Studies in Mycology 111: 1–17 (2025).

WESTERDIJK FUNGALBIO DIVERSITY INSTITUTE

Phylogenomics and functional annotation of 530 non-*Saccharomyces* yeasts from winemaking environments reveals their fermentome and flavorome

R. Franco-Duarte[1,2*], T. Fernandes[1,2], M.J. Sousa[1,2], P. Sampaio[1,2], T. Rito[1,2], P. Soares[1,2]

[1]*CBMA (Centre of Molecular and Environmental Biology), Department of Biology, University of Minho, Braga, Portugal;*
[2]*Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, 4710-057 Braga, Portugal.*

Corresponding author: R. Franco-Duarte, ricardofilipeduarte@bio.uminho.pt, ricardofrancoduarte@gmail.com

Abstract: The winemaking industry faces unprecedented challenges due to climate change and market shifts, with profound commercial and socioeconomic repercussions. In response, non-*Saccharomyces* yeasts have gained attention for their potential to both mitigate these challenges and enhance the complexity of winemaking. This study builds upon our previous cataloguing of 293 non-*Saccharomyces* yeast species associated with winemaking environments by rigorously analysing 661 publicly available genomes. By employing a bioinformatics pipeline with stringent quality control checkpoints, we annotated and evaluated these genomes, culminating in a robust dataset of 530 non-*Saccharomyces* proteomes, belonging to 134 species, accessible to the research community. Employing this dataset, we conducted a comparative phylogenomic analysis to decipher metabolic networks related to fermentation capacity and flavor/aroma modulation. Our functional annotation has uncovered distinctive metabolic traits of non-*Saccharomyces* yeasts, elucidating their unique contributions to enology. Crucially, this work pioneers the identification of a non-*Saccharomyces* 'fermentome', a specific set of six genes uniquely present in fermentative species and absent in non-fermentative ones, and an expanded set of 35 genes constituting the complete fermentome. Moreover, we delineated a 'flavorome' by examining 96 genes across 19 metabolic categories implicated in wine aroma and flavour enhancement. These discoveries provide valuable genomic insights, offering new avenues for innovative winemaking practices and research.

Key words: bioinformatics, fermentation, fungi, genomics, non-conventional yeasts, phylogeny.

Citation: Franco-Duarte R, Fernandes T, Sousa MJ, Sampaio P, Rito T, Soares P (2025). Phylogenomics and functional annotation of 530 non-*Saccharomyces* yeasts from winemaking environments reveals their fermentome and flavorome. *Studies in Mycology* **111**: 1–17. doi: 10.3114/sim.2025.111.01
Received: 4 March 2024; **Accepted:** 10 October 2024; **Effectively published online:** 19 February 2025
Corresponding editor: J. Houbraken

Studies in Mycology

INTRODUCTION

Winemaking is an activity with great economic value where yeasts play a key role regarding wine's sensorial features. Nowadays, a global trend is evident towards new and differentiated wines, but, simultaneously, aiming to keep the typicity and quality of the obtained products. However, wine regions are facing harsh challenges, mainly due to climate change and market evolution, with commercial and socioeconomic impacts. In the last decades, several approaches were developed to address these challenges in winemaking, and non-conventional yeasts have emerged as key components in these strategies (Mira de Orduña 2010). As complex entities, yeasts are selected according to multiple criteria in order to ensure control of the winemaking process. *Saccharomyces* continues to, nevertheless, stand as the main genus explored to perform alcoholic fermentation as it displays suitable adaptation to the high sugar content of grape must and consequently converts it into a panoply of diverse organoleptic compounds. Within this genus, *Saccharomyces cerevisiae* stands out as the yeast of excellence and is often called the wine yeast. The global phylogeny of *S. cerevisiae* strains shows particular clusters indicative of their geographical origin and fermentative capacities (Liti *et al.* 2009, Franco-Duarte *et al.* 2014, Legras *et al.* 2018). Additionally, recent research demonstrated the remarkable plasticity of the *S. cerevisiae* genome, highlighting either the fast adaptation of strains to many ecological niches, such as the ones associated with the human hosts (Legras *et al.* 2018). Additionally, our previous work (Franco-Duarte *et al.* 2015, 2016, 2017) identified specific pheno-metabolomic traits linked to *S. cerevisiae* adaptation.

Non-*Saccharomyces* yeasts, referred to as the yeast species found in wine production and other biotechnological relevant areas, excluding *S. cerevisiae* (Jolly *et al.* 2014), were considered, for many years, as spoilage microorganisms and sources of contamination due to the high levels of volatile acidity detected. However, in recent years, they have been broadly recognized as enhancing positively the complexity of the organoleptic profile of wines. This group of yeasts is now considered as an attractive biotechnological host, since these yeasts offer several advantages in wine production, including the release of primary aromatic compounds through hydrolysis by glycosidases, unique to these yeasts (Maicas & Mateo 2023); the addition of novel sensorial and aromatic traits to the matrix of beverages by producing esters, higher alcohols, and volatile thiols; and increased glycerol content while reducing acetic acid and ethanol production. Additionally, non-*Saccharomyces* yeasts secrete esterases, β-glucosidases, lipases, proteases and other compounds responsible for the enrichment of wine's aroma, decrease the spoilage risk, and complement *S. cerevisiae* fermentative performance (Ellis *et al.* 2022, Gunhan 2022). Therefore, these non-conventional yeasts display a great biotechnological potential, as evidenced by the commercialization of several non-*Saccharomyces* starter yeasts (Fernandes *et al.*

2021, Ianieva & Podgorsky 2021, Raymond Eder & Rosa 2021).

In the last years, an extensive debate addressed the concept of wine microbiome and, from this, meaningful results on the microbial complexity of the fermentation process were obtained (Barata *et al.* 2012, Bokulich *et al.* 2013, Liu *et al.* 2017). If we consider the wide variety of yeast species identified in those studies, it urges to know and understand their role and the correspondent inter and intra-species variation. The first utilization of a non-*Saccharomyces* strain in wine fermentation is thought to have occurred in the 1950s with the use of a *Torulaspora delbrueckii* strain (Cantarelli 1955), resulting in a significant reduction of the volatile acidity of the finished wine. Our research group has placed a major focus on the study of *T. delbrueckii* strains, mainly due to the incredible biochemical, genomic and phenotypic features of this non-*Saccharomyces* species (Fernandes *et al.* 2021, Santiago *et al.* 2021, Silva-Sousa *et al.* 2022, Silva *et al.* 2022). Additionally, other species were shown to have a particular significant impact on the sensory, chemical, and microbial properties of the final product. These include *Schizosaccharomyces pombe*, *Wickerhamomyces anomalus*, *Metschnikowia pulcherrima*, *Lachancea thermotolerans*, *Starmerella stellata*, *Pichia kluyveri*, and *Hanseniaspora spp*., as reviewed by Maicas and Mateo (2023).

In 2021, we presented, for the first time, an in-depth comprehensive assessment of 231 scientific papers published during the previous 80 years, reporting identification of non-*Saccharomyces* yeasts from grapes and/or grape must (Drumonde-Neves *et al.* 2021). In that work, we established a catalogue of 293 yeast species, with links to their respective isolation environment, geographical occurrence, grapevine cultivar, health status of the grapes, and also associating each species to the accepted taxonomic names and correspondent synonyms. From this, 58 % of the species were associated with more than one occurrence, with eight species being isolated in more than 50 isolation campaigns: *Hanseniaspora uvarum, Metschnikowia pulcherrima, Starmerella bacillaris, Pichia terricola, Torulaspora delbrueckii, Pichia kudriavzevii, Hanseniaspora guilliermondii* and *Lachancea thermotolerans*. The collected database constitutes an important landmark for researchers working with non-conventional yeasts, and a path was drawn to start clearing their full contribution to winemaking (isolated or in a consortium), and to shed light into a better understanding of the wine microbiome. However, even though microbial dynamics in natural fermentations have been studied in great detail, the majority of the reports are mainly descriptive, without focusing much on microbial interactions and in the study of the metabolic implications of each microorganism in a particular fermentation. Since we now know that those interactions impact greatly on the quality of the wine (Kosel *et al.* 2017, Liu *et al.* 2017), and that the production of specific metabolites depends on the species (and strains) present, it is time to change our focus to the study of yeasts' community dynamics and biochemical activities.

In an attempt to identify the genes required by yeasts, not only to grow, but to complete fermentation of a high sugar medium, Walker *et al.* (2014) defined the fermentome of *S. cerevisiae*, as consisting of a set of 93 genes whose deletion resulted in a fermentation process that were at least 20 % longer compared to the wild type strain. Authors defined this set of genes as crucial for understanding the mechanisms behind *S. cerevisiae* strains´ response and adaptation to stresses imposed during high-sugar fermentations. However, to date, no extension of this work to non-*Saccharomyces* strains has been performed. Equally, Belda *et al.* (2016) introduced the concept of wine yeasts flavorome, defining an enzymatic basis of this concept through the analysis of the production of enzymes of oenological interest across 15 yeast species, and establishing links for their phylogenetic placement. However, broader research of the wine non-*Saccharomyces* yeasts and of the extensive metabolic networks involved in enzymes production was lacking.

A great number of important works have been published in the last years, trying to finely detail the diversity and taxonomy of some of the non-conventional yeasts. Shen *et al.* (2018) analysed the genomes of 332 yeast species, scanning about one third of the known budding yeasts diversity at the time, and showed a phylogenetic division of the subphylum *Saccharomycotina* into 12 major clades. In a follow up study, Groenewald *et al.* (2023) officialised this new vision of this subphylum by proposing a new higher rank classification by the incorporation of taxonomic novelties: 7 new classes – *Saccharomycetes*, *Pichiomycetes*, *Sporopachydermiomycetes*, *Alloascoideomycetes*, *Dipodascomycetes*, *Trigonopsidomycetes*, and *Lipomycetes*; 12 new orders: *Saccharomycetales*, *Saccharomycodales*, *Phaffomycetales*, *Ascoideales*, *Serinales*, *Alaninales*, *Pichiales*, *Sporopachydermiales*, *Alloascoideales*, *Dipodascales*, *Trigonopsidales*, and *Lipomycetales*. Opulente *et al.* (2023) recently sequenced 953 additional strains of this subphylum, confirming the spread of the 1049 analysed species through the 12 proposed orders.

In the present work, we considered the set of 293 non-*Saccharomyces* yeast species present in grapes and/or grape musts identified in our previous study (Drumonde-Neves *et al.* 2021), and obtained 661 publicly available genomes. These species belong to two phyla (*Ascomycota* and *Basidiomycota*) and 5 subphyla, and are distributed across 9 classes, 16 orders and 22 families. This comprehensive genomic dataset allowed us to perform a comparative study unravelling metabolic networks related with fermentation capacity and flavour/aroma production. This study also aimed to identify key differences between these yeasts and *S. cerevisiae*. Through this analysis, we gained a deeper understanding of the unique metabolic capabilities and potential contributions of non-*Saccharomyces* yeasts in shaping the fermentation process and imparting flavour and aroma characteristics to wine, and to establish, for the first time, a putative core fermentome and core flavorome of non-*Saccharomyces* encountered in the wine microbiome.

MATERIALS AND METHODS

Dataset

From the list of 293 non-*Saccharomyces* species published in our previous work (Drumonde-Neves *et al.* 2021), we first excluded the ones clearly associated with human infections – *Candida albicans, C. glabrata, C. metapsilosis, C. orthopsilosis, C. parapsilosis, C. tropicalis* and *Cryptococcus neoformans* –, obtaining a final set of 286 species. Using this list, we searched two NCBI databases – Assembled Genomes (https://www.ncbi.nlm.nih.gov/genome/) and SRA (https://www.ncbi.nlm.nih.gov/sra) –, for the public availability of complete genomes for these species (April 2022). Genomes for all the strains of the considered species were downloaded, and when more than one deposit was available for the same strain, all of them were considered, being data filtered later using quality control parameters. In this way, a final set of 661 genomes was considered (Table S1) and analysed using the pipeline described in Fig. 1. Supplementary Tables have been deposited in FigShare (doi: 10.6084/m9.figshare.27208080).

Genome assembly and annotation

For the genomes downloaded from SRA database (raw data; without assembled version available), assembly was performed using SPAdes Genome Assembler software v. 3.15.4 (Bankevich *et al.* 2012), using default parameters. In total, 661 assembled genomes were annotated using AUGUSTUS software v.3.4.0 (Stanke & Morgenstern 2005), considering 16 different pre-trained models, chosen as belonging to the *Ascomycota* phyla (11) or the *Basidiomycota* phyla (5): *Ascomycota* – *S. cerevisiae* S288c, *C. albicans*, *Meyerozyma (Candida) guilliermondii*, *C. tropicalis*, *Debaryomyces hansenii*, *Eremothecium gossypii*, *Kluyveromyces lactis*, *Lodderomyces elongisporus*, *Scheffersomyces (Pichia) stipitis*, *Schizosaccharomyces pombe*, and *Yarrowia lipolytica*; *Basidiomycota* – *Cryptococcus neoformans*, *Coprinus*, *Laccaria bicolor*, *Phanerochaete chrysosporium* and *Ustilago maydis*. Results were manually reviewed to select the most robust annotation in terms of predicted coding genes. The potential coding regions reported by AUGUSTUS were extracted from the complete genomes to FASTA files.

Quality control

As quality control is a key part of this work, and to avoid later incongruencies associated with lack of genome quality and/or the use of different software, careful quality control checkpoints were always implemented throughout the entire pipeline while analysing the dataset. After obtaining the 661 genomes and the consensus annotated proteomes, three methods were employed to filter out genomes lacking quality, robustness and completeness (Fig. 1, Table S1):

i) Assessment of the genomes' completeness was performed with BUSCO (Benchmarking Universal Single-Copy Orthologs) software, v. 5.0 (Simão *et al.* 2015), using four different datasets: *Fungi_odb10* (composed by 758 genes), *Fungi_odb9* (250 genes), *Ascomycota_odb10* (1706 genes) and *Basidiomycota_odb10* (1764 genes). *Fungi_odb* databases were applied to all the genomes considered in this study. However, the datasets associated with the specific phyla were only applied to the genomes belonging to that taxonomic level;

ii) Genome assembly quality metrics were computed using QUAST v. 5.0.2 (Mikheenko *et al.* 2018);

iii) Internal quality control was also assessed using BLAST homology search: we used the reference proteome of *S. cerevisiae* S288c (NCBI accession number GCF_000146045.2) to perform pairwise comparisons using BLASTP and an e-value cutoff of 10^{-5} and a minimal query coverage of 15 % with all the proteins in our dataset. We detected which protein homologs were present in at least 95 % of our proteomes (201 in total). These represented proteins related with basal features that should be present in all the organisms. Following the definition of this set of proteins, we analysed the results considering the missing proportion of this set of genes in each individual genome. As these missing proteins should be present independently of the specific taxonomic groups, their absence likely indicates issues related with assembly and annotation quality. Multiple missing genes are indicative of general incompleteness.

Using this pipeline, from the initially considered 661 genomes, we excluded 131 genomes that failed the different thresholds. In general, the three quality checkpoints indicate quality issues within the same genomes (Table S1).

Homology analysis and Phylogenomics

A consensus proteome database was prepared by considering the 530 complete genomes (from 134 species) that passed the quality control. As in the quality control, BLASTP analysis was performed using the full proteome of the *S. cerevisiae* S288c as a query against the total database. Again, we used an e-value cutoff of 10^{-5}. The homologs detected in the 530 proteomes were selected as representative of the core genome of the analysed species. The effect of the quality control performed before was clear as the number of conserved proteins obtained increased to 1 131 proteins. Each homologous protein set were multiple-aligned using the MAFFT

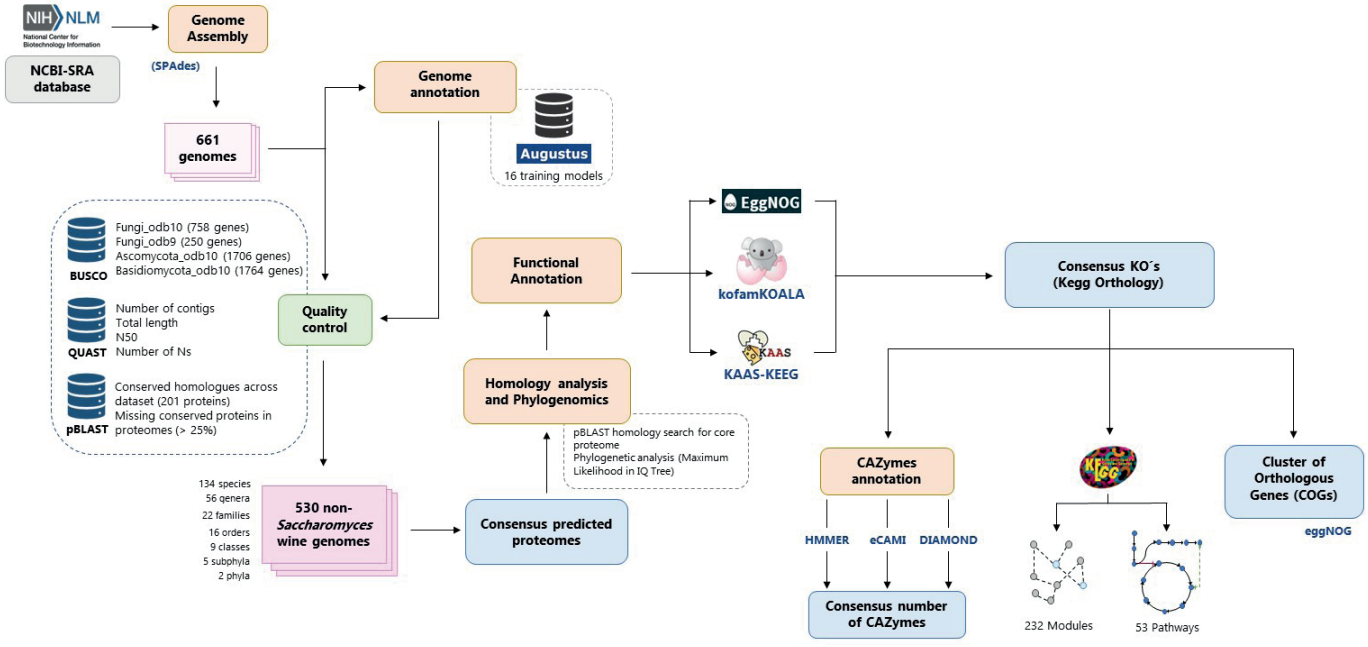

Fig. 1. Bioinformatic pipeline used to analyze the 661 genomes of non-*Saccharomyces* yeasts. The schematic representation of the pipeline includes all the quality control checkpoints used throughout the work, together with the software and the intermediate datasets obtained. For details about each software and parameters, see the Methods section.

algorithm in FasParser (https://github.com/Sun-Yanbo/FasParser) (Katoh *et al.* 2018) and the alignments were concatenated into a single long multiple alignment containing the conserved aligned proteome. We reconstructed the evolutionary relationship between the analyzed species/strains using maximum likelihood analysis in IQ-TREE (www.iqtree.org) (Nguyen *et al.* 2015), with the JTT model of amino acid evolution and gamma-distributed rates (four rates) with 10 000 bootstrap replicates. One outgroup, *Rhizophagus irregularis*, outside the groups *Ascomycota* and *Basidiomycota* that included all the species analysed in this work, was included. Phylogenetic tree was visualized using FigTree v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) and annotated using Interactive Tree Of Life (iTOL) software v. 6 (Letunic & Bork 2019).

Functional annotation

Functional genomic annotation was performed using three tools, for increased robustness, as being the three most used tools available for functional annotation: *i)* eggNOG-mapper v.5.0 (Jensen *et al.* 2008); *ii)* kofamKOALA v. 2022-04-03 (Aramaki *et al.* 2020); *iii*) KAAS-KEEG Automatic Annotation Server (Moriya *et al.* 2007). From each software, a list of annotated KO´s (KEEG Orthology), as representing the functional orthologs, was obtained, and an in-house script was built in Python to obtain the list of Consensus KO´s for each genome. The script would search for a consensus annotation in a way that: *(i)* if only a platform would provide a KO, that one would be considered; *(ii)* if two platforms would provide a KO and the third a different one, the KO that appeared in the two matching results was selected as consensus; *(iii)* if three different KO´s were obtained (or two different KO´s and a third with no result), it would be considered undefined. The results were then analyzed using KEGG Mapper-Reconstruct Pathway tool (Kanehisa and Sato 2020), allowing to perform the KO-based mapping against KEGG databases (Pathways and Modules), and to visualize reconstructed global maps of metabolic pathways. Results were also described considering Gene Ontology (GO) terms and clusters of orthologous groups (COGs) with their associated functional categories. Gene function predictions were also accomplished by assessing the Carbohydrate-Active EnZymes (CAZymes) database (Cantarel *et al.* 2009), using dbCAN2 software (Zhang *et al.* 2018), testing three different annotation tools to increase robustness, HMMER, eCAMI and DIAMOND, and compiling results to obtain the consensus number of CAZymes.

Fermentome and flavorome of non-*Saccharomyces* yeasts

To establish the fermentome of the non-*Saccharomyces* dataset, we scrutinized our BLASTP results to analyze if a particular gene was present or absent in the genome of a particular strain, separating yeasts according to their fermentative capacity as reported by Kurtzman *et al.* (2010). To unravel the flavorome of non-*Saccharomyces* yeasts, as being the set of genes related with the particular aroma/flavor characteristics that these yeasts bring to wine, a list of genes was compiled from relevant literature, but mainly from Martin *et al.* (2016), Godoy *et al.* (2021), Morata *et al.* (2020), Kutyna and Borneman (2018), Dzialo *et al.* (2017), Belda *et al.* (2017), Giorello *et al.* (2019), Mendes *et al.* (2017), Verstrepen *et al.* (2003) and Padilla *et al.* (2016). Genes were divided into categories, according to their metabolic function, and their presence was searched within the genomes of the yeasts catalogued previously as fermentative, using the results of BLASTP analysis.

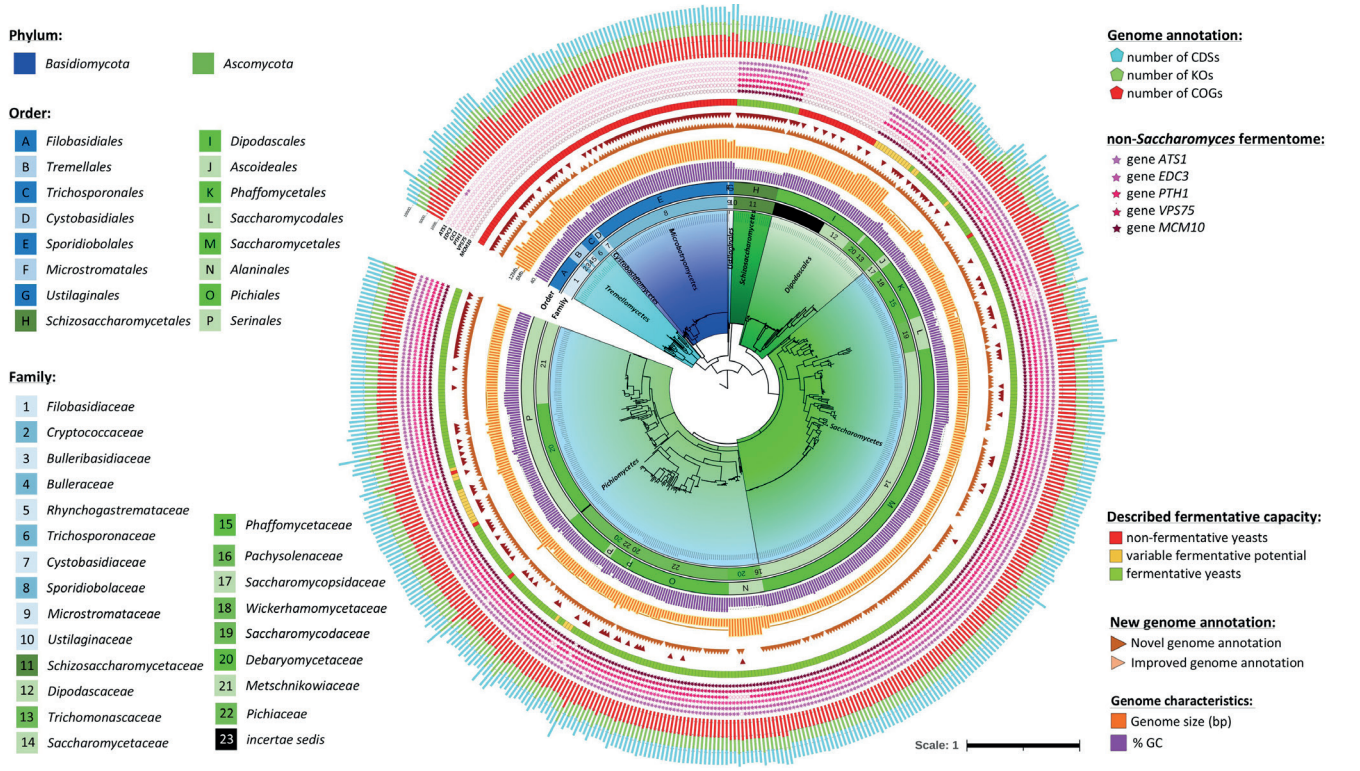

Fig. 2. Detailed phylogeny of the 530 genomes belonging to 293 non-*Saccharomyces* yeast species. Phylogeny was obtained after the alignment of 1209 core proteins, and maximum likelihood analysis with the JTT model of amino acid evolution and gamma-distributed rates (four rates) with 10 000 bootstrap replicates. Bootstrap values were omitted from the branches due to being > 99 %. *Rhizophagus irregularis* was used as outgroup to root the tree. Branches were coloured and annotated according to the taxonomic classification available in Supplementary Table S2, and the remaining annotations/colourings were done using results of genomes annotation (see methods for details).

Data analysis and visualization

Taxonomic information of the non-*Saccharomyces* considered in this study was collected from Goenewald *et al.* (2023), for species within *Saccharomycotina*, as being the currently accepted taxonomy for this subphylum (https://theyeasts.org/), and from NCBI-Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy), MycoBank (https://www.mycobank.org/) and IndexFungorum (http://www.indexfungorum.org/) for the remaining species. Discrepancies between databases were identified (Table S2), and the most concordant identification was used, with the exception of the *Saccharomycotina* yeasts.

Data was analyzed and visualized using the statistical analysis software MINITAB v. 19.2020, and Orange Data Mining v. 3.36.2 (Demšar *et al.* 2004). All scripts described in the previous sections were written in Python v. 3.10. Statistical analysis between taxonomic groups in terms of presence/absence of groups of genes was performed using Kruskal-Wallis H-test, in Python, by being the most suitable test for comparing multiple groups without assuming a normal distribution of data.

Data availability

All data associated with this study are provided in the main article, in the supplementary materials or on a public repository (DOI: 10.34622/datarepositorium/WPHMJL). In particular, the de-novo assembled genomes of the strains without assembly previously available, the annotation of the assembled genomes (in .gff format), the predicted proteomes of the 530 annotated genomes, the consensus KO´s (KEGG Orthologous) obtained after pairing of the three annotation software, and the list of 1131 consensus proteins used throughout the study are available in the repository. The detailed version of the phylogenetic tree shown in Fig. 2, including all the annotations and node support, are available on https://itol.embl.de/shared/2UnEoNdSx8Yui.

RESULTS AND DISCUSSION

Non-*Saccharomyces* wine yeasts database: quality control and genome annotation

In the present work, we searched public genomic databases for the 293 yeast species identified in the previous work as being associated with wine and vine, and collected 661 genomes, belonging to 178 species (Table S1). This corresponds to approximately 61 % of the 293 identified species, as no genome was available for the remaining species at the time of this work (April 2022). The subsequent step in our pipeline involved annotating the genomes of the 661 species. Given that this is a critical step where results can significantly vary between projects, mainly due to the use of different models to train the algorithm, we employed 16 training models to obtain the most robust set of annotated protein-coding genes. The best results were manually chosen to enable precise and accurate comparisons between the genomes, selecting annotations yielding a higher number of protein-coding genes. Although this approach, that doesn't consider a minimal length to consider the annotated proteins, might introduce false positives, it also facilitates the detection of microproteins, which have been identified as significantly important in *S. cerevisiae* (Sun *et al.* 2022). Furthermore, any false positives introduced at this stage are likely to be filtered out during functional analysis, thus minimizing their impact on subsequent analysis. As we collected data for the 661 genomes from various sequencing platforms, each with different depths and robustness, we implemented a meticulous quality-control pipeline. This involved combining different software and approaches to assess genome quality (Fig. 1, Table S1), aiming to prevent incongruencies in later analyses.

First, each assembled genome was analysed using BUSCO software against different databases, and low-quality genomes were filtered out, corresponding to the ones containing less than 80 % of the total single-copy BUSCO genes in each database: *fungi_odb10* - 607 genes; *fungi_odb9* - 290 genes; *ascomycota_odb10* - 1 365 genes; *basidiomycota_odb10* - 1 412 genes. After, BLAST homology search was performed to include an additional internal quality control checkpoint, using the reference proteome of *S. cerevisiae* S288c to perform pairwise comparisons, which not only offered an opportunity to test the quality of the assemblies but also the display of full annotatable genes. A core set of 201 conserved proteins was obtained, and genomes/proteomes showing less than 150 of the conserved homologous proteins across the dataset were excluded. Finally, QUAST results were used in the end to manually check for the genomes passing the previous filters, and for the ones leading to incongruent results between the different used parameters.

After filtering out the genomes failing the previous quality checkpoints, as lacking quality, robustness and completeness (Table S1), a final set of 530 non-*Saccharomyces* genomes/proteomes was considered (average N50 = 873,214; min/max N50 = 1,492/5,629,136; Average number of nucleotides = 13.7 Mb; min/max number of nucleotides = 7.2 Mb/30 Mb; Average %GC = 44.4 %; min/max %GC = 26.7 % / 68.0 %), available to be used by the scientific community (Figshare repository). These genomes belong to 134 yeast species (~46 % of the 293 yeast species previously found to be associated to wine and vine) – Table S2. This consensus set of proteomes constitutes the most complete dataset of wine non-*Saccharomyces* yeast genomes/proteomes to date, compiling data with great interest for further use in winemaking studies or in other areas of biotechnological relevance. These advances were possible due to the pipeline used, that maximizes quality control through a novel combination of methods, allowing increased robustness.

We also annotated for the first time the genome of 208 yeast strains, belonging to 69 species that never had published genome annotation before, to the extent of our knowledge (Table S3) and at the time of this study (April 2022). Additionally, we improved the genome annotation of 264 additional strains (from 91 species), in comparison with the annotation available at the moment. The best annotation results of *Ascomycota* yeasts were obtained using as training model the genome of *Lodderomyces elongisporus*, which has previously provided good results for the annotation of other Ascomycetous yeasts (Franco-Duarte *et al.* 2022). *Basidiomycota* yeasts' annotation showed the best results when the genome of *Cryptococcus neoformans* was used. The obtained average number of protein coding sequences per species varied between 4138 and 9686 proteins throughout the entire dataset, with an average number of 6160 proteins per species. Significant differences (unpaired t-test; $p<0.01$) were found between phyla: *Ascomycota* - 5856 coding regions per species on average; *Basidiomycota* - 7365 coding regions per species on average.

While we successfully analysed approximately 46 % of the 293 yeast species previously identified in association with wine and vine, continuous efforts are still necessary to enhance this percentage in the future. Recent large-scale sequencing projects by Shen *et al.* (2018), Li *et al.* (2021), Groenewald *et al.* (2023) and Opulente *et al.* (2023), among others, have significantly contributed to unraveling yeast diversity, particularly within *Ascomycota*, which will allow to greatly increase this value. However, our knowledge about Basidiomycota

yeasts remains limited. In the current study, 54 species in *Basidiomycota* had no available genome, and 24 species, despite having genomes in the databases, did not pass the quality control. This accounts for a total of 26 % (78 out of the 293) of species that could not be considered in this work.

Genome-scale robust phylogeny of non-*Saccharomyces* yeasts associated with winemaking environments

To establish the phylogenomics of the non-*Saccharomyces* wine yeasts, a proteome database was first compiled by considering the 3221 923 predicted protein coding sequences from the 530 genomes. The complete proteome of *S. cerevisiae* S288c was used in a BLASTP analysis against the database, and a total of 1131 conserved proteins was selected as the core proteome of the analysed isolates (shared in Figshare repository). The phylogenetic tree presented in Fig. 2 represents the alignment of the core concatenated proteins, and the reconstruction of strains' evolutionary relationships using maximum likelihood analysis.

As anticipated, a complete separation was achieved between *Ascomycota* and *Basidiomycota* isolates, indicated by green and blue branches in the tree (Fig. 2), respectively. Within the phylum *Basidiomycota*, our analysis distinctly segregated strains from 10 families, spanning 7 orders and 5 classes. Despite the relatively low number of isolates within the classes *Cystobasidiomycetes*, *Ustilaginomycetes*, and *Exobasidiomycetes*, we achieved a clear separation from other classes, emphasizing the necessity of obtaining additional high-quality genomes from these groups.

In the *Ascomycota*, a complete phylogenetic separation was observed when considering the four classes: *Schizosaccharomycetes*, *Saccharomycetes*, *Pichiomycetes*, and *Dipodascomycetes*. Even though the latter three classes were previously considered a single taxon – *Saccharomycetes* -, the work of Groenewald *et al.* (2023) proposed the division of this class. Our results align with this classification, demonstrating an almost complete phylogenetic separation between *Saccharomycetes*, *Pichiomycetes*, and *Dipodascomycetes*, as suggested by Groenewald *et al.* Furthermore, this separation extends to families and orders, also in line with their suggestion. One significant exception warrants further consideration in the future: isolates from the family *Debaryomycetaceae* (number 20 in Fig. 2), order *Serinales* (P in Fig. 2), were found to be distributed across different classes, particularly within *Dipodascales* – one *Candida incommunis* isolate – and within *Pichiomycetes*. However, even within the latter, they did not form an isolated cluster but rather mixed with species from the family *Pichiaceae*.

The genome size (represented by orange squares) and %GC content (depicted by purple squares) exhibited great heterogeneity across the genomes. In many instances, these variations were associated with the yeasts' taxonomic groups, such as in the class *Schizosaccharomycetes*, which displayed noticeable differences compared to other taxa. The substantial heterogeneity observed within members of the remaining families of *Ascomycota* has been extensively discussed in previous studies, encompassing analyses of complete nuclear genomes (Shen *et al.* 2018, 2020), the mitogenomes (Vieira *et al.* 2020, Christinaki *et al.* 2022), specific parts as the telomers (Peska *et al.* 2021), and even regarding the analysis of the sugar transporters (Donzella *et al.* 2023). This genetic diversity, particularly within *Saccharomycetes*, has been primarily attributed to a genome duplication event (WGD) that occurred approximately 100 million years ago. This event led to the introduction of genetic rearrangements, gene losses, and mutations in the genomes of these yeasts, followed by subsequent evolution and adaptation (Kellis *et al.* 2004, Marcet-Houben & Gabaldón 2015).

Functional genome annotation: metabolic traits across non-*Saccharomyces* yeasts

To functionally annotate the predicted open reading frames of the 530 non-*Saccharomyces* genomes, we obtained the consensus list of KO´s (functional orthologs on KEGG Orthology) for each genome (Table S4), from three tools: eggNOG-mapper, kofamKOALA and KAAS-KEGG. This marks the first instance, to the best of our knowledge, where a successful combination of these three widely-used platforms for yeast functional genomics has been attempted. By individually analysing results of each software, one can appreciate the importance of this validation step, as the outcomes varied considerably depending on the platform applied. As an example of extreme variations, one can consider the case of *W. anomalus*, where eggNOG-mapper and KAAS-KEGG predicted 4 530 and 4 334 KO´s respectively, while kofamKOALA predicted only 1 124 functional orthologs. These substantial differences are extended to the remaining species of this genus, and generally kofamKOALA consistently produced a smaller number of hits compared to the other two platforms. Given the absence of a consensus on the best platforms to use and recognizing that each employs different databases that could hypothetically be more suitable for a particular taxonomic group, our method allowed us to integrate information from all three platforms, obtaining a consensus list of annotated KO´s (Table S4 and Figshare repository). Results showed variations between 2 780 (*Hanseniaspora uvarum*) and 4 976 (*Filobasidium floriforme*) KO´s per species, on average. Unlike the disparities observed in total coding sequences and genome sizes between *Ascomycota* and *Basidiomycota*, the average number of KO´s showed no significant differences: *Ascomycota* - 3 808, *Basidiomycota* - 3 887. This list was also used to derive clusters of orthologous groups (COGs) along with their associated functional categories, and to calculate the percentage of coding genes involved in the breakdown of complex carbohydrates (CAZymes). The results unveiled significant inter-species variability, allowing for the prediction of each species' metabolic potential. However, no distinct patterns could be established within taxonomic groups. For instance, *Starmerella magnoliae* demonstrated only 44.1 % of its predicted coding genes associated with a specific COG category (Table S4), a trend reflected also in its low association with KEGG pathways and modules (Table S5 and S6). This variability extended to the analysis of CAZymes, where only 1 % of the genes showed a connection with carbohydrate breakdown activity. A possible explanation for this observation could be the known association of this yeast species with hymenopterans, flowers, and specific species of bees (Neto & Morais 2020). Its presence in winemaking environments might be linked to a migration from such environments to the vine, potentially explaining its apparent low metabolic activity. This pattern could be extended to the remaining *Starmerella* species, that exhibited a lower percentage of genes associated with a CAZyme (< 1.5 %) or a COG category (< 61 %). However, the most plausible explanation arises from recent findings by Gonçalves *et al.* (2019, 2022), indicating that yeasts belonging to the *Wickerhamiella* and *Starmerella* genera (W/S clade) exhibit a distinctive evolutionary history marked by the loss and subsequent reinstatement of alcoholic fermentation mediated by horizontal gene transfer events. Our functional annotation results, indicative of a lower metabolic potential for *Starmerella* yeasts, align with the

unusual metabolic features reported by Gonçalves *et al.* These yeasts show a preference for fructose over glucose as a carbon source, a trait known as fructophily. Furthermore, the authors demonstrated that the gene *SUC2*, crucial for sucrose assimilation, is almost exclusively present in the *Starmerella* subclade, having been horizontally acquired from bacteria. Regarding *Wickerhamiella*, although only one strain was available in our dataset, results also showed a small percentage of genes associated with a specific COG category (57.9 %). While slightly higher than that of *Starmerella* species, this value corroborates the distinct strategies employed by yeasts of the W/S clade to adapt to fermentative environments. This observation aligns with the fact that this clade has, by far, the most horizontal gene transfer events reported, originating from bacteria (Gonçalves & Gonçalves 2022).

In contrast, a specific strain of *T. delbrueckii* – PYCC2999 – exhibited an impressive 97.1 % of KO´s annotated as functional categories. This high percentage underscores the species' metabolic potential, indicating a diverse array of functional genes associated with various metabolic processes. The consistency of these high values across all *T. delbrueckii* strains considered in this work reinforces the species' significance and versatility. This observation has been reiterated in multiple studies conducted by our research group (Fernandes *et al.* 2021, Santiago *et al.* 2021, Silva-Sousa *et al.* 2022, Silva *et al.* 2022), providing robust evidence for *T. delbrueckii* as one of the most important non-*Saccharomyces* yeasts. This is particularly evident in its broad metabolic potential across diverse environments, including those relevant to industrial applications such as winemaking. Another noteworthy result pertains to the species *Sporisorium graminicola*, where 3.3 % of the genes were annotated as CAZymes - the highest value obtained in this study. This percentage is primarily attributed to the cataloguing of 117 genes as Glycoside Hydrolases (GH), suggesting the species' potential to degrade biomass compounds, an aspect not previously assessed.

Next, to unravel the metabolic signature of each strain, our analysis delved into the study of the KEGG pathways associated with each gene. Upon reconstructing the metabolic pathways of the 530 proteomes using KEGG Mapper-Reconstruct, notable disparities emerged between the 134 species and their respective taxonomic groups (Table S5). On average, *Ascomycota* yeasts exhibited 6946 genes annotated to KEGG pathways per species (7054 for *Saccharomycotina* and 6839 for *Taphrinomycotina*), whereas *Basidiomycota* displayed a higher average with 7380 genes (7467, 7088 and 7585 for *Agaricomycotina*, *Pucciniomycotina* and *Ustilaginomycotina*, respectively). While every *Basidiomycota* family had over 7000 functional orthologs annotated to at least one KEGG pathway, certain Ascomycota species, particularly those from the *Taphrinomycotina* subphylum and class *Schizosaccharomycetales*, exhibited values below 7000.

Figure 3 represents a PCA visualization of these results, illustrating the distribution of all species colored based on their taxonomic classification. The two principal components represent the number of genes associated with each KEGG category, explaining 25.5 % of variability. A clear separation is evident between *Ascomycota* and *Basidiomycota*. However, no well-defined separation is observed when analyzing other taxonomic levels, although with interesting particularities.

Schizosaccharomycetales yeasts tightly cluster in the centre of the PCA, embedded within *Saccharomycetales* yeasts. Yeasts from the *Saccharomycetes* class exhibit a distinct metabolic profile, evident in both components, apart from the remaining classes *Pichiomycetes* and *Dipodascomycetes*, thus supporting the findings of Groenewald *et al.* (2023). Within this class, *Saccharomycetales* and *Saccharomycodales* form a cluster, exhibiting almost complete

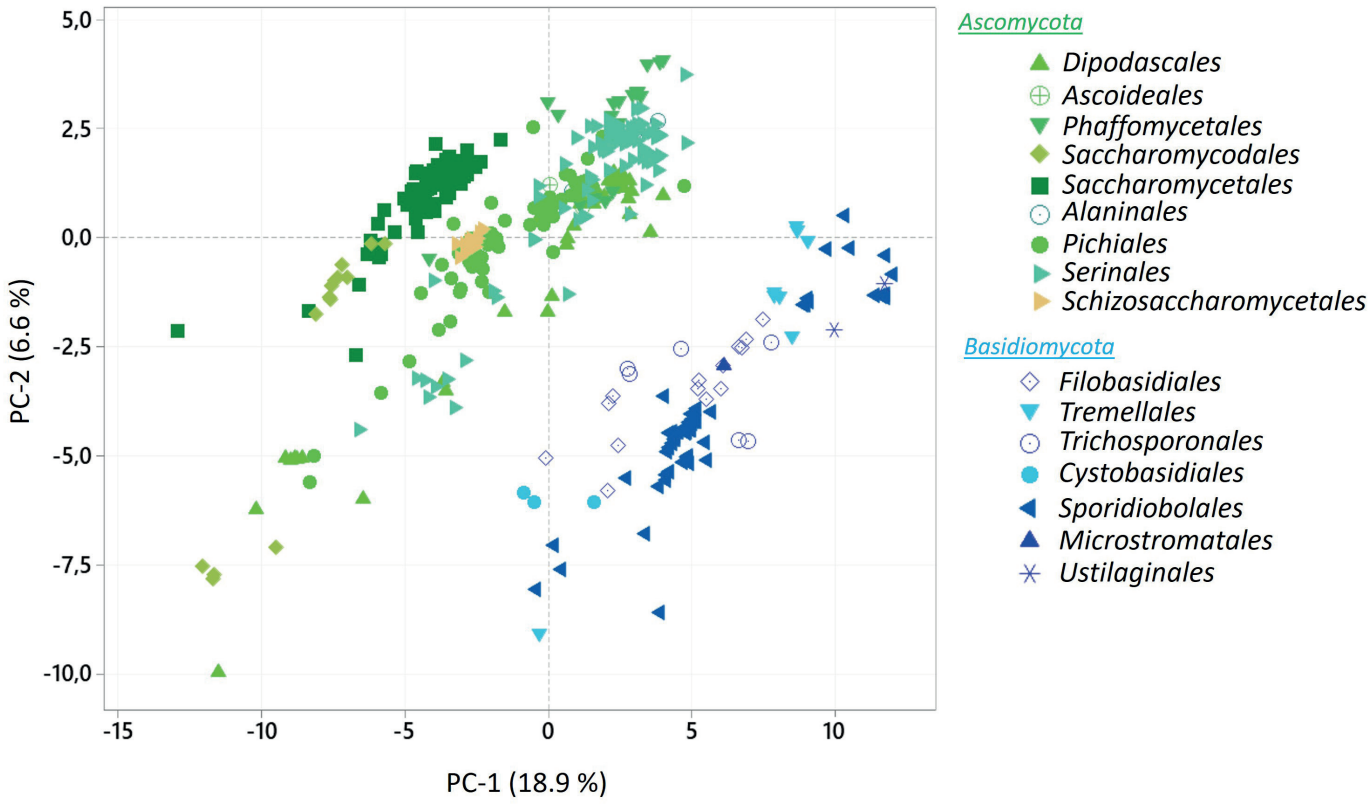

Fig. 3. PCA visualization of strains distribution, considering the number of genes associated to each metabolic KEGG category (Supplementary Tables S4 and S5), and coloured according to taxonomic level. Green – Phylum *Ascomycota*; Blue – Phylum *Basidiomycota*. Symbols represent different orders according with the legend. PC1 – 18.9 %; PC2 – 6.6 %.

separation, but remaining distinct from the remaining orders *Phaffomycetales* and *Ascoideales*. These latter orders appear to be closer to the orders of the *Pichiomycetes* class, particularly to *Serinales*. This observation aligns with the findings and discussion presented in Fig. 2 regarding the overlap of species from *Debaryomycetaceae* and *Metschnikowiaceae* (from *Serinales* / *Pichiomycetes*), and points to the need for further validation. This is especially important in light of the findings from Groenewald *et al.* (2023) and Opulente *et al.* (2023), but further investigation is warranted to deeper into the positioning of these two families.

Within *Basidiomycota*, two distinct groups of *Sporidiobolales* are discernible in the PCA visualization, primarily separated by the second principal component. The rationale behind this separation is not immediately evident, especially given that genomes of the same genus, such as *Rhodotorula mucilaginosa* and *R. toruloides*, are present in both groups. Upon closer scrutiny, it becomes apparent that *R. toruloides* consistently possesses a large number of genes annotated into various functional categories, including COG, CAZyme and specific KEGG pathways, compared to other *Rhodotorula* species. Specifically, it was shown before that *R. toruloides* strains demonstrate intrinsic metabolic properties, and distinctive metabolic trade-offs associated with protein and lipid production (Tiukova *et al.* 2019). These unique characteristics set them apart from other yeasts within the same genus and family.

Vishniacozyma victoriae, a species within *Tremellales*, occupies a unique position in the PCA analysis, as denoted by a light blue inverted triangle in the lower part of the image. This species has previously demonstrated distinct metabolic characteristics, including the individual presence of pathways associated with flavone and flavonol biosynthesis, as well as a lower number of differentially expressed stress-related genes (Baeza *et al.* 2022). The distinct metabolic profile of *V. victoriae* is further emphasized by the heatmap analysis presented in Table S6. In this analysis, the species exhibits clear and distinctive patterns, particularly in KEGG modules related to lipid, aromatic amino acid, and glycosaminoglycan metabolism. These findings reinforce the unique and specialized metabolic features of *V. victoriae* within the context of *Tremellales*, as shown in Table S5. Notably, this species was the only one within this family for which fewer than 7000 genes were associated with a KEGG pathway, in contrast with the family average of 7595 genes.

To explore whether the number of KO´s is correlated with genome size, as suggested by the annotation symbols of Fig. 2 (blue and green pentagons), we calculated a cumulative percentage based on the number of KO´s and the total number of genes, as illustrated in Fig. 4. Results reveal that, despite *Schizosaccharomycetales* have a lower number of genes annotated to KEGG pathways, they exhibited the second-highest percentage of genes in Fig. 4. This suggests that, notwithstanding their relatively small genome size (with an average of 12.6 Mb), especially when compared to orders from *Saccharomycotina*, such as *Dipodascales* (with an average of 18.6 Mb) or *Alaninales* (with an average of 13.1 Mb), they are among the functionally richest orders. Upon closer examination of the results, it becomes evident that the majority of genes in *Schizosaccharomycetales* are associated with Metabolism (depicted in blue; 47.9 %), and Human diseases (shown in green; 36.8 %). These percentages surpass those observed in other taxa, with the exception of *Pachysolenaceae*. The substantial number of genes linked to human diseases can be attributed to the potential utility of *Schizosaccharomyces* yeasts for studying human diseases, as demonstrated in previous studies (Sun *et al.* 2005). This functional richness is also reflected in the COG categories annotations (with an average of 89.5 % of genes annotated) and the CAZymes attribution (with an average of 2.3 % of the genes attributed to a CAZyme category) (Table S4).

Furthermore, we examined the number of complete KEGG modules, which are defined as "functional units of gene sets in metabolic pathways, including molecular complexes" (Kanehisa *et al.* 2021). This approach enabled us to identify sets of orthologous genes sharing the same biochemical steps for a specific function (Tables S6, S7). While the majority of the modules exhibited conservation among all yeasts, with either presence or absence in all species, some specific differences can be discerned in the heatmap presented in Table S7: *(i)* pectin (M00081) and D-Galacturonate (M00630) degradation genes are more abundant in almost all families of *Basidiomycetes*, whereas they are nearly absent in *Ascomycota*. *(ii)* Carbon fixation modules and biosynthesis of phytochemical compounds are also more represented in *Basidiomycota*, as expected. This aligns with the known metabolic diversity of *Basidiomycota*, which includes various ecological roles such as wood decay, mycorrhizal symbiosis, and plant pathogenesis. The higher representation of carbon fixation modules suggests an enhanced ability to assimilate carbon from different sources, while the increased presence of biosynthesis pathways for phytochemical compounds reflects the potential involvement of *Basidiomycota* in the synthesis of bioactive molecules relevant to plant interactions. *(iii)* Betaine biosynthesis (M00555), particularly the biochemical transformation of choline to betaine is exclusive of *Basidiomycota* (modules complete in almost all families), being completely absent from all *Ascomycota*. Betaine, as a metabolite, serves various roles, including osmoregulation and stress tolerance. The exclusive presence of this biosynthetic pathway in *Basidiomycota* suggests a specialized adaptation to environmental conditions or ecological niches unique to this fungal group (Senik *et al.* 2015). *(iv)* Within *Ascomycota*, orders *Saccharomycetales* and *Saccharomycodales* showed, generally, the highest differences in the analysed modules, aligning with other genomic distinctions discussed in previous sections and visible in Figs. 2 and 3. The results have the potential to unveil numerous other crucial differences between taxonomic levels, some of which may not have been revealed by yeast genomics studies to date. However, a more in-depth analysis is beyond the scope of this work and could be explored in future research.

The non-*Saccharomyces* yeasts fermentome

The term fermentome was defined by Walker *et al.* (2014), as the set of 93 genes whose deletion in *S. cerevisiae* led to a delay in the fermentation process. Recognizing the significance of these genes in understanding strain adaptation to fermentative conditions, we conducted a thorough examination of their presence in the genomes of our 530 non-*Saccharomyces* yeasts (Table S8). Out of the 93 genes, 32 were identified in all yeasts, regardless of their fermentative capacity, indicating their association with fundamental cellular functions. However, the remaining 61 genes were found to be absent in at least one strain or species. It is noteworthy that four genes, still annotated as dubious open reading frames, were absent from all the 530 genomes and appear to be exclusive to *S. cerevisiae*. As expected, yeasts from *Saccharomycetaceae* exhibited the presence of the highest number of genes from this list, particularly in the genera *Zygotorulaspora* (89 % of the genes), *Torulaspora* (88.7 %), *Kazachstania* (88.3 %), *Zygosaccharomyces* (87.6 %), *Lachancea* (87.6 %), and *Kluyveromyces* (86.4 %). Conversely, species *Filobasidium uniguttulatum*, *Papiliotrema laurentii* and *Vishniacozyma victoria*e showed the lowest number of these *S. cerevisiae* fermentation-related genes - 52 (56 %).

Table 1. Forty-one genes defined as the putative fermentome of non-*Saccharomyces* yeasts within the present work. Genes underlined represent the six genes present in more than 95 % of fermentative strains, and in less than 5 % of non-fermentative strains (strict fermentome). The remaining 36 genes were detected in more than 90 % of fermentative strains, and in less than 10 % of non-fermentative (less strict fermentome). Short description information was collected from *Saccharomyces* genome database (SGD: https://www.yeastgenome.org/). Further information about these genes is available at Table S10. *NA*- not available

Gene name	Short description	Gene name	Short description
MCM10	MiniChromosome Maintenance	***AIP5***	Actin Interacting Protein
CIC1	Core Interacting Component	***ECM21***	ExtraCellular Mutant
ATS1	Alpha Tubulin Suppressor	***RAP1***	Repressor/Activator site binding Protein
Vps75	Vacuolar Protein Sorting	***GEP3***	GEnetic interactors of Prohibitins
PTH1	Peptidyl-tRNA Hydrolase	***MNN1***	MaNNosyltransferase
EDC3	Enhancer of mRNA DeCapping	***BSC5***	Bypass of Stop Codon
LYS14	LYSine requiring	***BUL1***	Binds Ubiquitin Ligase
RTC3	Restriction of Telomere Capping	***BUL2***	Binds Ubiquitin Ligase
RTN1	ReTiculoN-like	***RPA34***	RNA Polymerase A
EMP47	*NA*	***HKR1***	*Hansenula mrakii* Killer toxin Resistant
RMD9	Required for Meiotic nuclear Division	***ROY1***	Repressor Of Ypt52
YBR238C	*NA*	***CAF130***	CCR4 Associated Factor
TYE7	Ty1-mediated Expression	***LDB19***	Low Dye Binding
DPB11	DNA Polymerase B (II)	***SWC5***	SWr Complex
MDS3	Mck1 Dosage Suppressor	***EOS1***	ER-localized and Oxidants Sensitive
PMD1	Paralog of MDS3	***RSM18***	Ribosomal Small subunit of Mitochondria
YPR084W	Putative protein of unknown function	***YMR209C***	*NA*
SYS1	Suppressor of Ypt Six	***SCD5***	Suppressor of Clathrin Deficiency
MNT2	MaNnosylTransferase	***KEI1***	Kex2-cleavable protein Essential for Inositol phosphorylceramide synthesis
MRPL15	Mitochondrial Ribosomal Protein, Large subunit	***STS1***	Sec Twenty-three Suppressor 1
EMP46	*NA*	***RGT1***	Restores Glucose Transport
		RGT1	Restores Glucose Transport

Given the lack of a direct extrapolation from the *S. cerevisiae* fermentome, established by Walker *et al.*, to the non-*Saccharomyces* fermentome, we conducted a detailed analysis using BLASTP. In this analysis, we employed the proteome of *S. cerevisiae* as a query to investigate the presence of all *S. cerevisiae* genes in each of the non-*Saccharomyces* genomes. This approach allowed us to explore patterns associated with the fermentative capacity of non-*Saccharomyces* yeasts, as outlined by Kurtzman *et al.* (2010). Although these yeasts originated from winemaking environments, only 71 % of them demonstrated the ability to ferment glucose, according to Kurtzman *et al.* (2010). An additional 6.4 % exhibited variable fermentative capacity. Consequently, the yeasts were categorized into three groups, as illustrated in Fig. 2: glucose-fermentative yeasts (green squares), non-fermentative yeasts (red squares), and those with variable fermentative character (yellow squares). Notably, 23 % of the strains were identified as non-fermentative for glucose.

Table S9 presents the comprehensive results, delineating the presence or absence of all genes in the 530 genomes. It delves into groups of genes present in fermentative strains but absent in the non-fermentative ones. The findings were meticulously analysed to derive a definite list of genes, constituting the fermentome of non-*Saccharomyces* yeasts (Table 1). A total of six genes were detected as being present in more than 95 % of the fermentative yeasts (more than 357 strains), but found in less than 5 % of the non-fermentative ones (less than 6 strains). This 5 % margin allows for the consideration of possible false positives or negatives obtained in previous genomic annotation steps.

Six genes met these criteria, and are identified in Table 1- *MCM10, CIC1, ATS1, VPS75, PTH1, EDC3* – and illustrated in Fig. 2 by pink stars (filled star represents the presence of the gene, and an empty star its absence), and were denominated as the putative strict fermentome of non-*Saccharomyces* yeasts. A very clear pattern is visible when comparing the presence of these genes with the described fermentative capacity of the yeasts, as indicated by the red (absent), yellow (variable) and green (positive) squares in Fig. 2. Strains with variable fermentative capacity also consistently exhibited the presence of these genes, akin to the fermentative ones. The functions of the six genes are described in Table S10. To the best of our knowledge, this study marks the first association of these six genes with the fermentative capacity of non-*Saccharomyces* strains, being present also in *S. cerevisiae*. The only exception could be gene *CIC1*, which was previously observed to be overexpressed in 14 winemaking *S. cerevisiae* strains during

the study of Carreto *et al.* (2008), using comparative genome hybridization on array (aCGH), although the study did not provide a detailed explanation for its role in fermentation.

The obtained results, despite signifying novel and crucial progress in understanding the fermentative capacity of certain non-*Saccharomyces* yeasts, need to be approached with due diligence. It is important to acknowledge that, while these species were isolated from grapes or wine must, their presence in an open environment suggests the potential for encountering them in various settings or on other fruits. Therefore, the fermentome of non-*Saccharomyces* yeasts defined in this study should be considered as putative, and not necessarily exclusive of wine environments, since these yeasts can also ferment other fruits. Validation through experimental studies, currently lacking for these species, is essential to confirm their status. Nevertheless, if we tailor the obtained results solely to the non-*Saccharomyces* species frequently found in wine, our findings demonstrate robustness. By focusing on the species reported in more than 50 isolation campaigns in the work of Drumonde-Neves *et al.* (2021) – *Lachancea thermotolerans*, *Pichia kluyveri*, *Schizosaccharomyces pombe*, *Starmerella stellata* and *Wickerhamomyces anomalus* (*Hanseniaspora* was not considered due to low-quality genomes) – the six genes identified as constituting the non-*Saccharomyces* fermentome were present in all strains from these species. This observation reinforces our results, although it does not negate the necessity to experimentally assess the role of these genes in the fermentative process carried out by these species.

Gene *MCM10* (minichromosome maintenance protein 10) promotes helicase activity, being important for replication initiation, and guarantying replisome stability (Campos *et al.* 2023). Several works state that this gene is an evolutionarily conserved component of the eukaryotic replication machinery (Du *et al.* 2012), despite its low domain and sequence conservation across major taxonomic groups. However, in the context of yeasts, evidence of the presence of this gene is limited to *S. cerevisiae* (Sawyer *et al.* 2004) and *Schizosaccharomyces pombe* (Gregan *et al.* 2003), to the best of our knowledge. In the current study, we observed that this gene was only present in the genome of 412 out of the 530 strains, spanning a total of 107 species out of 134, which includes all the fermentative species. These findings appear to challenge the theory suggesting the conservation of this gene in all eukaryotes. However, it is crucial to note that further experimental analysis is required to validate these results conclusively. The association of *MCM10* with fermentation has not been suggested previously. Our rationale draws from the study of Alver *et al.* (2014) which revealed a novel role for the N-terminus domain of *MCM10* in DNA damage resistance. The study demonstrated that its interaction with other genes enhances resistance to replication stress caused by external factors. Given that yeast exposure to glucose fermentative conditions induces replication stress (Santos *et al.* 2016), we speculate that the presence of this gene (in combination with the other genes identified as the fermentome) enables these species to withstand the stressful conditions of glucose fermentation.

The four remaining genes proposed as the putative non-*Saccharomyces* strict fermentome – *ATS1* (Alpha Tubulin Suppressor 1), *VPS75* (Vacuolar Protein Sorting), *PTH1* (Peptidyl-Trna Hydrolase) and *EDC3* (Enhancer of mRNA DeCapping) – were observed to be present in the vast majority of the fermentative species, in contrast to the non-fermentative ones. However, there is no apparent association with the fermentative capacity of those yeasts. These genes are not essential for cell survival and function, but they play known roles in the regulation and metabolic activity of the cell. All of them are regulated by transcription factors in response to stress, and all seem to be associated with stability of nucleic acid and proteins. In this way, they could be of great importance to increase yeasts resistance to the stress found during the fermentative process. A more in-depth analysis, supported by experimental validation, is needed to elucidate why these genes exhibit such a positive correlation with the yeast's capacity to ferment glucose, as suggested in our study.

A complementary analysis was undertaken by adjusting the thresholds to 90 % for the presence of genes in fermentative strains and increase it to 10 % for their presence in the non-fermentative strains. This approached yielded an additional set of

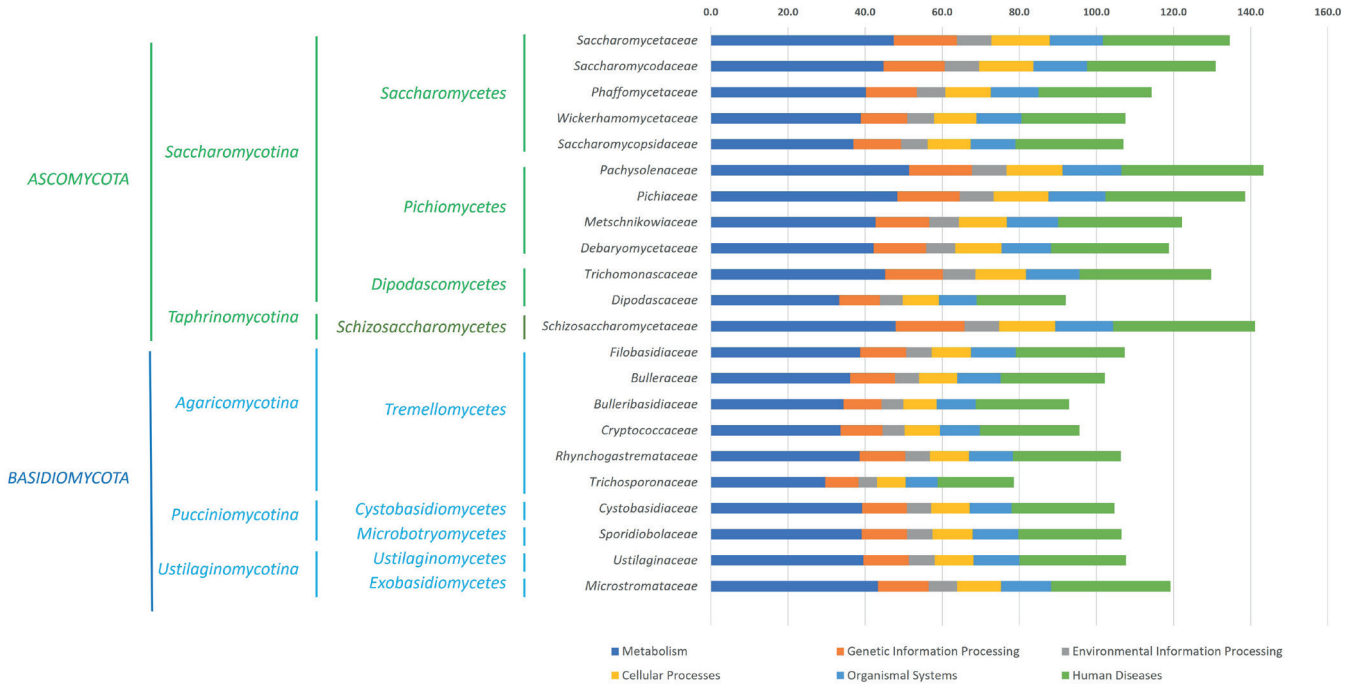

Fig. 4. Cumulative percentage of proteins annotated to each metabolic KEGG category (detailed results in Tables S4 and S5), divided by taxonomic level (from left to right: Phyla, Class, Order, Family).

Table 2. Ninety-six genes selected as the putative flavorome of non-*Saccharomyces* strains, as relevant for aroma and flavor production in wine and divided by main categories. Further information about these genes, together with knowledge about their presence in the genome of the 530 non-*Saccharomyces* yeasts are available in Table S11.

Main metabolic categories	Gene name
Higher alcohol - permeases of amino acids	***BAP2*** (YBR068C), ***MEP2*** (YNL142W), ***TAT1*** (YBR069C), ***MUP1*** (YGR055W), ***TAT2***(YOL020W), ***MUP3*** (YHL036W
Higher alcohol - aromatic amino acid transferase	***ARO8*** (YGL202W), ***ARO9*** (YHR137W)
Higher alcohol - branched chain amino acid transferase	***BAT1*** (YHR208W), ***BAT2*** (YJR148W), ***PAD1*** (YDR538W)
Higher alcohol - decarboxylase	***PDC1*** (YLR044C), ***PDC5*** (YLR134W), ***PDC6*** (YGR087C), ***THI3*** (YDL080C)
Higher alcohol - alcohol dehydrogenase	***ADH1*** (YOL086C), ***ADH2*** (YMR303C), ***ADH3*** (YMR083W), ***ADH4*** (YGL256W), ***ADH5*** (YBR145W), ***ADH6*** (YMR318C), ***ADH7*** (YCR105W), ***SFA1*** (YDL168W), ***GRE2*** (YOL151W), ***YPR1*** (YDR368W), ***SPE1*** (YKL184W), ***OYE2*** (YHR179W), ***HOM2*** (YDR158W)
Higher alcohol - aryl alcohol dehydrogenase	***AAD3*** (YCR107W), ***AAD4*** (YDL243C), ***AAD10*** (YJR155W), ***AAD14*** (YNL331C), ***AAD15*** (YOL165C), ***AAD16*** (YFL057C)
Higher alcohol - regulation	***ARO80*** (YDR421W), ***GAT2*** (YMR136W), ***GLN3*** (YER040W), ***GZF3*** (YJL110C), ***DAL80*** (YKR034W)
Acetate esters - alcohol acetyl transferases	***ATF1*** (YOR377W), ***ATF2*** (YGR177C), ***SLI1*** (YGR212W)
Ethyl esters - ethanol O-acyltransferase and esterase	***EEB1*** (YPL095C), ***EHT1*** (YBR177C), ***MGL2*** (YMR210W), ***IAH1*** (YOR126C)
Volatile organic acids - acetic acid	***ACS1*** (YAL054C), ***FPS1*** (YLL043W)
Volatile organic acids - aldehyde	***ALD2*** (YMR170C), ***ALD3*** (YMR169C), ***ALD4*** (YOR374W), ***ALD5*** (YER073W), ***ALD6*** (YPL061W), ***ILV2*** (YMR108W), ***ILV3*** (YJR016C), ***ILV5*** (YLR355C), ***BDH1*** (YAL060W)
p-ABA synthesis - synthesis of p-ABA from chorismate	***ABZ1*** (YNR033W), ***ABZ2*** (YMR289W)
Aromatic amino acid synthesis - synthesis of chorismate, phenyalanine, tryptophan and tyrosine	***ARO1*** (YDR127W), ***ARO2*** (YGL148W), ***ARO3*** (YDR035W), ***ARO4*** (YBR249C), ***TRP2*** (YER090W), ***TRP3*** (YKL211C), ***ARO7*** (YPR060C), ***PHA2*** (YNL316C), ***TYR1*** (YBR166C)
Benzyl alcohol / benzaldehyde synthesis - mandelate pathway	***ARO10*** (YDR380W), ***SCS7*** (YMR272C), ***DLD1*** (YDL174C), ***DLD2*** (YDL178W), ***DLD3*** (YEL071W)
Sulfur metabolism - Volatile thiols - conversion of conjugated precursors into varietal thiols	***GAP1*** (YKR039W), ***OPT1*** (YJL212C)
Sulfur metabolism - Volatile thiols - B-lyases cleaving the cysteinylated precursors releasing 4MMP	***BNA3*** (YJL060W), ***CYS3*** (YAL012W), ***GLO1*** (YML004C), ***IRC7*** (YFR055W), ***STR3*** (YGL184C), ***CYS4*** (YGR155W), ***STR2*** (YJR130C)
Sulfur metabolism - Volatile thiols - degradation of glutathionylated precursors	***DUG1*** (YFR044C), ***DUG2*** (YBR281C), ***DUG3*** (YNL191W), ***ECM38*** (YLR299W)
Sulfur metabolism - sulphate and sulfite reductases	***MET10*** (YFR030W), ***MET14*** (YKL001C), ***MET16*** (YPR167C), ***MET17*** (YLR303W), ***MET2*** (YNL277W), ***MET3*** (YJR010W), ***MET5*** (YJR137C), ***MET6*** (YER091C)
Glycerol production - redox reactions with flavor implications	***GPD1*** (YDL022W), ***GPD2*** (YOL059W)

36 genes (detailed in Table 1 and highlighted in yellow in Tables S9 and S10). While not regarded as part of the strict fermentome of non-*Saccharomyces* yeasts, these genes provide grounds for speculating about their potential significance in elucidating the fermentative behaviour of these yeasts. Further exploration of these genes in future studies could prove valuable.

The non-*Saccharomyces* yeasts flavorome

Previously, the metabolic potential of 15 non-*Saccharomyces* species (770 isolates) was investigated, focusing on the production of enzymes of enological significance, including glycosidases, *β*-lyases, pectinases, proteases and cellulases (Belda *et al.* 2016). This study marked the introduction of the term 'flavorome'. The investigation identified strain-dependent features as the primary contributors to variation, while pinpointing certain species, namely *W. anomalus*, *T. delbrueckii*, *L. thermotolerans* and *H. uvarum*, as the highest producers of these specific enzymes. Nevertheless, no attempt was made to establish a correlation between this capacity and the yeasts' genomic features. In the current work, we examined the dataset of 530 genomes to identity genes associated with the aroma and flavor traits that these yeasts bring to wine (Table 2 and S11). Nineteen major biotechnological categories were established, gathering 96 genes selected from pertinent literature (refer to the methods section), and grouped according to their involvement in the production of higher alcohols, acetate esters, ethyl esters, volatile organic acids, aromatic amino acids, benzyl alcohols, sulfur metabolism or glycerol.

Results (Table S11) show that genes from four categories are conservatively present in all the species (and in all the isolates), irrespective of the strains' taxonomic group or whether they are categorized as fermentative or not: *(i)* "higher alcohols-aromatic amino acid transferase" - *ARO8* and *ARO9* genes; *(ii)* "higher alcohols-branched chain amino acid transferase" - *BAT1* and *BAT2*; *(iii)* "higher alcohol-decarboxylase" - *PDC1, PDC5, PDC6* and *THI3*; and *(iv)* "Glycerol production-redox reactions with flavor implications" - *GPD1* and *GPD2*. While the involvement of these genes in shaping the aromatic and flavor profiles of wines is well established, it is intriguing to note that these genes are universally present across all yeast species. What adds complexity to the equation are the varying degrees of expression of these genes under fermentative conditions. It's not just the presence of these genes that matters

but rather how actively they are expressed during fermentation. This dynamic interplay between gene expression and fermentation conditions is what ultimately influences the distinct aromas and flavours imparted by each yeast strain to the wine. Unravelling the intricacies of these genetic mechanisms in different yeasts holds the key to gaining deeper insights into their contribution to aroma and flavour of wine.

When comparing our findings with those of Belda *et al.* (2016), we have confirmed the flavor and aroma potential of those species by showing hits for the majority of the analysed genes in our BLASTP analysis. Five strains of *W. anomalus* were analysed in the current work, and even though inter-strain differences were discernible, two strains exhibited the presence of all the 96 genes, with the exception of *AAD15* which will be discussed below. These results are also in line with the proposed potential of this species to enhance the aromatic qualities of wine (Sabel *et al.* 2014). Similarly, *L. thermotolerans*, recognized for its proficiency in producing various acetate esters and certain terpenes (Whitener *et al.* 2016), revealed the presence of all genes in two of the three tested strains. In this case, an exception was found for gene *PAD1*, a particularity found for several other species, which will also be further discussed below. *Torulaspora delbrueckii* was the species with the higher number of strains analysed – 74 – in the current manuscript, as a consequence of recent sequencing efforts within our research group. As anticipated, this extensive analysis revealed significant inter-strain variability. However, the majority of strains exhibited the presence of over 95 % of the scrutinized genes. This inter-strain variation, extensible also to their metabolic particularities, is further associated with the different origins, both geographical and technological, as previously demonstrated by our group (Silva *et al.* 2022). Additionally, the presence of gene clusters related to galactose and maltose utilization further supports this association. Only one strain of *H. uvarum* was analysed in the current work due to the failure of the majority of the genomes to pass the quality control. This particular strain exhibited a significant deficiency in a substantial portion of the assessed genes, a pattern consistent with other specific genomic features discussed earlier in this work. This species appears to possess distinctive features such as a small genome size and a limited number of coding genes. Therefore, additional efforts in the future will be essential to comprehensively investigate the potential significance of this yeast in winemaking through genomic analysis. While it is important to note that no definitive conclusions can be drawn due to the small sample size, our results align with certain distinctive characteristics observed in other *Hanseniaspora* species. Notably, the work of Seixas *et al.* (2017) highlighted that *H. guilliermondii* lacks enzymes required for glycerol or galactose catabolism, as well as for biotin and thiamine biosynthesis. Furthermore, it is deficient in specific fructose transport mechanisms.

Upon detailed examination of the 19 categories into which the 96 genes were divided (Table 2), 18 exhibited statistically significant differences after conducting the Kruskal–Wallis one-way analysis of variance (Fig. 5A). This analysis assessed the presence of the genes from each category across strains from different taxonomic levels. The category 'higher alcohol - alcohol dehydrogenase', comprised of *ADH* genes among others, revealed no statistical difference at

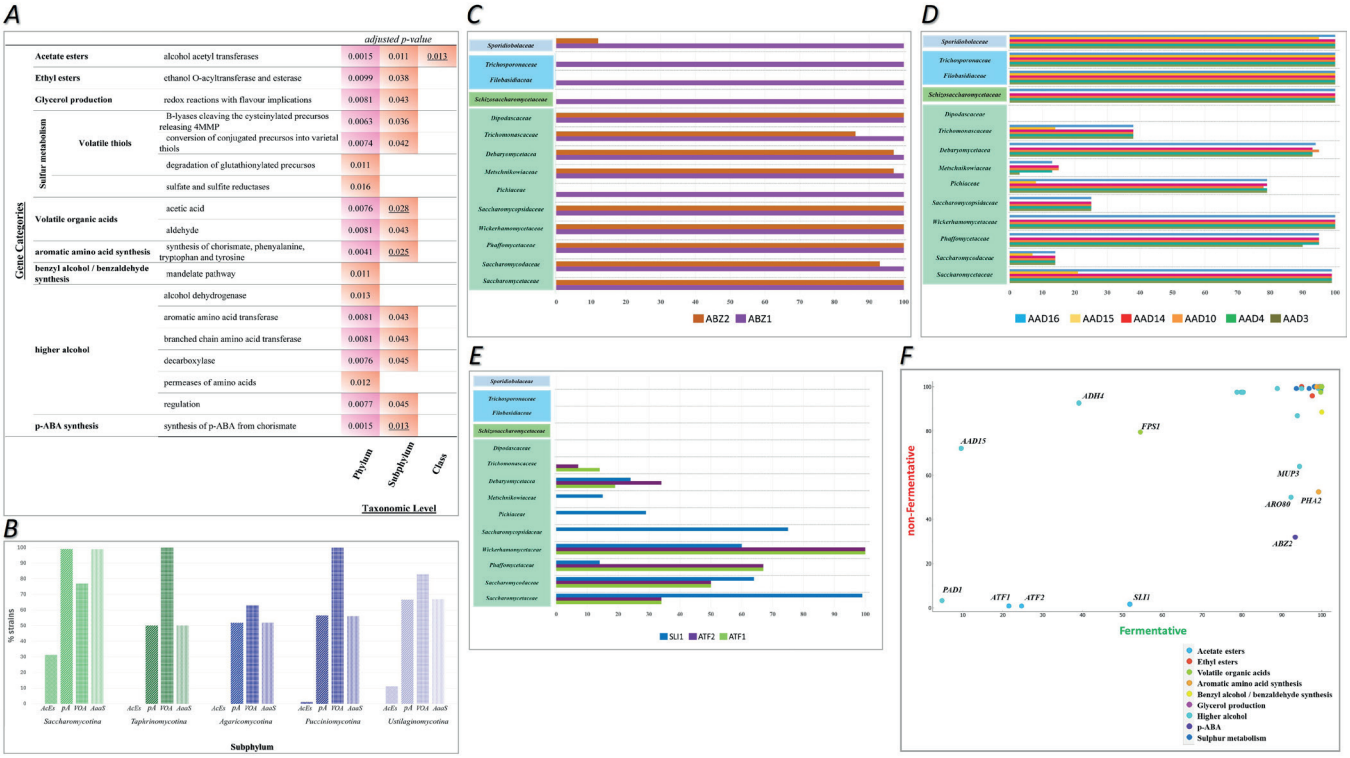

Fig. 5 Analysis of the flavorome of non-*Saccharomyces* yeasts, in particular emphasizing the presence and absence of the 96 genes related with aroma and flavour production during wine fermentation. **A.** Heatmap displaying the adjusted p-values from the Kruskal–Wallis one-way analysis of variance, which assesses the presence/absence of genes within each taxonomic group. Significant p-values are indicated with colour coding: < 0.05 in orange, and < 0.01 in pink. **B.** Average percentage of strains within each subphylum that encode genes from the indicated categories: AcEs – "Acetate esters - alcohol acetyl transferases"; pA – "p-ABA synthesis - synthesis of p-ABA from chorismite"; VOA – "Volatile organic acids - acetic acid"; AaaS – "aromatic amino acid synthesis - synthesis of chorismate, phenyalanine, tryptophan and tyrosine"; **C.** Percentage of strains within each family that possess genes *ABZ1* and *ABZ2*; **D.** Percentage of strains within each family that possess genes *AAD3, AAD4, AAD10, AAD14, AAD15* and *AAD16*; **E** Percentage of strains within each family that possess genes *SLI1, ATF1* and *ATF2*; **F** Scatter plot illustrating the distribution of each of the 96 genes in fermentative versus non-fermentative strains, with individual genes coloured based on overarching categories as listed in Table 2. For panels (C), (D) and (E), families represented by fewer than 4 strains were not included in the analysis.

any taxonomic level. Alcohol dehydrogenase (*ADH*) enzymes are integral to the essential biochemical machinery, facilitating the conversion of alcohols to aldehydes or ketones, with a concomitant reduction of NAD+ to NADH. This reaction is a pivotal step in the metabolic pathways of fermentation and respiration, and as such, *ADH* enzymes are broadly conserved across different species, including fungi, plants, and animals. The notable exception of this conservation was *ADH4*, which was absent in certain species, and revealed to be largely associated with non-fermentative species (Fig. 5F). As for the other categories under scrutiny, 'Acetate esters – alcohol acetyl transferases' was the sole group demonstrating statistically significant differences at the class level (Fig. 5A). Furthermore, at the subphylum level, categories 'Volatile organic acids – acetic acid', 'Aromatic amino acid synthesis – synthesis of chorismite, phenylalanine, tryptophan and tyrosine' and 'synthesis of p-ABA from chorismite' showed the most substantial differences. These were further elaborated upon in Fig. 5B, to elucidate variation within the five subphyla.

Results illustrate the percentage of strains within each subphylum that possess genes related to Acetate esters (AcEs), p-ABA synthesis (pA), Volatile organic acids (VOA), and Aromatic amino acid synthesis (AaaS). Genes associated with alcohol acetyl transferases activity (Acetate esters – AcEs) showed the most pronounced differences between subphyla, as highlighted also in Fig. 5A, being present in about 30 % of *Saccharomycotina* yeasts, but nearly absent in the majority of the other subphyla. This underscores the potential role of this subphylum in acetate ester production, a key contributor to wine aroma. During alcoholic fermentation, the activity of these enzymes can produce many important aroma compounds, including phenylethyl acetate, which imparts a floral aroma reminiscent of roses (Lilly *et al.* 2000). In the context of volatile organic acids, specifically the production of acetic acid, *Saccharomycotina* yeasts exhibit one of the lowest gene presence rates, only surpassed by *Agarimycotina*, suggesting their suitability for wine production as the production of acetic acid is considered a sensory defect. On the contrary, gene categories associated with advantageous traits for wine production, such as the synthesis of aromatic amino acids chorismite, phenylalanine, tryptophan and tyrosine, and the synthesis of para-aminobenzoate (p-ABA), were present in around 98 % of the isolates, unlike other subphyla which displayed significantly lower percentages. The disparity in p-ABA production, an important aromatic compound and precursor of other valuable products, was mainly attributed to the gene *ABZ2* (Fig. 5C), which exhibited considerable differences at the family level. Conversely, gene *ABZ1* was found to be conserved across all yeast strains. *ABZ2* catalyses the third step in p-ABA acid biosynthesis, and thus participates in folic acid biosynthesis (Perli *et al.* 2020). Its absence in certain yeast groups, while not directly impacting wine production, is known to influence yeast fermentative capacity.

One category that yielded variable and interesting results in terms of gene presence across the species, and worth of an extensive discussion, was "higher alcohol - aryl alcohol dehydrogenase" –, composed by genes *AAD3, AAD4, AAD10, AAD14, AAD15* and *AAD16* (Fig. 5D). These genes are characterized as putative aryl-alcohol dehydrogenases, but their role was not yet fully elucidated. The *Saccharomycetaceae* family, which includes yeast species mostly known for their role in wine fermentation, exhibits these genes in the majority of its strains (with the exception of *AAD15*, a paralog of *AAD3*, which was absent in 80 % strains, and that will be discussed further). Whitin the *Ascomycota*, the families *Debaryomycetaceae*, *Phaffomycetaceae*, *Wickerhamomyceteae* and *Schizosaccharomycetaceae* also exhibited a high presence of *AAD* genes, with more than 90 % of the strains containing these genes (Fig. 5D). Notably, the latter two families demonstrated the presence of *AAD* genes in all the analysed genomes. To the best of our knowledge, aryl-alcohol dehydrogenase activity has not been conclusively documented in yeasts from some of these families, as per the current literature. In contrast, some studies have even reported unsuccessful attempts to detect activity of this enzyme, even when specifically tested (Zang *et al.* 2022, Yu *et al.* 2023). Conversely, the families *Dipodascaceae*, *Metschnikowiaceae* and *Saccharomycodaceae* (which include the genera *Hanseniaspora* and *Saccharomycodes*) exhibited an almost complete absence of these genes. Within the later family, the sole exceptions were one strain of *H. vineae* – NRRL Y-17529 – and the strain of *H. lindneri* - CBS 285 – where all these genes (with the exclusion of *AAD15*) were present. Nevertheless, these genes have previously been identified in some Uruguayan strains of *H. vineae* (Martin *et al.* 2016), suggesting that our study is not the first to associate their presence in winemaking.

The *Saccharomycetaceae* family, which includes yeast species renowned for their fermentative skills and aromatic contributions, such as the genera *Torulaspora, Zygotorulaspora,* and *Lachancea*, also demonstrated a nearly ubiquitous presence of categorized as putative aryl-alcohol dehydrogenases, again with the exception of *AAD15*. Intriguingly, the gene *AAD15* gene stands out a peculiar one as already referred. Documented as a paralog of *AAD3,* it has been implicated in playing a vital ancestral role in detoxifying aromatic aldehydes in ligninolytic fungi (Yang *et al.* 2018). In *S. cerevisiae*, its function in the production of aroma metabolites via the Ehrlich pathway remains inconclusive, and this gene was not detected in the genome of several analysed strains. In our current work, gene *AAD15* was found only in the genome of 40 species (29 % of the total), with the interesting finding that it is present almost exclusively in the genome of non-fermentative species (Fig. 5F; 72 % of non-fermentative species versus 9.6 % of fermentative ones). This finding correlates with the fact that this gene was not part of the non-*Saccharomyces* fermentome defined in the previous section.

One interesting exception within the genes catalogued as 'higher alcohol – aryl alcohol dehydrogenase' is observed in the genus *Starmerella*. Among the five *Starmerella* species analysed in this study, three of them – *Starmerella apicola, Starmerella bacillaris* and *Starmerella magnoliae* – exhibited total absence of the genes from this category – *AAD3, AAD4, AAD10, AAD14, ADD15* and *AAD16*. In contrast, *Starmerella bombi* and *Starmerella stellata* harboured the entirety of these genes, with the previously noted exception of *AAD15*. The contrasting gene presence among these species, points to a potential divergence in metabolic capabilities concerning higher alcohol production within the genus, which may have implications for their distinct ecological niches and utility in industrial processes. In earlier studies, Gonçalves *et al.* (2018; 2020) showed that *Starmerella bombicola* – though not included in the present study – harbours an alcohol dehydrogenase gene of bacterial origin (*ADH1a*) that is functional in alcoholic fermentation. The authors propose an evolutionary model wherein orthologs associated with alcohol dehydrogenase were lost in an ancestor of the W/S clade, a hypothesis that adds depth to our understanding of fermentative evolution in these yeasts. Our results suggest that the absence of alcohol dehydrogenase genes, as observed in *S. bombicola*, may not be a universal trait across the entire W/S clade but rather confined to specific species, since the three species in our analysis – *Starmerella apicola, Starmerella bacillaris* and *Starmerella magnoliae* – exhibited the presence of all the *AAD*

and *ADH* genes. Identical results were obtained for gene *FPS1*, only detected in *Starmerella stellata* and *Starmerella bombi*, being absent in all strains from the other three *Starmerella* species. Gene *FPS1* is involved in the efflux of glycerol and xylitol, and this discrepancies within genus *Starmerella* can explain the remarkable different regulation of the glycerol pathway previously reported for some *Starmerella* species (Lemos Junior *et al.* 2018). To the best of our knowledge, these finding about *Starmerella* has not been previously reported and could hold significant importance.

Another category of genes worth to be examined in detail is the 'Acetate esters - alcohol acetyl transferases' set, in particular the genes *ATF1* and *ATF2*. These genes have been identified in *S. cerevisiae* as critical for acetate ester formation (Fujii *et al.* 1994), yet their presence in non-*Saccharomyces* yeasts remains unexplored. However, non-*Saccharomyces* yeasts were already described as producing relevant amounts of acetate esters, in particular of 2-phenylethyl acetate and isoamyl acetate (Rojas *et al.* 2001, Domizio *et al.* 2011). Our results revealed the presence of gene *SLI1*, a member of this category, in species across almost every family of *Saccharomycotina* (Fig. 5E), but the presence of genes *ATF1* and *ATF2* is predominantly found only within families of the *Saccharomycetes* class. This pattern becomes particularly evident when comparing fermentative with non-fermentative species, as shown in Fig. 5F. It is observed that these genes are only found in fermentative species. Notably, yeasts from genera *Cyberlindnera*, *Hanseniaspora, Hyphopichia, Kazachstania, Kluyveromyces, Lachancea, Nakaseomyces, Nakazawaea, Starmerella, Torulaspora* and *Zygotorulaspora* harbour these genes, albeit with considerable intra-species variation (as detailed in Table S11). These findings suggest that these species are worthy of further research to understand their potential contributions to acetate ester production.

Notably, *T. delbrueckii* entirely lacks the *ATF1* and *ATF2* genes, a finding previously reported by Tondini *et al.* (2019). The authors proposed that in *T. delbrueckii* the gene *SLI1* may be functionally active, potentially assuming an alternative role in acetate ester production. Our study corroborates this hypothesis and extends it to other species. Specifically, *Pichia kudriavzevii, Pichia membranifaciens, Pichia norvegensis, Pichia terricola, Saccharomycodes ludwigii, Saccharomycopsis fibuligera, Saturnispora zaruensis, Z. bailii, Z. bisporus, Z. mellis, Z. rouxii* and *Zygotorulaspora florentina* were found to possess only the *SLI1* gene within their genomes (Fig. 5E). The singular presence of *SLI1* in these genomes indicates an alternate genomic pathway for acetate ester synthesis, warranting further investigation into its metabolic implications.

One gene worth mentioning is *PAD1*, catalogued in 'higher alcohols – permeases of amino acids' category, which showed interesting patterns. According to the literature, this gene is associated with the decarboxylation of the aromatic carboxylic acids cinnamic acid, ferulic acid, and coumaric acid, and its activity was already shown in *C. albicans*, *C. dubliniensis*, *Debaryomyces hansenii*, and *Pichia anomala* (Stratford *et al.* 2007). In the current work only *Debaryomyces hansenii* was analysed, and our results confirmed the presence of this gene in all the examined isolates. Even though the two mentioned species of *Candida* were not included in this study, *C. intermedia* genome revealed the presence of the *PAD1* gene in all strains. Additionally, *Barnettozyma pratensis*, *H. vineae*, *Hyphopichia burtonii*, *Hyphopichia gotoi*, *Nakazawaea holstii*, *Papiliotrema laurentii*, *Vishniacozyma victoriae*, *W. anomalus*, *Z. bailii*, *Z. mellis*, *Z. rouxii*, and *Zygotorulaspora florentina*, were the only other species showing the presence of this gene. When comparing fermentative with non-fermentative yeasts (Fig. 5F), our results show that this gene was the one present in less species, independently of their fermentative capacity. This pattern is mainly visible when comparing fermentative with non-fermentative species (Fig. 5F), in which one can observe that these genes were predominant only in fermentative species.

CONCLUSIONS

In conclusion, this study marks a significant advancement in our comprehension of non-*Saccharomyces* yeasts and their potential roles in oenology, particularly by elucidating metabolic pathways associated with fermentation capacity and flavor/aroma synthesis. The identification of a fermentome present in almost all non-*Saccharomyces* species and strongly linked to fermentative abilities could redefine their contribution to vinification processes. Moreover, the characterization of a flavorome underscores the presence of genes that are critical to the aromatic complexity of wine, revealing the untapped potential of these yeasts to elevate wine quality. Notably, the diverse metabolic routes identified in species such as *T. delbrueckii*, and the complex genomic variations within *Starmerella*, underscore the profound connection between yeast genetics and their metabolic contributions, paving the way for innovative approaches to flavour profile enhancement.

Our revelations of gene sets constituting the fermentome and flavorome of non-*Saccharomyces* yeasts present in winemaking environments hold the promise to revolutionize wine production. These findings can contribute to the ability of wine production to adapt to climate change and evolving market demands. Furthermore, the bioinformatics pipeline utilized in this research, with its rigorous quality checkpoints and integration of multiple analytical tools and databases, sets a new standard for future phylogenomics and proteomics investigations in this field.

ACKNOWLEDGEMENTS

This work was supported by the UID/BIA/04050/2013 (POCI-01-0145-FEDER-007569) and UID/BIA/04050/2019 strategic programs (DOI: 10.54499/UIDB/04050/2020), by the project PTDC/BIA-MIC/32059/2017, and by the Infrastructure Network program LA/P/0069/2020 (DOI: 10.54499/LA/P/0069/2020I, funded by national funds through the FCT - Fundação para a Ciência e Tecnologia and by the ERDF-European Regional Development Fund through the COMPETE2020–Programa Operacional Competitividade e Internacionalização (POCI) and Sistema de Apoio à Investigação Científica e Tecnológica (SAICT). Ticiana Fernandes was supported by Fellowship 2021.04595.BD from Fundação para a Ciência e Tecnologia, Portugal (DOI: https://doi.org/10.54499/2021.04595.BD). Financial support was also granted by the FCT to Ricardo Franco-Duarte (CEEC-Ind 2022.00340.CEECIND, DOI: 10.54499/2022.00340.CEECIND/CP1718/CT0018, and Teresa Rito (2023.06638.CEECIND).

DECLARATION ON CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

Alver RC, Zhang T, Josephrajan A, *et al*. (2014). The N-terminus of Mcm10 is important for interaction with the 9-1-1 clamp and in resistance to DNA damage. *Nucleic Acids Research* **42**: 8389–8404.

Aramaki T, Blanc-Mathieu R, Endo H, *et al*. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**: 2251–2252.

Baeza M, Zúñiga S, Peragallo V, *et al*. (2022). Response to cold: A comparative transcriptomic analysis in eight cold-adapted yeasts. *Frontiers in Microbiology* **13**: 1–15.

Bankevich A, Nurk S, Antipov D, *et al*. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**: 455–477.

Barata A, Malfeito-Ferreira M, Loureiro V (2012). The microbial ecology of wine grape berries. *International Journal of Food Microbiology* **153**: 243–259.

Belda I, Ruiz J, Alastruey-Izquierdo A, *et al*. (2016). Unraveling the enzymatic basis of wine "Flavorome": A phylo-functional study of wine related yeast species. *Frontiers in Microbiology* **7**: 1–13.

Belda I, Ruiz J, Esteban-Fernández A, *et al*. (2017). Microbial contribution to wine aroma and its intended use for wine quality improvement. *Molecules* **22**: 1–29.

Bokulich NA, Ohta M, Richardson PM, *et al*. (2013). Monitoring seasonal changes in winery-resident microbiota. *PLoS ONE* **8**: e66437.

Campos LV, Van Ravenstein SX, Vontalge EJ, *et al*. (2023). RTEL1 and MCM10 overcome topological stress during vertebrate replication termination. *Cell Reports* **42**: 112109.

Cantarel BI, Coutinho PM, Rancurel C, *et al*. (2009). The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Research* **37**: 233–238.

Cantarelli C (1955). Studio comparativo dei lieviti apiculati dei generi *Kloeckera* (Janke) ed *Hanseniaspora* (Zikes). *Annals of Microbiology* **6**: 85.

Carreto L, Eiriz MF, Gomes AC, *et al*. (2008). Comparative genomics of wild type yeast strains unveils important genome diversity. *BMC Genomics* **9**: 1–17.

Christinaki AC, Kanellopoulos SG, Kortsinoglou AM, *et al*. (2022). Mitogenomics and mitochondrial gene phylogeny decipher the evolution of *Saccharomycotina* yeasts. *Genome Biology Evolution* **14**: 1–19.

Demšar J, Zupan B, Leban G, *et al*. (2004). Orange: From experimental machine learning. In: *Knowledge Discovery in Databases: PKDD 2004* (Boulicaut, JF, Esposito F, Giannotti F, *et al.*, eds). Lecture Notes in Computer Science vol 3202. Springer, Berlin, Heidelberg: 537–539.

Domizio P, Romani C, Lencioni L, *et al*. (2011). Outlining a future for non-*Saccharomyces* yeasts: Selection of putative spoilage wine strains to be used in association with *Saccharomyces cerevisiae* for grape juice fermentation. *International Journal of Food Microbiology* **147**: 170–180.

Donzella L, Sousa MJ, Morrissey JP (2023). Evolution and functional diversification of yeast sugar transporters. *Essays in Biochemistry* **67**: 811–827.

Drumonde-Neves J, Fernandes T, Lima T, *et al*. (2021). Learning from 80 years of studies: A comprehensive catalogue of non-*Saccharomyces* yeasts associated with viticulture and winemaking. *FEMS Yeast Research* **21**: 1–13.

Du W, Stauffer ME, Eichman BF (2012). Structural biology of replication initiation factor mcm10. *Subcellular Biochemistry* **62**: 197–216.

Dzialo MC, Park R, Steensels J, *et al*. (2017). Physiology, ecology and industrial applications of aroma formation in yeast. *FEMS Microbiology Reviews* **41**: S95–S128.

Ellis DJ, Kerr ED, Schenk G, *et al*. (2022). Metabolomics of non-*Saccharomyces* yeasts in fermented beverages. *Beverages* **8**: 1–14.

Fernandes T, Silva-Sousa F, Pereira F, *et al*. (2021). Biotechnological importance of *Torulaspora delbrueckii*: from the obscurity to the spotlight. *Journal of Fungi* **7**: 712.

Franco-Duarte R, Bessa D, Gonçalves F, *et al*. (2017). Genomic and transcriptomic analysis of *Saccharomyces cerevisiae* isolates with focus in succinic acid production. *FEMS Yeast Research* **17**: 1–12.

Franco-Duarte R, Bigey F, Carreto L, *et al*. (2015). Intrastrain genomic and phenotypic variability of the commercial *Saccharomyces cerevisiae* strain Zymaflore VL1 reveals microevolutionary adaptation to vineyard environments. *FEMS Yeast Research* **15**: 1–13.

Franco-Duarte R, Čadež N, Rito T, *et al*. (2022). Whole-genome sequencing and annotation of the yeast *Clavispora santaluciae* reveals important insights about its adaptation to the vineyard environment. *Journal of Fungi* **8**: 52.

Franco-Duarte R, Mendes I, Umek L, *et al*. (2014). Computational models reveal genotype-phenotype associations in *Saccharomyces cerevisiae*. *Yeast* **31**: 265–277.

Franco-Duarte R, Umek L, Mendes I, *et al*. (2016). New integrative computational approaches unveil the *Saccharomyces cerevisiae* pheno-metabolomic fermentative profile and allow strain selection for winemaking. *Food Chemistry* **211**: 509–520.

Fujii T, Nagasawa N, Iwamatsu A, *et al*. (1994). Molecular cloning, sequence analysis, and expression of the yeast alcohol acetyltransferase gene. *Applied and Environmental Microbiology* **60**: 2786–2792.

Giorello F, Valera MJ, Martin V, *et al*. (2019). Genomic and transcriptomic basis of *Hanseniaspora vineae*'s impact on flavor diversity and wine quality. *Applied and Environmental Microbiology* **85**: 1–20.

Godoy L, Acuña-Fontecilla A, Catrileo D (2021). Formation of aromatic and flavor compounds in wine: a perspective of positive and negative contributions of non-*Saccharomyces* yeasts. In: *Chemistry and Biochemistry of Winemaking, Wine Stabilization and Aging* (Cosme F, Nunes FM, Filipe-Ribeiro L, eds). Vol. i. IntechOpen: 13.

Gonçalves C, Ferreira C, Gonçalves LG, *et al*. (2019). A new pathway for mannitol metabolism in yeasts suggests a link to the evolution of alcoholic fermentation. *Frontiers in Microbiology* **10**: 1–15.

Gonçalves C, Marques M, Gonçalves P (2022). Contrasting strategies for sucrose utilization in a floral yeast clade. *mSphere* **7**: e0003522.

Gonçalves C, Wisecaver JH, Kominek J, *et al*. (2018). Evidence for loss and reacquisition of alcoholic fermentation in a fructophilic yeast lineage. *eLife* **7**: 1–28.

Gonçalves P, Gonçalves C (2022). Horizontal gene transfer in yeasts. *Current Opinion in Genetics & Development* **76**: 101950.

Gonçalves P, Gonçalves C, Brito PH, *et al*. (2020). The *Wickerhamiella/Starmerella* clade – A treasure trove for the study of the evolution of yeast metabolism. *Yeast* **37**: 313–320.

Gregan J, Lindner K, Brimage L, *et al*. (2003). Fission yeast Cdc23/Mcm10 functions after pre-replicative complex formation to promote Cdc45 Chromatin Binding. *Molecular Biology of the Cell* **14**: 3876–3887.

Groenewald M, Hittinger CT, Bensch K, *et al*. (2023). A genome-informed higher rank classification of the biotechnologically important fungal subphylum *Saccharomycotina*. *Studies in Mycology* **105**: 1–22.

Gunhan M (2022). With advantages and disadvantages the role of non-*Saccharomyces* yeast in the wine industry. *Turkish Journal of Food Agriculture Sciences* **4**: 13–17.

Ianieva O, Podgorsky V (2021). Enological potential of non-*Saccharomyces* yeast strains of enological and brewery origin from Ukrainian collection of microorganisms. *Mycology* **12**: 203–215.

Jensen LJ, Julien P, Kuhn M, *et al*. (2008). eggNOG: Automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research* **36**: 250–254.

Jolly NP, Varela C, Pretorius IS (2014). Not your ordinary yeast: non-*Saccharomyces* yeasts in wine production uncovered. *FEMS Yeast Research* **14**: 215–237.

Kanehisa M, Sato Y (2020). KEGG Mapper for inferring cellular functions from protein sequences. *Protein Science* **29**: 28–35.

Kanehisa M, Sato Y, Kawashima M (2021). KEGG mapping tools for uncovering hidden features in biological data. *Protein Science* **31**: 47–53.

Katoh K, Rozewicki J, Yamada KD (2018). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* **20**: 1160–1166.

Kellis M, Birren BW, Lander ES (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**: 617–624.

Kosel J, Cadež N, Schuller D, *et al*. (2017). The influence of *Dekkera bruxellensis* on the transcriptome of *Saccharomyces cerevisiae* and on the aromatic profile of synthetic wine must. *FEMS Yeast Research* **17**: 1–11.

Kurtzman CP, Boekhout T, Fell JW (eds) (2011). *The Yeasts - A Taxonomic Study*, 5th edn. Elsevier, United Kingdom.

Kutyna DR, Borneman AR (2018). Heterologous production of flavour and aroma compounds in *Saccharomyces cerevisiae*. *Genes (Basel)* **9**: 326.

Legras JL, Galeote V, Bigey F, *et al*. (2018). Adaptation of *S. cerevisiae* to fermented food environments reveals remarkable genome plasticity and the footprints of domestication. *Molecular Biology and Evolution* **35**: 1712–1727.

Lemos Junior WJF, Duarte VS, Treu L, *et al*. (2018). Whole genome comparison of two *Starmerella bacillaris* strains with other wine yeasts uncovers genes involved in modulating important winemaking traits. *FEMS Yeast Research* **18**: 1–11.

Letunic I, Bork P (2019). Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research* **47**: 256–259.

Li Y, Steenwyk JL, Chang Y, *et al*. (2021). A genome-scale phylogeny of the kingdom *Fungi*. *Current Biology* **31**: 1653–1665.e5.

Lilly M, Lambrechts MG, Pretorius IS (2000). Effect of increased yeast alcohol acetyltransferase activity on flavor profiles of wine and distillates. *Applied and Environmental Microbiology* **66**: 744–753.

Liti G, Carter DM, Moses AM, *et al*. (2009). Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.

Liu Y, Rousseaux S, Tourdot-Maréchal R, *et al*. (2017). Wine microbiome: a dynamic world of microbial interactions. *Critical Reviews in Food Science and Nutrition* **57**: 856–873.

Maicas S, Mateo JJ (2023). The life of *Saccharomyces* and non-*Saccharomyces* yeasts in drinking wine. *Microorganisms* **11**: 1178.

Marcet-Houben M, Gabaldón T (2015). Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology* **13**: 1–26.

Martin V, Giorello F, Fariña L, *et al*. (2016). *De novo* synthesis of benzenoid compounds by the yeast *Hanseniaspora vineae* increases the flavor diversity of wines. *Journal Agriculture Food Chemistry* **64**: 4574–4583.

Mendes I, Sanchez I, Franco-Duarte R, *et al*. (2017). Integrating transcriptomics and metabolomics for the analysis of the aroma profiles of *Saccharomyces cerevisiae* strains from diverse origins. *BMC Genomics* **18**: 1–13.

Mikheenko A, Prjibelski A, Saveliev V, *et al*. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**: i142–i150.

Mira de Orduña R (2010). Climate change associated effects on grape and wine quality and production. *Food Research International* **43**: 1844–1855.

Morata A, Escott C, Bañuelos MA, *et al*. (2020). Contribution of non-*Saccharomyces* yeasts to wine freshness. A review. *Biomolecules* **10**: 34.

Moriya Y, Itoh M, Okuda S, *et al*. (2007). KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* **35**: 182–185.

Neto DJ da C, Morais PB de (2020). The vectoring of *Starmerella* species and other yeasts by stingless bees in a Neotropical savanna. *Fungal Ecology* **47**: 100973.

Nguyen LT, Schmidt HA, Von Haeseler A, *et al*. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 268–274.

Opulente DA, LaBella AL, Harrison M-C, *et al*. (2024). Genomic factors shape carbon and nitrogen metabolic niche breadth across *Saccharomycotina* yeasts. *Science* **384**: eadj4503.

Padilla B, Gil JV, 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**: 1–20.

Perli T, Wronska AK, Ortiz-Merino RA, *et al*. (2020). Vitamin requirements and biosynthesis in *Saccharomyces cerevisiae*. *Yeast* **37**: 283–304.

Peska V, Fajkus P, Bubeník M, *et al*. (2021). Extraordinary diversity of telomeres, telomerase RNAs and their template regions in *Saccharomycetaceae*. *Scientific Reports* **11**: 1–11.

Raymond Eder ML, Rosa AL (2021). Non-conventional grape varieties and yeast starters for first and second fermentation in sparkling wine production using the traditional method. *Fermentation* **7**: 321.

Rojas V, Gil JV, Piñaga F, *et al*. (2001). Studies on acetate ester production by non-*Saccharomyces* wine yeasts. *International Journal of Food Microbiology* **70**: 283–289.

Sabel A, Martens S, Petri A, *et al*. (2014). *Wickerhamomyces anomalus* AS1: a new strain with potential to improve wine aroma. *Annals of Microbiology* **64**: 483–491.

Santiago C, Rito T, Vieira D, *et al*. (2021). Improvement of *Torulaspora delbrueckii* genome annotation: towards the exploitation of genomic features of a biotechnologically relevant yeast. *Journal of Fungi* **7**: 287.

Santos J, Leitão-Correia F, Sousa MJ, *et al*. (2016). Nitrogen and carbon source balance determines longevity, independently of fermentative or respiratory metabolism in the yeast *Saccharomyces cerevisiae*. *Oncotarget* **7**: 23033–23042.

Sawyer SL, Cheng IH, Chai W, *et al*. (2004). Mcm10 and Cdc45 cooperate in origin activation in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **340**: 195–202.

Seixas I, Barbosa C, Salazar SB, *et al*. (2017). Genome sequence of the nonconventional wine yeast *Hanseniaspora guilliermondii* UTAD222. *Genome Announcement* **5**: 4–5.

Senik SV, Maloshenok LG, Kotlova ER, *et al*. (2015). Diacylglyceryltrimethylhomoserine content and gene expression changes triggered by phosphate deprivation in the mycelium of the basidiomycete *Flammulina velutipes*. *Phytochemistry* **117**: 34–42.

Simão FA, Waterhouse RM, Ioannidis P, *et al*. (2019). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210−3212.

Shen X-X, Steenwyk J, LaBella A, *et al*. (2020). Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum *Ascomycota*. *Science Advances* **4**: eabd0079.

Shen XX, Opulente DA, Kominek J, *et al*. (2018). Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**: 1533−1545.

Silva-Sousa F, Fernandes T, Pereira F, *et al*. (2022). *Torulaspora delbrueckii* phenotypic and metabolic profiling towards its biotechnological exploitation. *Journal of Fungi* **8**: 569.

Silva M, Pontes A, Franco-Duarte R, *et al*. (2022). 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.

Stanke M, Morgenstern B (2005). AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* **33**: 465–467.

Stratford M, Plumridge A, Archer DB (2007). Decarboxylation of sorbic acid by spoilage yeasts is associated with the PAD1 gene. *Applied and Environmental Microbiology* **73**: 6534–6542.

Sun N, Jang J, Seunghee L, *et al*. (2005). The first two-dimensional reference map of the fission yeast, *Schizosaccharomyces pombe* proteins. *Proteomics* **5**: 1574–1579.

Sun Y, Huang J, Wang Z, *et al*. (2022). Identification of microproteins in *Saccharomyces cerevisiae* under different stress conditions. *Journal of Proteome Research* **21**: 1939–1947.

Tiukova IA, Prigent S, Nielsen J, *et al*. (2019). Genome-scale model of *Rhodotorula toruloides* metabolism. *Biotechnology and Bioengineering* **116**: 3396–3408.

Tondini F, Lang T, Chen L, *et al*. (2019). Linking gene expression and oenological traits: comparison between *Torulaspora delbrueckii* and *Saccharomyces cerevisiae* strains. *International Journal of Food Microbiology* **294**: 42–49.

Verstrepen KJ, Van Laere SDM, Vanderhaegen BMP, *et al*. (2003). Expression levels of the yeast alcohol acetyltransferase genes *ATF1*, *Lg-ATF1*, and *ATF2* control the formation of a broad range of volatile esters. *Applied and Environmental Microbiology* **69**: 5228–5237.

Vieira D, Esteves S, Santiago C, *et al*. (2020). Population analysis and evolution of *Saccharomyces cerevisiae* mitogenomes. *Microorganisms* **8**: 1001.

Walker ME, Nguyen TD, Liccioli T, *et al*. (2014). Genome-wide identification of the Fermentome; genes required for successful and timely completion of wine-like fermentation by *Saccharomyces cerevisiae*. *BMC Genomics* **15**: 1–17.

Whitener MEB, Stanstrup J, Panzeri V, *et al*. (2016). Untangling the wine metabolome by combining untargeted SPME–GCxGC-TOF-MS and sensory analysis to profile Sauvignon blanc co-fermented with seven different yeasts. *Metabolomics* **12**: 1–25.

Yang D-D, de Billerbeck GM, Zhang J-J, *et al*. (2018). Deciphering the origin, evolution, and physiological function in the yeast *Saccharomyces cerevisiae*. *Applied and Environmental Microbiology* **84**: 1–16.

Yu X, Huang T, Huang Z, *et al*. (2023). Effects of six commercially available koji (Chinese Xiaoqu) on the production of ethyl acetate, ethyl lactate, and higher alcohols in Chinese Baijiu (distilled spirit) brewing. *Heliyon* **9**: e17739.

Zang J, Yu D, Li T, *et al*. (2022). Identification of characteristic flavor and microorganisms related to flavor formation in fermented common carp (*Cyprinus carpio* L.). *Food Research International* **155**: 111128.

Zhang H, Yohe T, Huang L, *et al*. (2018). DbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **46**: W95–W101.

Supplementary Material: https://studiesinmycology.org/

The supplementary tables are deposited at the Figshare repository: https://doi.org/10.6084/m9.figshare.28278929.

Table S1. Quality control of the 661 genomes considered in the present study. Dark Grey - genomes excluded from further analysis. Red - particular parameters used to exclude the genomes (see methods for details). Light Grey - genomes further excluded due to lack of quality in the subsequent analysis. (provided as a separate Excel file)

Table S2. Taxonomic information of the 134 yeast species considered in this work, in particular regarding discrepancies found in different taxonomic databases, and the aggregation of different synonyms under the same accepted name.

Table S3. Comparison of number of Protein Coding Sequences (CDS) after genome annotation with our pipeline, in comparison with annotations available in NCBI and in literature. Numbers in red indicate the highest number of genes identified in the total number of annotations compared. Red background additionally indicates novel annotation done in this work for that particular strain.

Table S4. Functional annotation of the 530 non-*Saccharomyces* proteomes considered in this study (see methods for details).

Table S5. Number of genes (average of all the strains of each species (first row) ± standard deviation (second row)) annotated into each KEGG Pathway, organized by taxonomic level

Table S6. Heatmap analysis of the KEGG Pathway modules completeness in the 530 non-*Saccharomyces* genomes.

Table S7. Heatmap analysis of the KEGG Pathway modules completeness in the different yeast families (numbers on the third row), orders (second row), and phyla (first row). * - families with only one species represented

Table S8. Information about the 93 genes identified as the "Fermentome of *S. cerevisiae*", according with Walker et al. 2014, and their presence in the genome of 530 non-*Saccharomyces* yeasts. Blue cells - gene present; Grey cells - absence of the gene.

Table S9. BLAST results of S. cerevisiae proteins in the non-*Saccharomyces* genomes. ok - homologous protein detected (Blast e-value > e-6). absent - protein not detected (Blast e-value < e-6).

Table S10. 41 genes defined as the Fermentome of non-*Saccharomyces* yeasts within the present work. Genes highlighted in orange represent the ones present in more than 95% of fermentative strains, and in less than 5% of non-fermentative strains (script Fermentome). The remaining ones, highlighted in yellow, were detected in more than 90% of fermentative strains, and in less than 10% of non-fermentative strains.

Table S11. Presence of the 96 genes in the 530 non-*Saccharomyces* genomes, selected from the literature as being relevant for aroma and flavour production in wine. ok - gene identified in the genome; Empty grey cell - gene not identified. Identification of fermentative yeast species: red - no; orange - variable; green – yes.