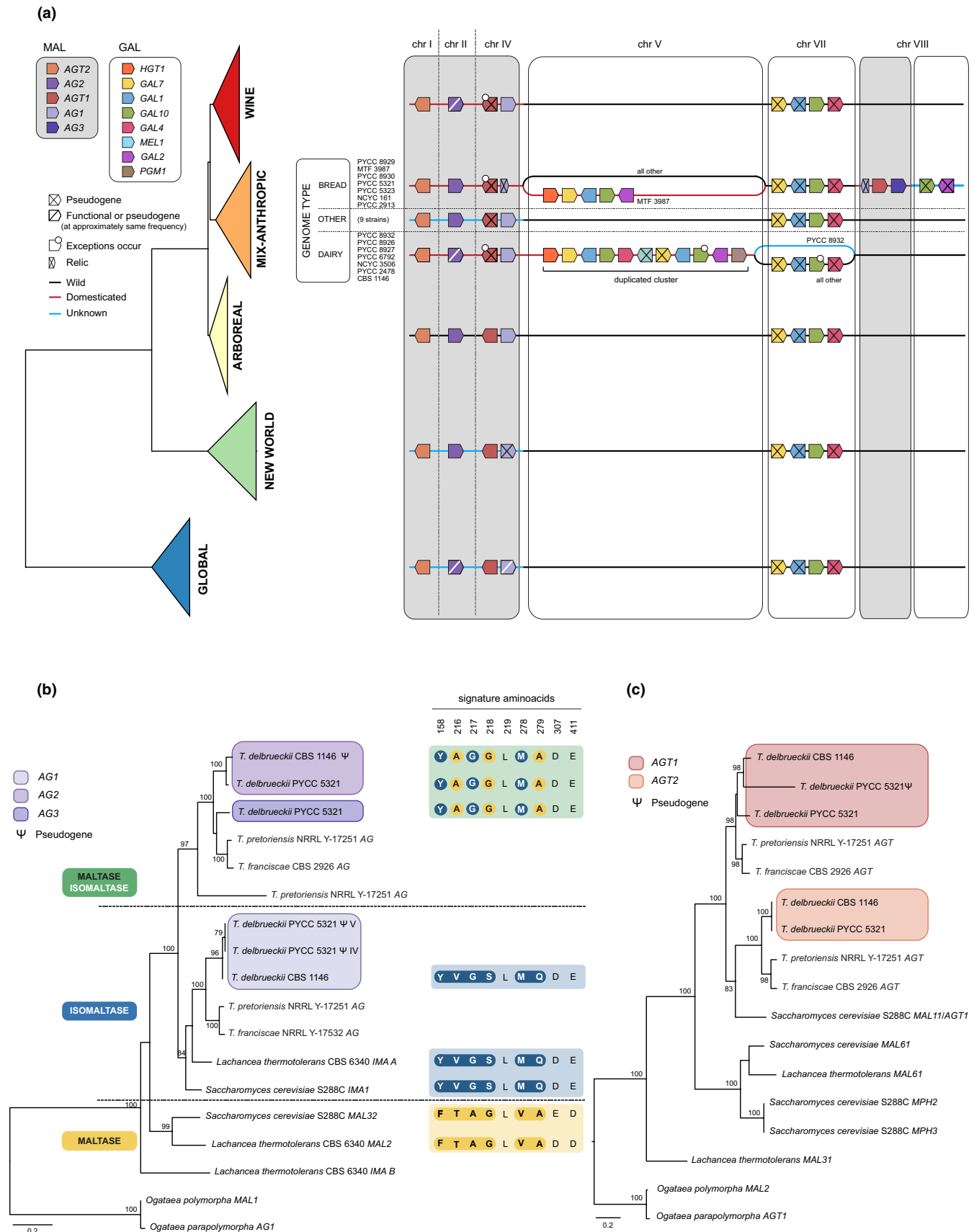
3.8 | Expansion and inactivation of *GAL* and *MAL* genes

Given previous indications gathered in earlier stages of this study, we analysed in more detail the genes implicated in galactose and maltose metabolism. It was recently reported that species of *Torulopsis* have large clusters of genes necessary for galactose metabolism (Venkatesh et al., 2020). These clusters have the *GAL 7-1-10-4* and the *GAL 7-1-10-2* configuration (other genes, such as *HGT1*, *MEL1* and *PGM1*, are also part of the *GAL* cluster). It was also proposed that the type strain of *T. delbrueckii*, but not 14 other strains of the species, received from *T. franciscae* a cluster present in chromosome V with the same gene organization both in *T. franciscae* and in the type strain of *T. delbrueckii*. Moreover, all the 15 strains analysed by Venkatesh et al. had another cluster in chromosome VII, in which *GAL10* was the only intact gene (*GAL 7-1-10-4* cluster), and that was consequently viewed as a primordial

cluster that has now almost disappeared. We analysed the presence/absence of these clusters and detected the ancestral cluster of chromosome VII in all but one genome (Figure 4a; Figure S4). Moreover, we detected the cluster transferred from *T. franciscae* not only in the type strain but also in six additional strains (Figure 4a; Figure S4). Contrary to what is observed in *T. franciscae*, we detected this cluster duplicated in chromosome V of *T. delbrueckii*. Interestingly, all the strains harbouring this transferred cluster were positioned at the base of the Mix-Anthropoc clade and this subgroup contained all the dairy-related strains included in our study. It thus appears that this acquisition is related to the colonization of a particular anthropic environment, that of milk products, being absent in other human-associated strains such as those from wine and bread dough, and in wild strains. We also found in chromosome V an incomplete version of the cluster, in a single strain (MTF 3987), and in chromosome VIII syntenic inactivated remnants of *GAL2* and *GAL10* in seven strains. Interestingly, all these cases involved strains of the same clade but not related to dairy products (Figure 4a). A phylogeny of the *GAL2* and *GAL10* genes separated the dairy and nondairy alleles (Figure S5). A detailed inspection of the alignment revealed extensive gene degeneration associated with loss of function mutations in the nondairy strains.

In *Saccharomyces*, α -glucosides are hydrolysed by two classes of enzymes. Maltases degrade maltose and maltotriose that have α -1,4 linkages, and isomaltases degrade isomaltose, palatinose, and methyl α -D-glucoside that have α -1,6 linkages. Like for *GAL* genes, clusters of *MAL* genes, located in subtelomeric regions, have been documented in *S. cerevisiae* (Brown et al., 2010) and in *T. delbrueckii* (Viigand et al., 2018). For *T. delbrueckii*, a single cluster containing an α -glucosidase (*AG1*) and an α -glucoside transporter (*AGT1*), two genes that are divergently transcribed, was detected in the genome of the type strain (Viigand et al., 2018). However, in another strain (PYCC 5321), the disruption of the transporter gene *AGT1* reduced but did not prevent maltose transport, and therefore the presence of at least two functional transporter genes was hypothesized (Alves-Araújo et al., 2004). Here we confirmed the presence of a cluster of divergently transcribed *AG1-AGT1* in all the 66 genomes analysed (Figure 4a). Moreover, we failed to detect an orthologue of *MAL*-activator genes (*MAL3x*), the third component of the *S. cerevisiae* cluster, besides the genes coding for the enzyme (*MAL2x*) and transporter (*MAL1x* or *AGT1*), a situation already reported for the type strain of *T. delbrueckii* (Viigand et al., 2018).

FIGURE 4 Occurrence and evolutionary relationships of *GAL* and *MAL* genes in the five clades of *Torulopsis delbrueckii*. (a) Occurrence and synteny of *GAL* and *MAL* genes and clusters. Single or multiple gene combinations per clade are indicated. Horizontal lines represent chromosomes and are colour-coded according to inferred wild or domesticated state. (b) Maximum-likelihood phylogeny of maltase (*MAL*, *AG*) and isomaltase (*IMA*) genes of selected species of the Saccharomycetaceae, depicting the phylogenetic relationships of *AG1*, *AG2* and *AG3* of *T. delbrueckii* (outgroup *Ogataea* spp.; LG+I+G4 model of sequence evolution, fast bootstrap with NNI optimization). Signature amino acids that correlate to substrate specificity for maltases and isomaltases, numbered as in *Saccharomyces cerevisiae* *IMA1*, are compared for maltase-, isomaltase- and maltase-isomaltase-like sequences. Relevant amino acid signatures for maltase and isomaltase activity are highlighted in yellow and blue, respectively, and mixed maltase-isomaltase affinity is colour-coded in green. Note that A216 is associated with a maltase in *Lipomyces starkeyi* and *Shizosaccharomyces pombe* (Viigand et al., 2018). (c) Maximum-likelihood phylogeny (constructed as in (b)) of α -glucoside transporter genes prepared as in (b), depicting the phylogenetic relationships of *AGT1*, *AGT2* of *T. delbrueckii* [Colour figure can be viewed at wileyonlinelibrary.com]



Besides detecting the two genes (AGT1 and AG1) in chromosome IV, we detected in all genomes an additional transporter gene (AGT2) in chromosome I and an additional enzyme-coding gene, AG2,

in chromosome II (Figure 4a). Moreover, for a restricted number of seven genomes, all in the Mix-Anthrope clade and most in the Bread subclade, an additional cluster was detected in chromosome VIII.

This cluster corresponds to a duplication of the two genes detected in chromosome IV, inserted in chromosome V in an inverted position, and to a distinct gene, AG3. Except for AGT2 and AG3 (chromosome VIII), all other genes have become pseudogenes in several strains, although a clear correlation between the inactivation pattern and the phylogeny could not be discerned. The phylogenetic analysis of the AG genes is shown in Figure 4b for *T. delbrueckii* and its closest relatives, *T. franciscae* and *T. pretoriensis*, and for *Lachancea thermotolerans* and *S. cerevisiae*. While the latter two species contain one or more versions of a maltase-like and isomaltase-like gene, in *T. delbrueckii* and its siblings the orthologues of *S. cerevisiae* MAL32 (maltase) have been lost and only the isomaltase-like genes were retained. A phylogenetic and functional relationship could be inferred for *S. cerevisiae* IMA1 and *T. delbrueckii* AG1. Taking into consideration the nine amino acid residues that are located close to the active site pocket of the enzyme and that define signatures correlated, at least to some extent, with substrate specificity for maltases or isomaltases (Viigand et al., 2018; Voordeckers et al., 2012; Yamamoto et al., 2004), we observed that three main substitutions occurred between *T. delbrueckii* AG1 and AG2-AG3 sequences (Figure 4b). Whereas the AG1 amino acid sequence had a subset of three signature positions typical of the *S. cerevisiae* isomaltase, the AG2-AG3 sequences had distinct amino acids at those positions and the changes were compatible with an increase of maltase activity, as seen in *S. cerevisiae* and other yeast species (Viigand et al., 2018). These changes do not necessarily imply that the isomaltase activity was lost. Rather, they probably caused slight shifts in the binding preference that gave rise to an enzyme with mixed isomaltase-maltase activity.

A similar analysis was carried out for the AGT protein sequences of the α -glucoside transporter. We found that *T. delbrueckii* and its siblings *T. franciscae* and *T. pretoriensis* had two AGT genes, AGT1 and AGT2, and that the *S. cerevisiae* AGT1 (MAL11) gene was more related to the *Torulaspora* AGT2 gene (Figure 4c).

In conclusion, our analysis of MAL genes in *T. delbrueckii* revealed a complex situation with expansion and, in some cases, subsequent inactivation of some (but not all) of the enzyme- and transporter-encoding genes. Genes coding for isomaltases and maltases-isomaltases were detected. The *T. delbrueckii* gene coding for an enzyme with presumed maltase activity was not phylogenetically related to the *S. cerevisiae* maltase gene although, by convergent evolution, it gained the signature amino acids relevant for maltose hydrolysis. AGT2 was the single gene present in a functional state in all strains. For the European strains, only those in the Arboreal clade consistently lacked gene inactivations, whereas all and 90% of those in the Mix-Anthropic and Wine clade had at least one gene degenerated, respectively. Gene inactivations were also lacking in the New World clade. Gene inactivations could be a sign of genome-level adjustments caused by adaptation processes related to the transition from wild to artificial environments. Moreover, in the Bread subclade of the Mix-Anthropic clade, contrary to all other groups, an additional cluster in chromosome VIII was detected. It contained AG3, which codes for an enzyme that, based on signature nucleotides, seems to have a mixed maltose-isomaltose hydrolysing ability.

Given that the strains in this subclade are exposed to maltose-rich environments, this acquisition is likely to be related to domestication in the breadmaking environment.

4 | DISCUSSION

Here, we performed the first comprehensive phylogenomic analysis of the genus *Torulaspora* and clarified the delimitation of *Torulaspora delbrueckii*. In addition, we revealed that some strains formerly classified in this species do in fact belong to distinct cryptic species, thus expanding considerably the documented diversity in this genus. We were able to detect five clades in *T. delbrueckii*, two of them corresponding to strains associated with human activities, even though our strain data set might be hampered by the uneven geographical representation of the strains currently available. A previous genotyping study of *T. delbrueckii* using microsatellite markers revealed a similar population structure with two presumed domesticated groups, the Wine group and a group gathering strains from different sources that matches the Mix-Anthropic clade (Albertin et al., 2014). We detected very high levels of homozygosity in *T. delbrueckii* genomes that typically have a frequency of heterozygous sites <1%. Such a pattern is compatible with a haploid genome. The phylogenetic and populational inference of two domesticated lineages, Wine and Mix-Anthropic, was corroborated by a suite of gene-level changes that sustain a broad hypothesis of domestication (i.e., changes linked to adaptation to human-driven fermentations). For example, wine strains of *T. delbrueckii*, in contrast to their wild closest relatives, have nonfunctional aquaporins, suggesting an adaptation to withstand the high osmotic stresses in wine must, as has been documented for *Saccharomyces cerevisiae* (Will et al., 2010). Moreover, within the Mix-Anthropic clade, dairy strains have an enlarged repertoire of genes for galactose metabolization, while bread strains show the same for maltose utilization. Together, these results provide genome-level evidence for domestication in *T. delbrueckii*, similarly to what was recently reported for *Brettanomyces bruxellensis* (Roach & Borneman, 2020) and *Kluyveromyces lactis* (Varela et al., 2019), thus expanding our knowledge on the domestication of non-*Saccharomyces* yeasts.

Striking differences are apparent when the domestication trajectories of *T. delbrueckii* and *S. cerevisiae* are compared. In *S. cerevisiae*, wine fermentation in the West, and Sake or other cereal-based fermentations in the East, played a central role in shaping primarily domesticated genotypes, from which domesticates emerged secondarily, exemplified by the genotypes of (ale) beer or bread strains (Barbosa et al., 2018; Fay et al., 2019). By contrast, in *T. delbrueckii*, a "non-wine" lineage, the Mix-Anthropic clade, emerges as a more specialized and diverse lineage. We suggest that the recurrent detection, in strains of this clade, of genomic changes compatible with adaptations to thrive in milk and in bread dough is a sign of recent and ongoing specializations. Moreover, since closely related strains are also found in other anthropic niches, we hypothesize that this clade combines enhanced attributes for the colonization of the

human ecosystem, a remarkable ecological plasticity that might be related to its recent and incomplete specialization. Thus, the Mix-Anthropic clade appears as a unique hotspot of domestication in this species.

We determined a pangenome of 5617 gene clusters in *T. delbrueckii*, 4289 (76.4%) of which make up the core genome and 934 (16.6%) the variable genome. A similar study conducted for *S. cerevisiae* and involving 14 times more genomes yielded a pangenome of 7796 genes, 36.7% of which belong to the variable genome (Peter et al., 2018). Thus, in *T. delbrueckii*, the weight of the variable genome in the species pangenome is smaller. This might be related to a more incipient stage in the domestication path, as adaptation to different anthropic niches in *S. cerevisiae* appears to have contributed to gene diversity across lineages (Peter et al., 2018).

In contrast to *S. cerevisiae*, presence of which is critical for successful wine fermentation or bread leavening, *T. delbrueckii* cannot be viewed as essential in wine fermentations, although its presence might impart important benefits, as recent research has emphasized (Pacheco et al., 2012; Padilla et al., 2016). Although the situation in artisanal corn and rye bread fermentations might be different, with *T. delbrueckii* being frequently found among the starter cultures (Almeida & Pais, 1996b) and likely to play a relevant role, a detailed assessment of the contribution of *T. delbrueckii* is not yet available. Therefore, the direct influence on the characteristics of the fermented product, an important hallmark of a domesticated microbe, cannot be fully ascertained at present for *T. delbrueckii*. We have recently proposed the concept of a quasidomesticated for such cases of evident genetic and genomic changes caused by adaptation to anthropic niches that cannot be associated with a clear role in the outcome of the fermentation (Pontes et al., 2019). Even if quasidomestication is the most appropriate framework to understand the genomic changes observed in *T. delbrueckii*, the improved understanding of phylogeny, population structure and genomic changes that we provide in this study can inform new programmes of strain improvement with specific aims such as oenology or bread leavening. For example, given that the genetic machinery for utilization of α -glucosides can be present in different states in different strains, a rational programme for strain selection can now be implemented.

Finally, this study opens a wider view of yeast domestication beyond *S. cerevisiae*. Even if these two species share relevant physiological features and overlap in natural and artificial niches, we have provided evidence for considering that the emergence of domesticated lineages of *T. delbrueckii* did not follow the trajectory known for *S. cerevisiae*. The early stages of domestication detected in *T. delbrueckii* are not shaped around alcoholic fermentations, in general, and wine fermentations in particular. Although wine strains can be recognized, they appear to be poorly differentiated from their wild relatives. Rather, adaptation to the dairy environment by the re-acquisition of the GAL operon was probably a first major domestication step, which is corroborated by the position of the dairy strains

at the base of the Mix-Anthropic clade. This transition might have facilitated the colonization of other anthropic niches, thus explaining the diverse substrates from which the representatives of the Mix-Anthropic clade have been isolated. We suggest that a secondary transition to maltose-rich environments then occurred, with the concomitant rapid loss of most of the *de novo* acquired GAL genes, providing a remarkable example of radical and contradictory changes driven by domestication. Another significant feature is the fact that in *T. delbrueckii* these early domestication events are not associated with major chromosomal structural alterations, such as variations in ploidy, aneuploidy or large changes in genome content, as was described for domesticated strains of *S. cerevisiae* (Peter et al., 2018). It is also remarkable that, in contrast to *S. cerevisiae* where dairy and bread populations are well demarcated (Legras et al., 2018), multiple domestication trajectories are seen in the same clade, thus suggesting a more recent timing of the domestication events or intrinsic features of the domestication process in *T. delbrueckii* that await detailed scrutiny. Together, these findings have implications for the rational improvement of biotechnology-relevant microorganisms and they also offer a first insight on a domestication process in an early stage. Moreover, they expand our views on the mechanisms of microbe domestication and on the trajectories leading to adaptation to anthropic niches.

ACKNOWLEDGMENTS

We thank Carole Camarasa (INRA, Montpellier, France), Javier Ruiz (Complutense University of Madrid, Spain), Filomena Duarte (EVN - Estação Vitivinícola Nacional, INIAV, Dois Portos, Portugal), and ECOFILTRA for kindly providing strains included in this study. Dr Jo Dicks, Quadram Institute Bioscience, UK, is gratefully acknowledged for making available genome sequences of NCYC strains prior to publication.

CONFLICT OF INTEREST

The authors have no conflicts of interest.

AUTHOR CONTRIBUTIONS

Conceptualization: J.P.S., M.J.S., P.H.B. Investigation: A.P., M.S., P.H.B., P.S., R.F.-D. Visualization: A.P., M.S. Funding: J.P.S., M.J.S. Writing: J.P.S., P.H.B. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The data presented in this study (sequencing reads) can be accessed on NCBI Bioprojects: PRJEB46640 (SAMEA8995727–SAMEA8995744 and SAMEA8997517–SAMEA8997546), PRJEB49385 (SAMEA11106836–SAMEA11106847) and PRJEB49423 (SAMEA11951367–SAMEA11951374). Accession numbers for assembled genomes are given in Table S1.

ORCID

José Paulo Sampaio  <https://orcid.org/0000-0001-8145-5274>
 Patrícia H. Brito  <https://orcid.org/0000-0002-2457-7520>

REFERENCES

- Albertin, W., Chasseriaud, L., Comte, G., Panfili, A., Delcamp, A., Salin, F., Marullo, P., & Bely, M. (2014). Winemaking and bioprocesses strongly shaped the genetic diversity of the ubiquitous yeast *Torulaspora delbrueckii*. *PLoS One*, 9(4), e94246. <https://doi.org/10.1371/journal.pone.0094246>
- Almeida, M. J., & Pais, C. (1996a). Leavening ability and freeze tolerance of yeasts isolated from traditional corn and rye bread doughs. *Applied and Environmental Microbiology*, 62(12), 4401–4404. <https://doi.org/10.1128/aem.62.12.4401-4404.1996>
- Almeida, M. J., & Pais, C. S. (1996b). Characterization of the yeast population from traditional corn and rye bread doughs. *Letters in Applied Microbiology*, 23(3), 154–158. <https://doi.org/10.1111/J.1472-765X.1996.TB00053.X>
- Alves-Araújo, C., Almeida, M. J., Sousa, M. J., & Leão, C. (2004). Freeze tolerance of the yeast *Torulaspora delbrueckii*: Cellular and biochemical basis. *FEMS Microbiology Letters*, 240(1), 7–14. <https://doi.org/10.1016/j.femsle.2004.09.008>
- Alves-Araújo, C., Hernandez-Lopez, M., Sousa, M., Prieto, J., & Rande-Gil, F. (2004). Cloning and characterization of the gene encoding a high-affinity maltose transporter from *Torulaspora delbrueckii*. *FEMS Yeast Research*, 4(4–5), 467–476. [https://doi.org/10.1016/S1567-1356\(03\)00208-3](https://doi.org/10.1016/S1567-1356(03)00208-3)
- Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, 36(7), 2251–2252. <https://doi.org/10.1093/bioinformatics/btz859>
- Azzolini, M., Fedrizzi, B., Tosi, E., Finato, F., Vagnoli, P., Scrinzi, C., & Zapparoli, G. (2012). Effects of *Torulaspora delbrueckii* and *Saccharomyces cerevisiae* mixed cultures on fermentation and aroma of Amarone wine. *European Food Research and Technology*, 235(2), 303–313. <https://doi.org/10.1007/s00217-012-1762-3>
- 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. <https://doi.org/10.1089/cmb.2012.0021>
- Barbosa, R., Pontes, A., Santos, R. O., Montandon, G. G., de Ponzes-Gomes, C. M., Morais, P. B., Gonçalves, P., Rosa, C. A., & Sampaio, J. P. (2018). Multiple rounds of artificial selection promote microbe secondary domestication – The case of cachaça yeasts. *Genome Biology and Evolution*, 10(8), 1939–1955. <https://doi.org/10.1093/gbe/evy132>
- Basenko, E. Y., Pulman, J. A., Shanmugasundram, A., Harb, O. S., Crouch, K., Starns, D., Warrenfeltz, S., Aurrecochea, C., Stoeckert, C. J., Kissinger, J. C., Roos, D. S., & Hertz-Fowler, C. (2018). FungiDB: An integrated bioinformatic resource for fungi and oomycetes. *Journal of Fungi*, 4(1), 39. <https://doi.org/10.3390/jof4010039>
- Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezzer, T. G., Fan, J., Castro, L. G., ... Teodoro, D. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1), D480–D489. <https://doi.org/10.1093/nar/gkaa1100>
- Blaisonneau, J., Sor, F., Cheret, G., Yarrow, D., & Fukuhara, H. (1997). A circular plasmid from the yeast *Torulaspora delbrueckii*. *Plasmid*, 38(3), 202–209. <https://doi.org/10.1006/plas.1997.1315>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Brito, P. H., Chevreux, B., Serra, C. R., Schyns, G., Henriques, A. O., & Pereira-Leal, J. B. (2018). Genetic competence drives genome diversity in *Bacillus subtilis*. *Genome Biology and Evolution*, 10(1), 108–124. <https://doi.org/10.1093/gbe/evx270>
- Brown, C. A., Murray, A. W., & Verstrepen, K. J. (2010). Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Current Biology*, 20, 895–903. <https://doi.org/10.1016/j.cub.2010.04.027>
- Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One*, 12(10), e0185056. <https://doi.org/10.1371/journal.pone.0185056>
- Carreté, L., Ksiezopolska, E., Pegueroles, C., Gómez-Molero, E., Saus, E., Iraola-Guzmán, S., Loska, D., Bader, O., Fairhead, C., & Gabaldón, T. (2018). Patterns of genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. *Current Biology*, 28(1), 15–27.e7. <https://doi.org/10.1016/j.cub.2017.11.027>
- Carvalho, C., Tomás, A., Libkind, D., Imanishi, Y., & Sampaio, J. P. (2018). *Zygotorulaspora chibaensis* sp. nov. and *Zygotorulaspora danielsina* sp. nov., novel ascomycetous yeast species from tree bark and soil. *International Journal of Systematic and Evolutionary Microbiology*, 68(8), 2633–2637. <https://doi.org/10.1099/ijsem.0.002889>
- Castelli, T. (1954). Les agents de la fermentation vinaire. *Archiv Für Mikrobiologie*, 20(4), 323–342. <https://doi.org/10.1007/BF00690877>
- Cheeseman, K., Ropars, J., Renault, P., Dupont, J., Gouzy, J., Branca, A., Abraham, A.-L., Ceppi, M., Conseiller, E., Debuchy, R., Malagnac, F., Goarin, A., Silar, P., Lacoste, S., Sallet, E., Bensimon, A., Giraud, T., & Brygoo, Y. (2014). Multiple recent horizontal transfers of a large genomic region in cheese making fungi. *Nature Communications*, 5, 2876. <https://doi.org/10.1038/ncomms3876>
- Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., ... Wong, E. D. (2012). *Saccharomyces* genome database: The genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1), D700–D705. <https://doi.org/10.1093/nar/gkr1029>
- Contreras-Moreira, B., & Vinuesa, P. (2013). GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology*, 79(24), 7696–7701. <https://doi.org/10.1128/AEM.02411-13>
- Crisuolo, A., & Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*, 10(1), 210. <https://doi.org/10.1186/1471-2148-10-210>
- Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernysky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491–501. <https://doi.org/10.1038/ng.806>
- Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature*, 418(6898), 700–707. <https://doi.org/10.1038/nature01019>
- Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform*, 23, 205–211.
- Fay, J. C., Liu, P., Ong, G. T., Dunham, M. J., Cromie, G. A., Jeffery, E. W., Ludlow, C. L., & Dudley, A. M. (2019). A polyploid admixed origin of beer yeasts derived from European and Asian wine populations. *PLoS Biology*, 17(3), e3000147. <https://doi.org/10.1371/journal.pbio.3000147>
- Gibbons, J. G., Salichos, L., Slot, J. C., Rinker, D. C., McGary, K. L., King, J. G., Klich, M. A., Tabb, D. L., McDonald, W. H., & Rokas, A. (2012). The evolutionary imprint of domestication on genome variation and function of the filamentous fungus *Aspergillus oryzae*.

- Current Biology*: CB, 22(15), 1403–1409. <https://doi.org/10.1016/j.cub.2012.05.033>
- Gonçalves, M., Pontes, A., Almeida, P., Barbosa, R., Serra, M., Libkind, D., Hutzler, M., Gonçalves, P., & Sampaio, J. P. (2016). Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Current Biology*, 26, 2750–2761. <https://doi.org/10.1016/j.cub.2016.08.040>
- Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., Stritt, C., Roulin, A. C., Schackwitz, W., Tyler, L., Martin, J., Lipzen, A., Dochy, N., Phillips, J., Barry, K., Geuten, K., Budak, H., Juenger, T. E., Amasino, R., ... Vogel, J. P. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications*, 8(1), 1–13. <https://doi.org/10.1038/s41467-017-02292-8>
- Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>
- Hernandez-Lopez, M. J., Prieto, J. A., & Randez-Gil, F. (2003). Osmotolerance and leavening ability in sweet and frozen sweet dough. Comparative analysis between *Torulaspora delbrueckii* and *Saccharomyces cerevisiae* baker's yeast strains. *Antonie van Leeuwenhoek*, 84(2), 125–134. <https://doi.org/10.1023/A:1025413520192>
- Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, 35(2), 518–522. <https://doi.org/10.5281/zenodo.854445>
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jeremiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6), 587–589. <https://doi.org/10.1038/nmeth.4285>
- Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of Molecular Biology*, 428(4), 726–731. <https://doi.org/10.1016/j.jmb.2015.11.006>
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Kotzekidou, P. (1997). Identification of yeasts from black olives in rapid system microtitre plates. *Food Microbiology*, 14(6), 609–616. <https://doi.org/10.1006/fmic.1997.0133>
- Lachance, M. A. (1995). Yeast communities in a natural tequila fermentation. *Antonie van Leeuwenhoek*, 68(2), 151–160. <https://doi.org/10.1007/BF00873100>
- Lachance, M. A., Lee, D. K., & Hsiang, T. (2020). Delineating yeast species with genome average nucleotide identity: a calibration of ANI with haplontic, heterothallic *Metschnikowia* species. *Antonie van Leeuwenhoek*, 113(12), 2097–2106. <https://doi.org/10.1007/s10482-020-01480-9>
- Lee, I., Kim, Y. O., Park, S. C., & Chun, J. (2016). OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology*, 66(2), 1100–1103. <https://doi.org/10.1099/IJSEM.0.000760>
- Legras, J.-L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., Sanchez, I., Couloux, A., Guy, J., Franco-Duarte, R., Marcet-Houben, M., Gabaldon, T., Schuller, D., Sampaio, J. P., Dequin, S., & Wittkopp, P. (2018). Adaptation of *S. cerevisiae* to fermented food environments reveals remarkable genome plasticity and the footprints of domestication. *Molecular Biology and Evolution*, 35(7), 1712–1727. <https://doi.org/10.1093/molbev/msy066>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9), 2178–2189. <https://doi.org/10.1101/gr.1224503>
- Limtong, S., & Koowadjanakul, N. (2012). Yeasts from phylloplane and their capability to produce indole-3-acetic acid. *World Journal of Microbiology and Biotechnology*, 28(12), 3323–3335. <https://doi.org/10.1007/s11274-012-1144-9>
- Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., Polouchine, N., Shakhova, V., Grigoriev, I., Lou, Y., Rohksar, D., Lucas, S., Huang, K., Goodstein, D. M., Hawkins, T., Plengvidhya, V., ... Mills, D. (2006). Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 103(42), 15611–15616. <https://doi.org/10.1073/PNAS.0607117103>
- Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, 210(2), 719–731. <https://doi.org/10.1534/GENETICS.118.301336>
- Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., & Thomas, P. D. (2019). Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols*, 14(3), 703–721. <https://doi.org/10.1038/s41596-019-0128-8>
- 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. *Molecular Biology and Evolution*, 32(1), 268–274. <https://doi.org/10.1093/molbev/msu300>
- Pacheco, A., Santos, J., Chaves, S., Almeida, J., Leo, C., & Joo, M. (2012). The emerging role of the yeast *Torulaspora delbrueckii* in bread and wine production: Using genetic manipulation to study molecular basis of physiological responses. In A. A. Eissa (Ed.), *Structure and function of food engineering*. IntechOpen.
- Padilla, B., Gil, J. V., & Manzanares, P. (2016). Past and future of non-*Saccharomyces* yeasts: From spoilage microorganisms to biotechnological tools for improving wine aroma complexity. *Frontiers in Microbiology*, 7, <https://doi.org/10.3389/fmicb.2016.00411>
- Papalexandratou, Z., Falony, G., Romanens, E., Jimenez, J. C., Amores, F., Daniel, H. M., & De Vuyst, L. (2011). Species diversity, community dynamics, and metabolite kinetics of the microbiota associated with traditional ecuadorian spontaneous cocoa bean fermentations. *Applied and Environmental Microbiology*, 77(21), 7698–7714. <https://doi.org/10.1128/AEM.05523-11>
- Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.-M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., ... Schacherer, J. (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*, 556(7701), 339–344. <https://doi.org/10.1038/s41586-018-0030-5>
- Pontes, A., Čadež, N., Gonçalves, P., & Sampaio, J. P. (2019). A quasidegenerate relic hybrid population of *Saccharomyces cerevisiae* × *S. paradoxus* adapted to olive brine. *Frontiers in Genetics*, 10, 449. <https://doi.org/10.3389/fgene.2019.00449>
- Pontes, A., Hutzler, M., Brito, P. H., & Sampaio, J. P. (2020). Revisiting the taxonomic synonyms and populations of *Saccharomyces cerevisiae* – Phylogeny, phenotypes, ecology and domestication. *Microorganisms*, 8, 903. <https://doi.org/10.3390/microorganisms8060903>
- Poplin, R., Ruano-Rubio, V., DePristo, M., Fennell, T., Carneiro, M., Van der Auwera, G., Kling, D., Gauthier, L., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M., Neale, B., MacArthur, D., & Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178, <https://doi.org/10.1101/201178>
- Proux-Wéra, E., Armisen, D., Byrne, K. P., & Wolfe, K. H. (2012). A pipeline for automated annotation of yeast genome sequences by a

- conserved-synteny approach. *BMC Bioinformatics*, 13(1), 1–12. <https://doi.org/10.1186/1471-2105-13-237>
- R Development Core Team (2019). *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.r-project.org/>
- Renault, P., Miot-Sertier, C., Marullo, P., Hernández-Orte, P., Lagarrigue, L., Lonvaud-Funel, A., & Bely, M. (2009). Genetic characterization and phenotypic variability in *Torulaspora delbrueckii* species: Potential applications in the wine industry. *International Journal of Food Microbiology*, 134(3), 201–210. <https://doi.org/10.1016/j.ijfoodmicro.2009.06.008>
- Roach, M. J., & Borneman, A. R. (2020). New genome assemblies reveal patterns of domestication and adaptation across *Brettanomyces* (*Dekkera*) species. *BMC Genomics*, 21(1), 1–14. <https://doi.org/10.1186/S12864-020-6595-Z>
- Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6), 1117–1123. <https://doi.org/10.1101/gr.089532.108>
- Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3), 693–702. <https://doi.org/10.1534/genetics.113.154138>
- Steensels, J., Gallone, B., Voordeckers, K., & Verstrepen, K. J. (2019). Domestication of industrial microbes. *Current Biology*, 29, R381–R393. <https://doi.org/10.1016/j.cub.2019.04.025>
- Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., MargaritRos, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., ... Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39), 13950–13955. <https://doi.org/10.1073/pnas.0506758102>
- van Breda, V., Jolly, N., & van Wyk, J. (2013). Characterisation of commercial and natural *Torulaspora delbrueckii* wine yeast strains. *International Journal of Food Microbiology*, 163(2–3), 80–88. <https://doi.org/10.1016/j.ijfoodmicro.2013.02.011>
- Varela, J. A., Puricelli, M., Ortiz-Merino, R. A., Giacomobono, R., Braun-Galleani, S., Wolfe, K. H., & Morrissey, J. P. (2019). Origin of lactose fermentation in *Kluyveromyces lactis* by interspecies transfer of a neo-functionalized gene cluster during domestication. *Current Biology*, 29(24), 4284–4290.e2. <https://doi.org/10.1016/j.CUB.2019.10.044>
- Venkatesh, A., Murray, A. L., Coughlan, A. Y., & Wolfe, K. H. (2020). Giant GAL gene clusters for the melibiose-galactose pathway in *Torulaspora*. *Yeast*, 38(1), yea.3532. <https://doi.org/10.1002/yea.3532>
- Viigand, K., Põšnograjeva, K., Visnapuu, T., & Alamäe, T. (2018). Genome mining of non-conventional yeasts: Search and analysis of MAL clusters and proteins. *Genes*, 9(7), 354. <https://doi.org/10.3390/genes9070354>
- Villalba, M. L., Susana Sáez, J., del Monaco, S., Lopes, C. A., & Sangorrín, M. P. (2016). TdKT, a new killer toxin produced by *Torulaspora delbrueckii* effective against wine spoilage yeasts. *International Journal of Food Microbiology*, 217, 94–100. <https://doi.org/10.1016/j.ijfoodmicro.2015.10.006>
- Vinuesa, P., Ochoa-Sánchez, L., & Contreras-Moreira, B. (2018). GET_PHYLOMARKERS, a Software package to select optimal orthologous clusters for phylogenomics and inferring pan-genome phylogenies, used for a critical geno-taxonomic revision of the genus *Stenotrophomonas*. *Frontiers in Microbiology*, 9, <https://doi.org/10.3389/FMICB.2018.00771>
- Voordeckers, K., Brown, C. A., Vanneste, K., van der Zande, E., Voet, A., Maere, S., & Verstrepen, K. J. (2012). Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biology*, 10(12), e1001446. <https://doi.org/10.1371/journal.pbio.1001446>
- Welthagen, J. (1998). Yeast profile in Gouda cheese during processing and ripening. *International Journal of Food Microbiology*, 41(3), 185–194. [https://doi.org/10.1016/S0168-1605\(98\)00042-7](https://doi.org/10.1016/S0168-1605(98)00042-7)
- Will, J. L., Kim, H. S., Clarke, J., Painter, J. C., Fay, J. C., & Gasch, A. P. (2010). Incipient balancing selection through adaptive loss of aquaporins in natural *Saccharomyces cerevisiae* populations. *PLoS Genetics*, 6(4), e1000893. <https://doi.org/10.1371/journal.pgen.1000893>
- Winzeler, E. A., Castillo-Davis, C. I., Oshiro, G., Liang, D., Richards, D. R., Zhou, Y., & Hartl, D. L. (2003). Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics*, 163(1), 79–89. <https://doi.org/10.1093/GENETICS/163.1.79>
- Yamamoto, Y., Nakayama, A., Yamamoto, Y., & Tabata, S. (2004). Val216 decides the substrate specificity of alpha-glucosidase in *Saccharomyces cerevisiae*. *European Journal of Biochemistry*, 271(16), 3414–3420. <https://doi.org/10.1111/J.1432-1033.2004.04276.X>
- 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 van Leeuwenhoek*, 110(10), 1281–1286. <https://doi.org/10.1007/S10482-017-0844-4>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Silva, M., Pontes, A., Franco-Duarte, R., Soares, P., Sampaio, J. P., Sousa, M. J., & Brito, P. H. (2023). A glimpse at an early stage of microbe domestication revealed in the variable genome of *Torulaspora delbrueckii*, an emergent industrial yeast. *Molecular Ecology*, 32, 2396–2412. <https://doi.org/10.1111/mec.16428>