FEMS Yeast Research, 2026, **26**, foag057
https://doi.org/10.1093/femsyr/foag057
Published: 29 October 2026
Research article – Industrial and environmental biotechnology

Whole genome sequencing, assembly, and functional annotation of *Hyphopichia burtonii* Y951 isolated from *Vitis vinifera* L. cv Cabernet Sauvignon grape must

Evelyn Maluleke and Mathabatha Evodia Setati *

South African Grape and Wine Research Institute, Stellenbosch University, Private Bag X1, Stellenbosch, 7599, South Africa
*Corresponding author. South African Grape and Wine Research Institute, Stellenbosch University, JH Neethling Building, Victoria Street, Stellenbosch, 7600, South Africa. E-mail:setati@sun.ac.za
Editor: Francisco Cubillos

Abstract

***Hyphopichia burtonii* is a non-conventional yeast present in diverse environments, including bread, cured meat, grape must, zebra dove droppings, and corn weevil. Strains derived from different environments display different traits. However, limited information exists on the genomic diversity of this species. Here, we present the genome of *H. burtonii* Y951 strain isolated from grape must and proven to have antifungal activity against various grapevine fungal pathogens. The genome was sequenced using PacBio long-read technology, and a high-quality de novo genome assembly was generated using the Flye assembler. The genome was found to be 12.4 Mb, with a GC content of 34.96% and 99.3% completeness based on BUSCO assessment. Gene prediction identified a total of 5249 protein-encoding genes, which were functionally classified using gene ontology terms, revealing clusters of genes grouped by their biological functions. The data further reveal diverse genes involved in carbohydrate metabolism, fermentation, and aroma compounds production, as well as a repertoire of biosynthetic gene clusters relevant in fungal interactions. Genome rearrangements provide the basis for intraspecific variability in some metabolic traits. This comprehensive genomic characterization provides valuable insight into the molecular basis of *H. burtonii's* biology and establishes a resource for future functional and applied research.**

Keywords *De novo* assembly, genome assembly, genome annotation, terpene clusters, NRPs-like clusters, Serinales

Background

Hyphopichia burtonii (formerly *Pichia burtonii*) is a non-conventional yeast species belonging to the family Debaryomycetaceae, within the order Serinales (formerly referred to as the CUG-Ser clade). It is a ubiquitous yeast, widely distributed across various natural and anthropogenic environments (Kurtzman et al. 2011). Indeed, *H. burtonii* has been isolated from substrates with low water activity such as bealmijang meju (Kim et al. 2011), Japanese dry noodles (Ohisa et al. 2012), stored white rice (Lee et al. 2019), bread (Garcia et al. 2019), and cured meats (Simoncini et al. 2007), as well as other environments, including grape must (Ghosh et al. 2015), corn weevil (Seco et al. 2024), dairy (Groenewald and Smith 2010), Chamroensakchai et al. 2021), zebra dove droppings (Indoung et al. 2021), marine environments (Gadanho and Sampaio 2005), and fish feed (Pinheiro et al. 2018). Previous studies have shown that different *H. burtonii* isolates can display contrasting phenotypes depending on their source and ecological contexts. Isolates from bread have been associated with spoilage, whereas strains recovered from grape must, corn weevil, and zebra dove droppings have been reported to inhibit grape spoilage fungi, corn-associated mycotoxigenic fungi, and undesirable *Penicillium* species on maturing hams, respectively (Virgili et al. 2012, Maluleke et al. 2022, Seco et al. 2024). In addition, *H. burtonii* is known for its tolerance to high-salt environments and its metabolic flexibility (Lee et al. 2021).

In our previous work, *H. burtonii* Y951 demonstrated desirable biocontrol potential, with antifungal activity against various phytopathogens including three strains of *Botrytis cinerea*, one of *Aspergillus niger*, and one of *Alternaria alternata* (Maluleke et al. 2022). In addition, *H. burtonii* Y951 showed high levels of cell wall-degrading enzymes, particularly chitinases and β-1,3-glucanases, when co-cultured with *B. cinerea*. These cell wall degrading enzymes may contribute to the proposed biocontrol mechanism by compromising fungal cell-wall integrity and restricting pathogen growth. The fermentation and oenological traits of *H. burtonii* Y951 have also been evaluated in monoculture and sequential inoculation with *Saccharomyces cerevisiae* (Rollero et al. 2018). Broadly, the strain was found to be a poor fermenter, with low consump-

Received: 8 April 2026. **Revised:** 14 October 2026. **Accepted:** 20 October 2026
© The Author(s) 2026. Published by Oxford University Press on behalf of FEMS. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site-for further information please contact journals.permissions@oup.com

tion of amino acids such as valine, leucine, and phenylalanine, as well as ammonium, but was associated with a higher production of higher alcohols and increased synthesis of ethyl esters that impart fruity flavours to wine (Rollero et al. 2018). Overall existing data suggest that this strain can contribute positively to grape and wine quality by suppressing grape phytopathogens and enhancing the organoleptic profile of wines without competing with *S. cerevisiae*. Both the studies that evaluated the biocontrol potential and oenological traits of *H. burtonii* Y951 were conducted separately from genome sequencing. The present study therefore builds on, but does not reproduce, the earlier biocontrol or fermentation experiments and rather focuses on the genomic sequencing, functional annotation, and comparative analysis of *H. burtonii* Y951.

Genomic data on *H. burtonii* remain limited, however available data on few different strains of *H. burtonii* indicate Its genome size is ˜12.3–12.5 Mb, with a GC content of around 35%, and contains about 2100 predicted protein-coding genes that are homologous to other yeasts within the order Serinales (Jeong et al. 2023, Indoung et al. 2025). Comparative genomics suggest that *H. burtonii* shares ˜85% nucleotide identity with closely related species such as *H. pseudoburtonii*, exhibiting chromosomal rearrangements indicative of ongoing divergence (Lee et al. 2021, Jeong et al. 2023).

Functionally, the genome encodes enzymes involved in carbohydrate metabolism, stress response mechanisms, and secondary metabolite production, all contributing to its survival under high salinity and osmotic stress conditions (Lee et al. 2021). Transcriptomic data reveal that *H. burtonii* responds rapidly to osmotic stress by upregulating genes related to ribosome biogenesis, transmembrane transport, and glycerol biosynthesis, with constitutive activation of the Hog1 MAP kinase pathway, which enhances its salt tolerance (Lee et al. 2021).

Additionally, the species possesses a diverse array of transporter proteins, including ABC transporters and amino acid permeases, facilitating ion homeostasis and osmoprotection. Enzymes involved in carbohydrate utilization and secondary metabolism further support its adaptation to saline environments. Genomic resources for *H. burtonii* are publicly available through repositories such as Ensembl Fungi and NCBI, providing a basis for further exploration of its unique osmotolerance mechanisms (Jeong et al. 2023). This study aimed to generate and analyse the first whole-genome sequence and de novo assembly of the grape must strain *H. burtonii* Y951, thereby establishing a genomic framework for investigating its biology, ecological adaptation, and broader functional potential, including traits relevant to biocontrol and fermentation. In addition to biocontrol-related traits, fermentative and flavour-associated metabolic potential was also considered as a secondary functional aspect given its isolation from grape must.

Although the biocontrol phenotype of *H. burtonii* Y951 was demonstrated experimentally in our previous study (Maluleke et al. 2022), the present work is based on genome sequencing, annotation, and comparative genomic analyses. Accordingly, the genomic features identified here should be regarded as candidate determinants that may be associated with previously observed biocontrol activity and do not establish causal or functional relationships with the observed phenotype.

Methods

Yeast strain

Hyphopichia burtonii Y951 isolated from *Vitis vinifera* L. cv Cabernet Sauvignon grape must (Ghosh et al. 2015) was sourced from the culture collection of the South African Grape and Wine Research Institute at Stellenbosch University. The strain was subcultured onto Wallerstein Nutrient agar (Merck Millipore, South Africa) to confirm culture purity and for routine short-term storage. For long-term preservation, the strain was stored at −80°C in cryogenic vials in a final concentration of 25% (v/v) glycerol as a cryoprotectant.

DNA extraction and PacBio HiFi genome sequencing

Genomic DNA was extracted using the Quick-DNA Miniprep Kit (Zymo Research, USA) according to the manufacturer's instructions. DNA quality and quantity were assessed at Inqaba Biotechnical Industries (Pty) Ltd (Pretoria, South Africa) using a Qubit Fluorometer with the Qubit 1X dsDNA HS Assay Kit (Thermo Fisher Scientific) for concentration, and an Agilent Femto Pulse system for size distribution, ensuring that most DNA fragments exceeded 50 kb. Purity was verified by spectrophotometric measurement of the A260/A280 ratio. For library preparation, 5 μg of high-molecular weight genomic DNA was diluted in 1X Elution buffer to a concentration of 83.3 ng/μl in a final volume of 60 μl, then sheared to a target fragment size of 15–18 kb using the Megaruptor 3 system (Diagenode) with a two-cycle protocol (speed settings 31 and 32). SMRTbell libraries were constructed using the SMRTbell Express Template Prep Kit 2.0 (PacBio), following the manufacturer's workflow: removal of single-strand overhangs (incubation at 37°C for 15 min), DNA damage repair (37°C for 30 min), end-repair and A-tailing (20°C for 10 min, then 65°C for 30 min), and ligation of Overhang Adapter v3. Following ligation, libraries were purified with 1X AMPure PB beads and size-selected using the PippinHT system (Sage Science) with 0.75% agarose gel cassettes to enrich for fragments within the desired range. The final libraries were again purified with AMPure PB beads and quantified prior to sequencing. Sequencing was performed on a PacBio Sequel II platform using the Sequel II Sequencing Kit 2.0 and SMRT Cell 8 M, with circular consensus sequencing mode to generate HiFi long reads. This approach yielded mean read accuracies exceeding 99%, with read lengths suitable for de novo genome assembly applications. The obtained reads were processed and analysed using the pipeline described in Fig. 1.

Genome assembly

De novo assembly strategies were employed for fungal genome reconstruction. Raw PacBio HiFi sequencing data was processed using the Galaxy Europe platform (https://usegalaxy.eu/). Quality control of raw reads was performed using FastQC (Andrews 2010) to assess read quality metrics. Input FASTQ files were assembled using Flye v2.9 (Lin et al. 2016) assembly software with default parameters optimized for HiFi reads, generating a draft genome assembly. Post-assembly polishing was conducted using the original HiFi reads to correct potential assembly errors. Assembly quality

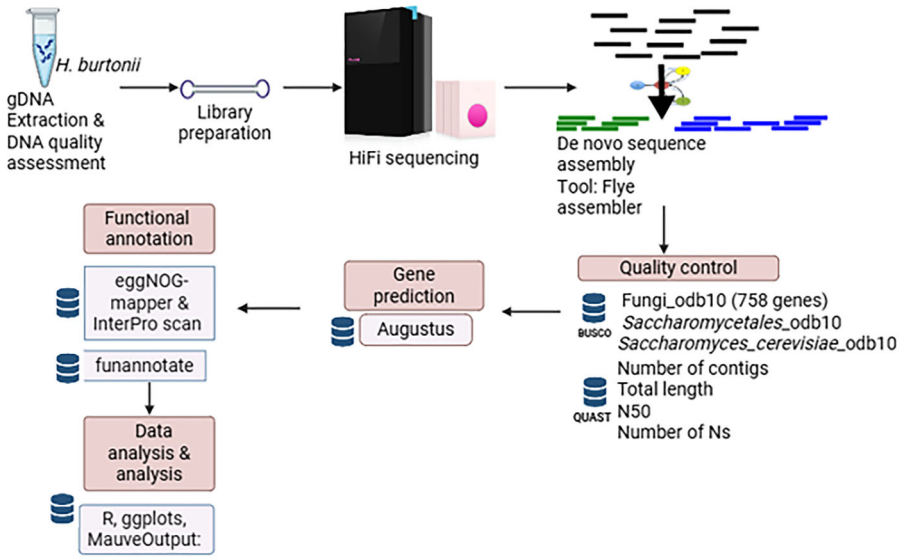


Figure 1 Bioinformatic pipeline used to analyse the genome of *Hyphopichia burtonii* Y951. The schematic representation of the pipeline includes all the quality control checkpoints used throughout the analysis, together with the software and the intermediate datasets obtained. For details about each software and parameters, see the methods and materials section.

was comprehensively evaluated through multiple analytical approaches. QUAST v5.2.0 (Mikheenko et al. 2018) was utilized to assess fundamental assembly metrics, producing detailed reports containing key statistics, including total assembly size, N50 values, L50 statistics, GC content distribution, and contig enumeration. Genome assembly completeness was rigorously evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.6.1 (Simão et al. 2015), employing the fungi_odb10 and *saccharomycetes*_odb10 database under default genome mode settings to assess the presence and completeness of conserved orthologous genes expected in fungal genomes.

Genome annotation

Gene prediction and structural annotation were performed using AUGUSTUS (Stanke and Morgenstern 2005) , using *Saccharomyces cerevisiae* S288C and *Candida albicans* as reference gene models. These species were selected as well-annotated yeast models commonly used for gene prediction in fungal genomes, although they are phylogenetically distant from *H. burtonii* and may introduce bias due to its placement within the order Serinales. The predicted gene structures were generated in General Feature Format Version 3 (GFF3). Repeat elements were identified and masked using RepeatMasker (Smit et al. 2013) prior to gene prediction to improve annotation accuracy. Comprehensive functional annotation was conducted using the Funannotate v1.8.17 pipeline (Palmer and Stajich 2020), following the standard clean-sort-mask workflow as specified in the Funannotate user documentation. The annotation process incorporated multiple evidence sources through downstream Funannotate predict and update commands, integrating *ab initio* gene predictions, homology-based evidence, and protein domain identification. Final annotations were generated in both GenBank and GFF3 formats with functional assignments linked to established databases, including UniProt (The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025), Pfam (Mistry et al. 2021), and gene ontology (GO) terms (Ashburner et al. 2000). Secondary metabolite gene cluster prediction was performed using the antibiotics and Secondary Metabolites Analysis SHell (antiSMASH) v7.1.0 implemented through the Galaxy Europe web platform, using the default settings (Medema et al. 2011) with detection strictness configured to relaxed settings according to default parameters, enabling identification of both well-characterized and novel biosynthetic gene clusters (BGCs). Quality assessment of predicted protein sequences was conducted using BUSCO analysis employing the *Saccharomycetales* and fungi_odb10 database (Manni et al. 2021) to evaluate annotation completeness and accuracy. Non-coding RNAs (ncRNAs) were identified using specialized tools optimized for yeast genomes via the Europe Galaxy platform (usegalaxy.eu). tRNAs were predicted using the tRNAscan/EuFinDtRNA/tRNA covariance model hybrid (Lowe and Eddy 1997; Chan et al. 2021), rRNAs using Barrnap v0.9 (Seeman 2013), and other ncRNAs using cmscan (Nawrocki and Eddy 2013) against the Rfam covariance model database (September 2023 release). The *H. burtonii* genome assembly (12 420 859 bp across 9 contigs) was analyzed using default parameters via the Galaxy web interface. High-quality alignments were selected using stringent filtering criteria: tRNAs (Cove score >50), rRNAs (*E*-value $<$1e-5), cmscan (pass $=1$, inc $=$ '!').

Additional functional characterization was performed using EggNOG-mapper v2.1.4–2 (Huerta-Cepas et al. 2016) for orthology assignment and functional classification, providing Clusters of Orthologous Groups (COG) of proteins (Galperin et al. 2025) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations. Genome analysis and manipulation were facilitated using various utilities from the GenomeTools v1.6.1 package (Gremme et al. 2013), with all parameters not explicitly specified maintained at their respective default settings. Carbohydrate active enzymes (CAZymes) were predicted on the dbCAN3 server (Zheng et al. 2023) using HMMER and DIAMOND tools, with the HMMER:dbCAN

and HMMER:dbCAN-sub set to filter hits with *e*-values below 10–5 and the DIAMOND:CAzy set at e value 10–102. Only hits supported by at least two of the three tools were retained. The predicted proteins were classified according to the CAZyme database into the following families: auxiliary activities (AA), carbohydrate-binding modules (CBM), carbohydrate esterases (CE), glycoside hydrolases (GH), and glycosyltransferase (GT). Protein sequences of CAZymes families reported to be associated with biological control were compared with protein sequences available on UniProt using BlastP and their homology with those of selected biocontrol yeasts were recorded. In addition, comparative analysis of the CAZyme repertoire of *H. burtonii* Y951 was performed using publicly available CAZy annotations from selected biocontrol and fermentation-associated yeast-like fungus (*Aureobasidium pullulans*), and yeasts (*Metschnikowia fructicola*, *Metschnikowia pulcherrima*, *Candida oleophila*, and *Wickerhamomyces anomalus*). Overall CAZyme family composition and the copy numbers of key lytic enzyme families involved in fungal cell wall degradation (GH5, GH16, GH17/GH55, GH72, and GH18) were compared across species.

Comparative genome analysis

Comparative genome analysis between *H. burtonii* Y951 and *H. burtonii* NRRL Y-1933 was performed to assess differences in assembly contiguity and overall genome organization. *H. burtonii* NRRL Y-1933 was used as the reference genome for all comparative alignment and synteny analyses presented in this study. Assembly statistics, including contig number, N50, genome size, and BUSCO completeness, were calculated to contextualize differences between the two genomes. Pairwise genome alignment was conducted using ProgressiveMauve (version 2.4.0) to visualize conserved genomic regions between assemblies. The alignment was used for graphical representation of conserved syntenic blocks, and outputs were exported as high-resolution figures from the Mauve interface.

Gene level synteny analysis

To assess genomic correspondence between the *H. burtonii* strain Y951 assembly and the *H. burtonii* NRRL Y-1933 reference genome, predicted proteins from strain Y951 were mapped against the NRRL Y-1933 reference proteins. For each retained match, the corresponding reference accession genomic coordinates and the strand orientation were recorded and duplicate records were removed. The resulting mapping dataset was processed in Galaxy Europe platform using Unique (Galaxy Version 9.11 + galaxy0) and sort (Galaxy Version 9.11 + galaxy0). Query Tabular (Galaxy Version 3.3.2) was used to determine the distribution of mapped genes among reference genomic sequences, the predominant reference sequence for each Y951 contig, the proportion of genes mapping to that sequence, and gene-order and strand consistency. Gene order was assessed by comparing the relative positions of mapped genes on the NRRL Y-1933 reference with their order on the Y951 contigs. Since the Y951 genome was assembled from PacBio HiFi reads using Flye and remains at the contig level, the analysis was interpreted as gene-level/local syntenic correspondence, rather than chromosome scale synteny or structural variation analysis.

Analysis of genes involved in the fermentome and flavourome

Previous research evaluating the fermentation performance of *H. burtonii* Y951 revealed that this yeast is a weak fermenter, with low nitrogen consumption and considerable contribution to phenyl ethanol, isoamyl alcohol, and ethyl esters in wine and low production of short and medium chain fatty acids (Rollero et al. 2018b). Referring to the recent work by (Franco-Duarte R et al. 2025), we evaluated the genome after functional annotation for the presence of genes defined as the putative fermentome and flavourome of non-*Saccharomyces* yeasts.

Data visualization

Genomic data visualization was performed using multiple specialized tools and platforms. Assembly statistics and quality metrics were visualized using QUAST-generated plots and custom R scripts employing ggplot2 for statistical graphics. Pairwise genome alignment visualization was performed using ProgressiveMauve (version 2.4.0) as described above. The resulting alignment was visualized using Mauve's graphical interface and exported as high-resolution figures. Functional annotation results were summarized using bar charts and pie charts generated in R, showing distributions of functional categories, pathway assignments, and protein domains.

Results

Genome assembly

In the present study, a high-quality genome assembly of *H. burtonii* Y951 was generated using long-read sequencing data obtained from the Pacific Biosciences (PacBio) revio platform. A total of 385 097 HiFi Reads with a mean length of 6388 bp were generated (see Supplementary Table 1). The raw sequencing data have been deposited in the National Centre for Biotechnology Information (NCBI) Sequence Read Archive (SRA) run accession SRR36539643 under BioProject accession PRJNA1391825 and BioSample number SAMN54234055. The genome assembly is available in GenBank under accession number SUB16549568.

The raw HiFi PacBio reads were assembled de novo using the Flye assembler. The genome assembly of *H. burtonii* Y951 resulted in a total length of 12.42 Mb distributed across 9 contigs, with no gaps (0 Ns/100 kb), indicating a highly contiguous assembly. Since no chromosomal-level assignment data were generated, the 9 contigs should be interpreted as contigs rather than definitive chromosomes The largest contig is 2.75 Mb, and the N50 and L50 values are 1.86 Mb and 3 (Table 1), respectively, demonstrating a high level of continuity. The GC content of the assembly is 34.96%, which is consistent with other yeasts in the order Serinales.

The completeness assessment using BUSCO (*saccharomycetes*_odb10 dataset, 2137 BUSCOs) revealed 99.3% completeness, with 99.1% of the BUSCO genes found as single copies and only 0.2% duplicated BUSCOs (Table 1). Only 0.1% fragmented and 0.7% missing BUSCO genes were detected, indicating that the assembly is nearly complete with minimal gaps or missing genomic regions.

Table 1 An overview of the assembly statistics and genome features for *H. burtonii* Y951.

Metric	Value
Assembly size (bp)	12 420 859
Number of contigs	9
Largest contig (bp)	2 746 947
GC content (%)	34.96
N50 (bp)	1 861 796
L50	3
BUSCO completeness (%)	99.3 [S: 99.1, D: 0.2]

In this study, we used infernal tools and stringent filtering criteria to identify non-coding sequences. The hybrid approach revealed 241 ncRNA loci totalling 37 605 bp (0.303% of genome length), dominated by tRNAs, followed by rRNAs and snoRNAs (Table 2). Among the rRNAs, 5S rRNA had the highest copy number. This analysis provides an overview of predicted ncRNA features that may be involved in gene regulation and essential cellular processes. However, these annotations are based on sequence prediction and do not demonstrate ncRNA expression or biological activity. Therefore, functional conclusions about the identified ncRNAs cannot be drawn without experimental validation.

Assembly comparison showed clear differences in contiguity between the two genomes (Table 3). The Y951 assembly was highly contiguous (9 contigs; N50 = 2.1 Mb), while NRRL Y-1933 was more fragmented (153 contigs; N50 = 180 kb). Both genomes had similar genome sizes and BUSCO completeness (Table 3), reflecting differences in sequencing and assembly strategies.

Comparative genome alignment between *H. burtonii* Y951 and NRRL Y-1933 revealed conserved syntenic blocks across large portions of both genomes, possibly indicating overall conservation of gene order between the two assemblies (Fig. 2). The alignment showed that major genomic regions could be aligned into locally collinear blocks, suggesting broad correspondence of genomic architecture between the strains. However, the continuity and arrangement of some blocks differed between the two genomes, which is likely influenced by differences in assembly contiguity and fragmentation, particularly in the more fragmented NRRL Y-1933 assembly. These differences affect the visual continuity of aligned regions and limit interpretation of fine-scale genome organization. Overall, the Mauve alignment (Fig. 2) provides a qualitative representation of shared genomic structure between the two assemblies. To the best of our knowledge, this represents the first high quality de novo genome assembly for *H. burtonii*. However, the absence of publicly available long-read sequencing data from additional phenotypically distinct strains currently limits broader comparative analyses.

To further quantify the genomic correspondence observed between the *H. burtonii* Y951 contigs and *H. burtonii* NRRL Y-1933 reference sequences, gene mapping was performed. Seven contigs showed >96% of mapped genes assigned to a single dominant reference genome sequence, whereas contigs 4 and 5 showed lower correspondence 64.36% and 68.67%, respectively (Supplementary Table 2). Both forward and reverse orientations were observed, supporting substantial local gene-level syntenic correspondence between the Y951 assembly and the NRRL Y-1933 reference.

Genome annotation and functional classification of predicted genes

Gene prediction was conducted using two reference yeasts, *S. cerevisiae* and *C. albicans* to determine which training model produced the most complete annotation for *H. burtonii.* These species were selected because they are well-annotated yeast references commonly used in gene prediction pipelines, although we acknowledge that their evolutionary distance from *H. burtonii* may introduce bias in gene structure prediction. The highest number of protein-encoding genes (5249) were predicted in the *H. burtonii* genome using *S. cerevisiae* as the reference (Table 4) and this annotation set was retained for downstream analyses. Compared to *S. cerevisiae*, only 4398 protein encoding genes were predicted when using *C. albicans* as the reference genome. These genes were shared with those predicted with *S. cerevisiae* as a reference. Moreover, using *C. albicans* allowed for prediction of 682 that were unique protein encoding genes not detected with the *S. cerevisiae* reference (Supplementary Table 3).

Of the 5249 genes predicted with *S. cerevisiae* as the reference, 5100 genes (97.16%) showed homology to entries in the non-redundant protein (nr) database, indicating a high degree of annotation completeness and confidence in gene identification. Functional classification using EggNOG identified 5098 genes (97.10%), while 4484 genes (85.43%) were associated with the InterPro (IPR) database, revealing conserved protein domains and motifs. Pfam domain analysis revealed 4107 genes (78.24%) with recognizable domains, such as protein kinases or transmembrane domains, further supporting their functional roles. GO terms were attributed to 4089 genes (77.89%), categorizing genes based on biological processes, cellular components, and molecular functions (Table 4).

Further pathway analysis using the KEGG mapped 2512 genes (47.90%) to known metabolic pathways. Examples include glycolysis, the citric acid cycle, and amino acid biosynthesis pathways. COG classification assigned 4000 genes (76.20%) to functional categories, such as 'translation, ribosomal structure and biogenesis' or 'carbohydrate transport and metabolism'. Specialized functional databases identified 59 transcription factor (TF) genes (1.12%) in the TF database, 139 genes (2.65%) in the carbohydrate-active enzymes (CAZy) database (e.g. glycoside hydrolases and glycosyltransferases), and 185 peptidase genes (3.52%) in the MEROPS database (Table 4).

Genes annotated were assigned to a broad range of COG functional categories, including both individual and combined categories (Fig. 3). The largest number of genes, 1222 fell into category S, representing genes of unknown function. This substantial fraction highlights the considerable fraction of the genome that remains functionally uncharacterized, underscoring the potential for discovering novel biological roles. The next most abundant categories were posttranslational modification, protein turnover, and chaperones (O) and intracellular trafficking, secretion, and vesicular transport (U), each comprising 405 genes, which suggests a strong capacity for protein processing and secretion. These

Table 2 Statistics of different types of ncRNA in the *H. burtonii* Y951 genome.

RNA Type	Tool	Copies	Avg length (bp)	Total length (bp)	% genome
tRNAs	tRNAscan hybrid v2.0.12	120	72	8640	0.070
5S rRNA	Barrnap v0.9	39	119	4641	0.037
5.8S rRNA	Barrnap v0.9	4	156	624	0.005
18S rRNA	Barrnap v0.9	3	1800	5400	0.043
28S rRNA	Barrnap v0.9	3	3400	10 200	0.082
snoRNA	cmscan (Rfam 14.9)	50	120	5880	0.047
snRNA	cmscan (Rfam 14.9)	5	92	1561	0.013
mi RNA	cmscan (Rfam 14.9)	17	132	659	0.005
Total	–	241	–	37 605	0.303

Table 3 Comparison of genome assembly statistics for *H. burtonii* genomes.

Metric	*H. burtonii* Y951	*H. burtonii* NRRL Y-1933
Sequencing technology	PacBio HiFi	Illumina HiSeq
Assembly method	Flye	AllPaths-LG/SPAdes
Number of contigs	9	243
Contig N50	2.1 Mb	114.1 kb
Contig L50	3	34
Genome size	11.2 Mb	13.1 Mb
BUSCO completeness (%)	99.2	96.8
GC content (%)	34.96	35.0

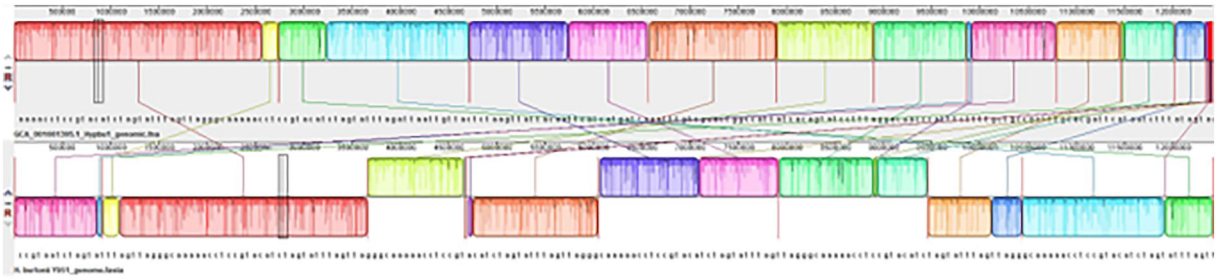


Figure 2 Mauve alignment of GCA_001661395.1_*Hypbu1* (top) and *H. burtonii* Y951 (bottom) genomes. Coloured blocks represent locally colinear homologous regions shared between the two genomes; blocks connected by the same colour indicate aligned syntenic segments. Blocks shown in the same orientation reflect conserved gene order, whereas blocks that appear inverted indicate regions of rearrangement. Crossing alignment lines denote inversions and translocations.

were followed by transcription (K), with 361 genes, reflecting extensive regulatory capacity. Genes involved in translation, ribosomal structure, and biogenesis (J; 358 genes), amino acid transport and metabolism (E; 302 genes), and carbohydrate transport and metabolism (G; 270 genes) were also well represented, indicating robust metabolic versatility and capacity for protein synthesis. Additionally, 266 genes were associated with RNA processing and modification (A) and 265 genes with signal transduction mechanisms (T), suggesting a significant capacity for environmental responsiveness. The remaining categories exhibited progressively fewer genes, ranging from 232 genes involved in energy production and conversion (C) to a single gene related to cell motility (N).

GO analysis of 2831 genes which could be assigned gene names based on homology to *S. cerevisiae* genes revealed that most of the genes could be categorized into various biological processes. In particular, five cellular processes, including metabolic process, organism substance metabolic process, cellular metabolic process, primary metabolic process, and nitrogen compound metabolic process were the most significantly enriched (Fig. 4a). The data further show that ˜1600 of the genes were enriched in 10 molecular functions dominated by binding activity, followed by catalytic activity, organic cyclic compound, heterocyclic compound, and ion binding (Fig. 4b). The most enriched GO terms in cellular components were cellular anatomical entity and intracellular anatomical structure while intracellular non-membrane-bounded organelle was the least enriched (Fig. 4c). KEGG enrichment analysis revealed 11 pathways among which the metabolic pathways as well as biosynthesis of secondary metabolites were the most enriched (Fig. 4d).

Table 4 Annotation of the predicted genes using a variety of databases with *S. cerevisiae* S288C and *C. albicans* models as reference.

Database	Full name	Count using *S. cerevisiae* S288C	%	Count using *C. albicans*	%
Total	Predicted genes	5249	100	4398	100
nr	Non-redundant protein database	5100	97.16	–	–
EggNOG	Functional annotation (EggNOG)	5098	97.10	0	0.0
IPR	InterPro database	4484	85.43	0	0.0
Pfam	Pfam protein families	4107	78.24	3514	79.9
GO	Gene ontology	4089	77.89	0	0.0
KEGG	Kyoto Encyclopedia of Genes and Genomes	2512	47.90	–	–
COG	Cluster of Orthologous Groups	4000	76.20	–	–
TF	Transcription factor database	59	1.12	–	–
CAZy	Carbohydrate-active enzymes database	139	2.65	118	2.7
MEROPS	Peptidase database	185	3.52	168	3.8

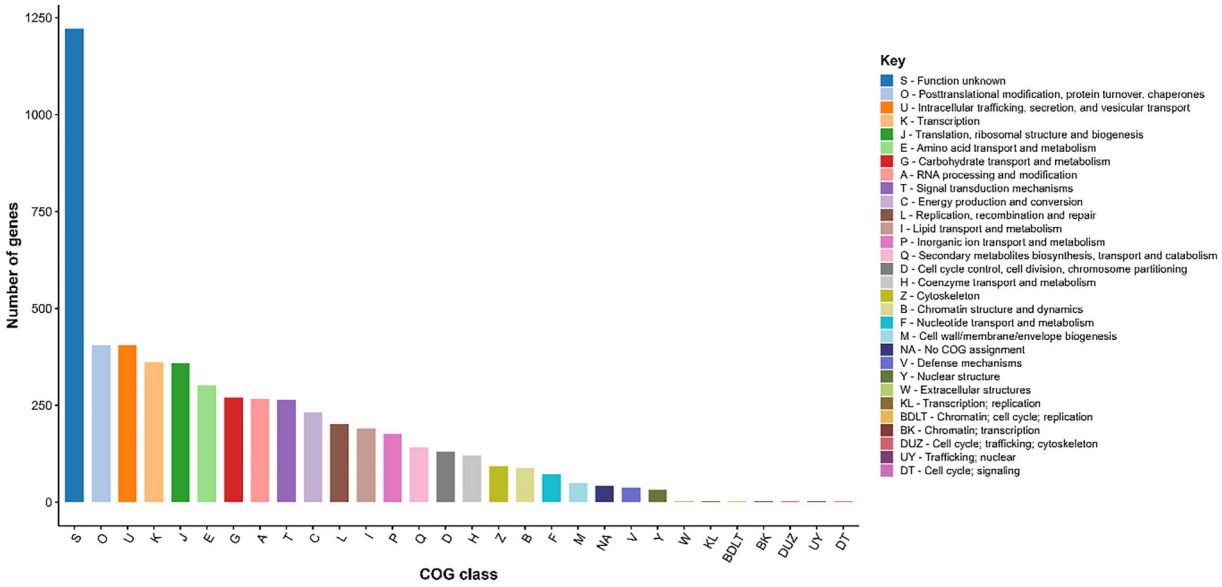


Figure 3 Functional annotation of predicted genes based on COG categories combining the genes annotated using *S. cerevisiae* reference genome with those annotated and unique to *C. albicans* reference genome. Each bar represents the combined number of genes assigned to specific functional category. Categories with a single gene are not shown.

Predicted enzymes involved in carbohydrate metabolism

A total of 145 carbohydrate-active enzyme (CAZyme)-encoding genes were identified and categorized into five CAZyme families (Fig. 5). Glycosyltransferases (GTs) constituted the largest group, comprising 67 genes distributed across 22 enzyme families, with GT71 (12 genes) being the most represented, followed by GT9 (9 genes) and GT2 (6 genes). Glycoside hydrolases (GHs) represented by 61 genes spanning 22 families, were dominated by GH76 (7 gene copies), followed by GH16 (6 copies) as well as GH5 and GH17 (5 copies each). The genomic distribution of CAZyme genes across contigs was not specifically resolved in this study, and therefore tandem duplication versus dispersed genomic localization could not be determined. Auxiliary activities (AAs) were represented by 12 genes across 6 families, with AA1 (3 genes) the most abundant. CEs and CBMs were the smallest groups, represented by 2 families each, totalling 2 and 3 genes, respectively. Several families were represented by single genes, indicating considerable functional diversity within the CAZyme repertoire. Comparison with published genomes of biocontrol and environmental yeasts shows that *H. burtonii* Y951 shares several conserved fungal cell wall-degrading glycoside hydrolase families with related species, including GH16, GH17, GH72, and GH18, though copy number varies considerably among species (Supplementary Tables 4 and 5). Notably, *H. burtonii* Y951 lacks GH55 entirely, a family also absent in *M. pul-*

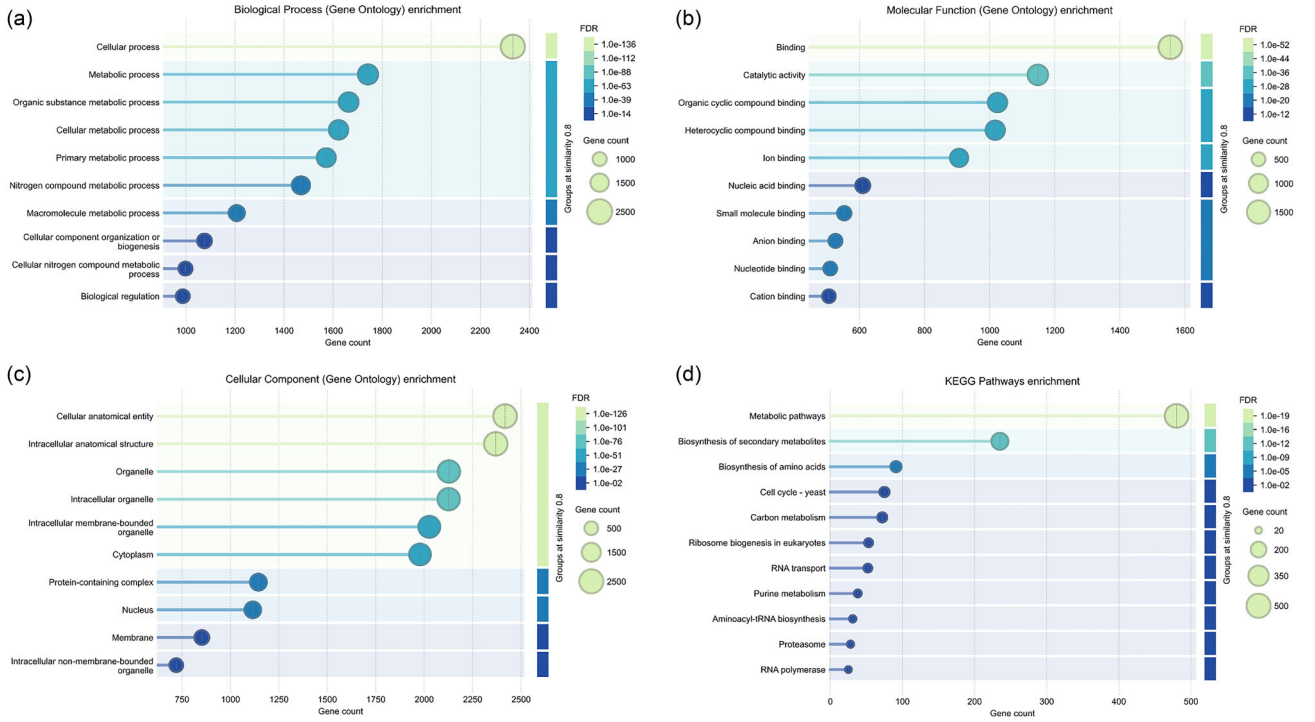


Figure 4 GO classification of 2831 annotated genes showing enriched biological processes (a), molecular function (b), and cellular component (c), showing the enriched categories on the *y*-axis and the number of genes assigned to each category on the *x*-axis. The bubble size represents the number of genes. KEGG enrichment analysis (d) shows enriched pathways on the *y*-axis.

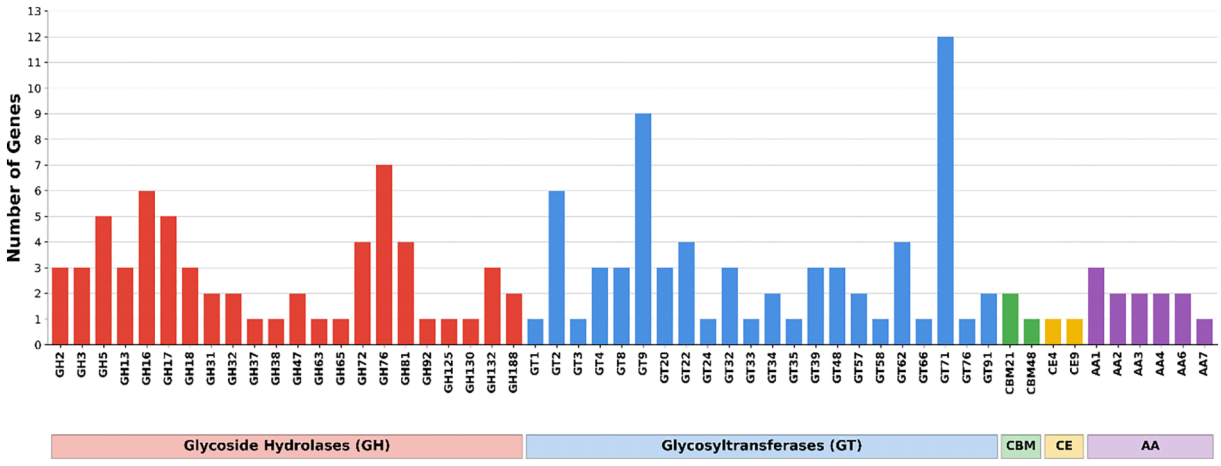


Figure 5 Distribution of CAZyme gene families by gene count. Bar chart depicting the abundance and diversity of 106 carbohydrate-active enzyme genes categorized into four major families.

cherrima and *W. anomalus* but present in *A. pullulans* (4 copies) and *M. fructicola* (6 copies), while GH18 chitinase copy number ranges from 3 in *H. burtonii* Y951 to as many as 28 in *M. fructicola* (Supplementary Table 5). Among the predicted CAZymes, using blastp in UniProt showed that three of the GH5 sequences had similarity with glucan-1,3-β-glucosidases (EC:3.2.1.58, Exo-1,3-β-glucanase) from *C. oleophila* and *W. anomalus*. Importantly, contig_11–1234 displayed 61% similarity to the EXG1 of *C. oleophila* as well as 42% and 58% similarity to EXG1 and EXG2 of *W. anomalus*, respectively.

Predicted secondary metabolite clusters

In this study, comprehensive genome mining and annotation using antiSMASH revealed the presence of multiple secondary metabolite BGCs in the fungal genome (Fig. 6). These BGCs represent three main types of pathways: terpene precursor, terpene, and non-ribosomal peptide synthetase (NRPS)-like clusters, each commonly associated with the production of bioactive compounds. This classification highlights the diversity of secondary metabolite biosynthesis potential within

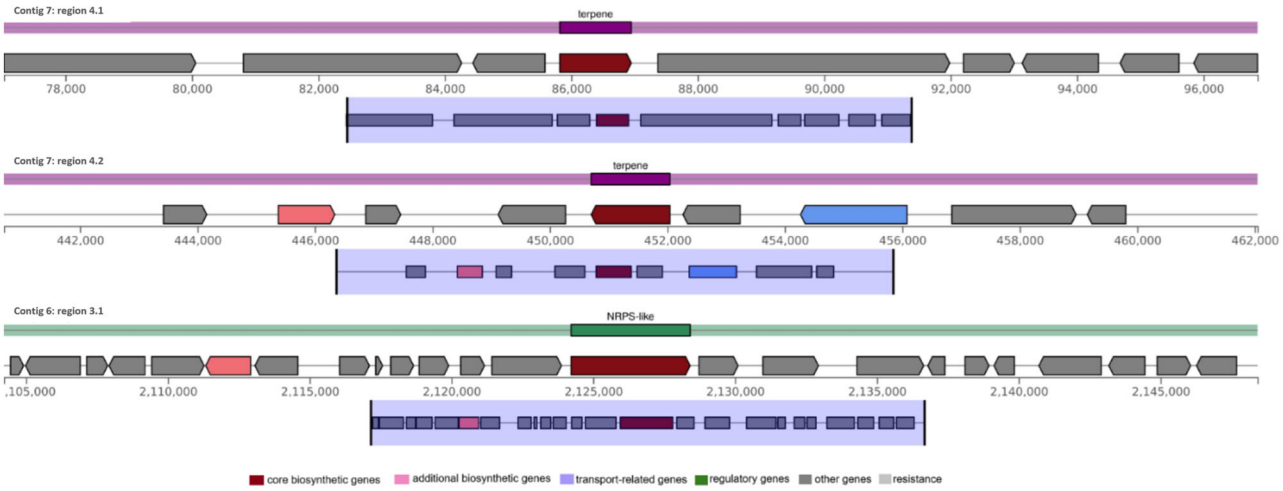


Figure 6 Predicted secondary metabolite BGCs identified by antiSMASH. Key clusters include: core biosynthetic, additional biosynthetic, transport-related, regulatory, other genes and resistance genes. Genomic coordinates are indicated above each cluster. Terpene and NRPS-like clusters are annotated.

the genome. The metabolic pathways and BGCs described here were predicted from genome annotation and comparative genomic analyses. The gene expression, biochemical activities, and metabolic products were not evaluated in the current study.

Specifically, antiSMASH analysis identified two terpene BGCs located on Contig 7 (regions 4.1 and 4.2) (Fig. 6). These clusters represent a putative genomic basis for the antifungal potential of *H. burtonii* Y951. The terpene clusters displayed a modular organization consisting of core biosynthetic genes flanked by accessory genes, including additional biosynthetic and transport-related genes. AntiSMASH predicted the potential biosynthesis of sesquiterpenes, diterpenes, sesterterpenes, and triterpenes. Complementary KEGG pathway analysis confirmed the presence of genes encoding the complete terpenoid backbone biosynthesis pathway, including the mevalonate pathway and the synthesis of geranyl diphosphate, farnesyl diphosphate, geranylgeranyl diphosphate, and hexaprenyl diphosphate (Fig. 7).

In addition to the terpene BGCs, antiSMASH identified a prominent NRPS-like BGC on Contig 6 (region 3.1) (Fig. 6). This cluster contained a core biosynthetic gene flanked by multiple accessory genes, suggesting the potential to produce non-ribosomal peptide-derived secondary metabolites that may contribute to the ecological fitness and antifungal activity of *H. burtonii* Y951. NRPS clusters represent secondary metabolite BGCs that may contribute to the production of bioactive compounds with potential ecological or antagonistic functions. The NRPS-like cluster contained genes encoding proteins with a ThiF domain, DLH, AMP-binding domain, PP-binding domain, PDEase_I, and Ras family domains. ThiF encodes a sulphur carrier protein ThiS adenyltransferase, while DLH is a dienelactone hydrolase family. Ras family proteins and AMP-binding proteins are regulatory proteins, involved in cell proliferation and stress response for the Ras proteins, and metabolism and signalling for the AMP-binding proteins. The PP-binding domain provides a phosphopantetheine attachment site while the PDEase_I codes for a 3'5-cyclic nucleotide phosphodiesterase.

Putative fermentome and flavourome relevant genes

In the current study, we evaluated the genome of *H. burtonii* Y951 for genes involved in various metabolic process. Similar to the BGCs, the expression of the genes as well as their biochemical activities and metabolic products were not evaluated. Among them, genes involved in glucose and galactose catabolism, the synthesis of vitamins, branched chain amino acids, as well as ergosterol were confirmed to be present in the genome (Supplementary Table S6). We further searched for the fermentome and flavourome associated genes as recently reported by Franco-Duarte et al. (2025). Only 15 of the genes defined as the putative fermentome of non-*Saccharomyces* yeasts were detected in the genome of *H. burtonii* Y951 (Table 5).

An evaluation of the genes associated with the non-*Saccharomyces* yeasts flavourome revealed that *H. burtonii* harbours a broad complement of genes encoding enzymes involved in the production of higher alcohols, acetate esters, ethyl esters, volatile organic acids, benzyl alcohols, and volatile thiols (Table 6). Most importantly, genes involved in the metabolism of aromatic amino acids (e.g. phenylalanine, tryptophan, and tyrosine) were more dominant in the genome. Among the transporters, genes for the uptake of ammonia (*MEP2*), all L-amino acids (*GAP1*) and methionine (*MUP1* and *MUP2*) were present in the genome (Table 6), while those for uptake of branched chain amino acids (*BAP2* and *TAT2*) were not detected. In addition, *PDC1*, *PDC5*, and *PDC6* encoding pyruvate decarboxylase as well as genes encoding aryl alcohol dehydrogenases were also missing. The genome contained all the genes for the oxidative decarboxylation of pyruvate to acetyl-coA and for the super-pathway of branched chain amino acid biosynthesis (Supplementary Table S2).

Discussion

Hyphopichia burtonii is a yeast species frequently isolated from natural environments where it demonstrates remarkable compet-

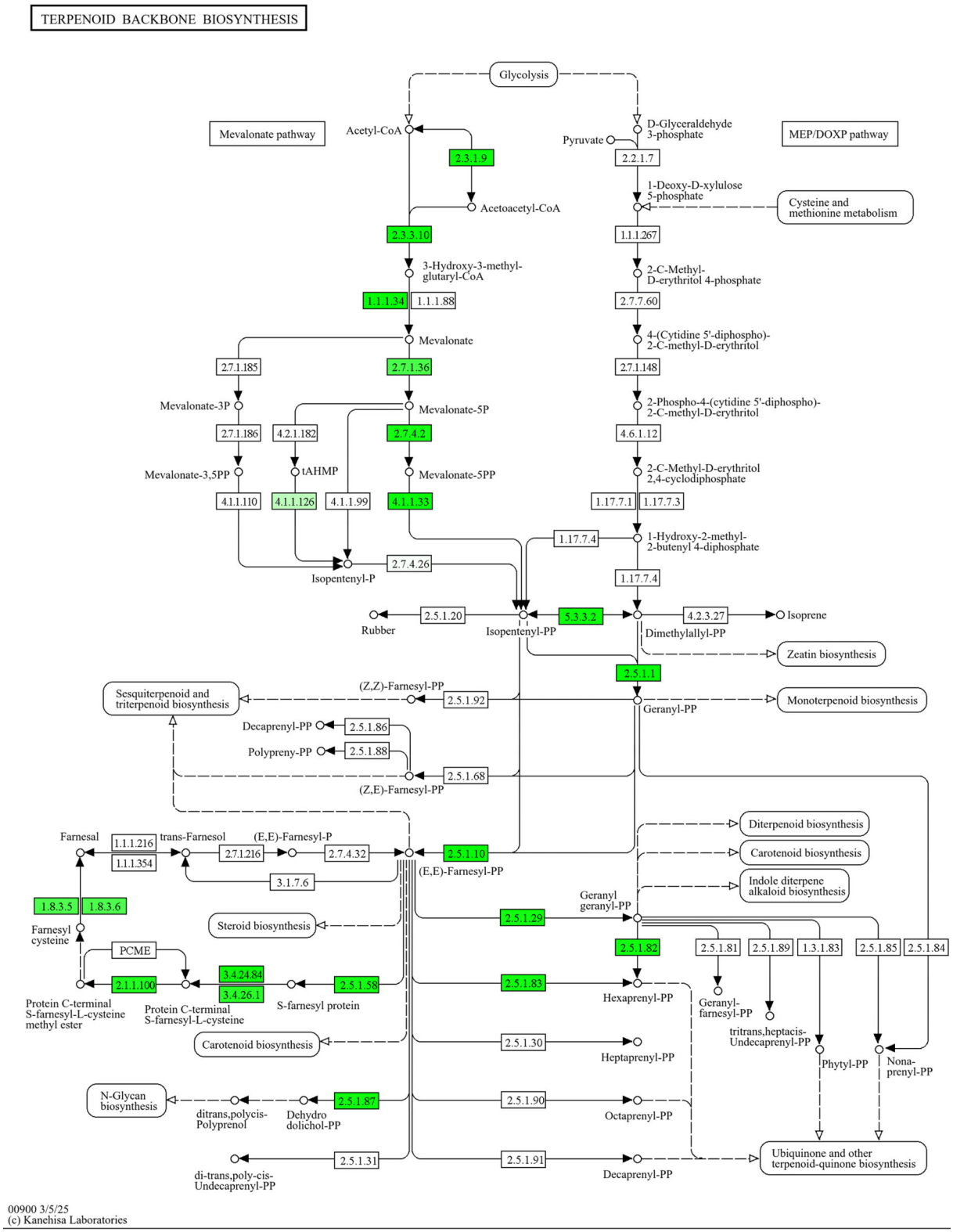


Figure 7 KEGG mapping of the terpenoid backbone biosynthesis pathway identified in *H. burtonii* Y951, the enzymes encoded by the genes present in the genome are highlighted .

Table 5 Predicted genes associated with the putative non-*Saccharomyces* yeasts fermentome.

Gene name	Short description
CIC1	Core interacting component
ATS1	Alpha tubulin suppressor
EDC3	Enhancer of mRNA decapping
BUL1	Binds ubiquitin ligase
LYS14	LYSine requiring
RPA34	RNA polymerase A
RMD9	Required for meiotic nuclear division
LBD19	Low dye binding
GEP3	Genetic interactors of prohibitions
SWC5	SWr complex
RAP1	Repressor/activator site binding protein
DPB11	DNA polymerase B (II)
STS1	Sec twenty-three suppressor 1
MDS3	Mck1 dosage suppressor
ECM21	Extracellular mutant

Table 6 Predicted genes associated with the putative flavourome non-*Saccharomyces* yeasts.

Main metabolic categories	Gene name
Higher alcohol—permeases of amino acids	*MEP2, MUP1, MUP3*
Higher alcohol—aromatic amino acid transferase	*ARO8, ARO9*
Higher alcohol—branched chain amino acid transferase	*BAT2, PAD1*
Higher alcohol—alcohol dehydrogenase	*ADH1, ADH2, GRE2, SPE1*
Higher alcohol—regulation	*ARO80, GLN3*
Acetate esters—alcohol acetyl transferases	*ATF1*
Ethyl esters—ethanol O-acyltransferase and esterase	*EHT1, IAH1*
Volatile organic acids—acetic acid	*ACS1, ACS2*
Volatile organic acids—aldehyde	*ALD2, ALD5, ALD6, ILV1, ILV2, ILV3, ILV5, ILV6*
p-ABA synthesis of p-ABA from chorismate, phenylalanine, tryptophan, and tyrosine	*ABZ1, ABZ2*
Aromatic amino acid synthesis—synthesis of chorismate, phenylalanine, tryptophan, and tyrosine	*ARO1, ARO2, ARO3, ARO4, ARO7, TRP1, TRP2, TRP3, TRP5, TRP5, PHA2, TYR1*
Benzyl alcohol/benzaldehyde synthesis—mandelate pathway	*ARO10, SCS7, DLD1, DLD2*
Sulphur metabolism—Volatile thiols—conversion of conjugated precursors into varietal thiols	*GAP1*
Sulphur metabolism—Volatile thiols β-lyase cleaving the cysteinylated precursors releasing 4MMP	*CYS4, STR2, STR3*
Sulphur metabolism—Volatile thiols—degradation of glutathionylated precursors	*DUG3*
Sulphur metabolism—sulphate and sulfite reductases	*MET2, MET3, MET5, MET6, MET10, MET14. MET16*
Glycerol production—redox reactions with flavour implications	*GPD1*

itiveness under diverse and often stressful conditions (Simoncini et al. 2007, Ghosh et al. 2015, Seco et al. 2024). Consequently, strains of this species have been the subject of genomic studies to decipher the genetic determinants of traits such as diverse substrate utilization (Riley et al. 2016), unique osmotolerance mechanisms (Lee et al. 2021), fermentation traits (Franco-Duarte et al. 2025), and drug resistance mechanisms (Indoung et al. 2025). In the current study, the genome of *H. burtonii* Y951 isolated from *Vitis vinifera* L. cv. Cabernet Sauvignon grape must, was sequenced and assembled using PacBio HiFi long-read sequencing, chosen for its superior sequence accuracy relative to other long-read platforms prone to higher error rates (Hon et al. 2020). The nearly complete genome assembly of Y951, with extensive contigs, high N50, and >97% BUSCO completeness, demonstrates its integrity and suitability for functional analyses. Such a complete and continuous assembly provides a reliable foundation for downstream functional and comparative analyses, enabling future work to elucidate the biological attributes underpinning this strain's metabolic capabilities and adaptability. Consistent with this, previously reported genome sequences such as strain Tip11 (isolated from zebra dove droppings) were included for contextual comparison and exhibit lower completeness (BUSCO 75.5%), further highlighting differences in available genome quality among reported *H. burtonii* assemblies (Indoung et al. 2025) as well as strain KJJ43 (isolated from Nuruk samples) with 92.8% BUSCO completeness (Lee et al. 2021), the assembly of strain Y951 enables more accurate gene prediction and annotation. *Hyphopichia burtonii* Y951, KJJ43 (Lee et al. 2021) and NRRL Y-1933 (Riley et al. 2016; Franco-Duarte et al. 2025) share similar genome sizes and annotation metrics, indicating conserved attributes across environmental strains that support shared physiology. The genome of *H. burtonii* carries 49 rRNA, 120 tRNA, 50 snoRNA, and 22 snRNA genes giving it a full machinery involved in regulation of ribosomal function, RNA processing and modification, pre-mRNA splicing as well as translation (Qi et al. 2025). Among the rRNAs, 5S rRNA was the most abundant with 39 copies. This is closer to the genome of *H. burtonii* NRRL Y-1933 which was reported to carry 41 copies of dispersed 5S rRNA while the 18S-5.8S-26S was found to be arranged as tandem RNA arrays (Riley et al. 2016). Overall, the genome exhibits similar abundance and diversity patterns of predicted ncRNAs as other yeasts of the Serinales order such as *Candida oleophila* (Sui et al. 2020), *Clavispora santaluciae* (Franco-Duarte et al. 2022), and *Metschnikowia bicuspidata* (Jiang et al. 2022).

The comparative analysis between Y951 and NRRL Y-1933 revealed differences in assembly contiguity, reflecting the use of different sequencing technologies and assembly strategies. While both genomes show comparable completeness, these dif-

ferences limit direct comparison of genome organization beyond assembly-level metrics. Consistent with this limitation, the Mauve alignment showed conserved syntenic blocks between the two assemblies. However, due to differences in assembly contiguity, the alignment should be interpreted as a qualitative representation of shared genomic structure rather than evidence of structural variation (Riley et al. 2016).

Functional annotation assigned GO terms to nearly 78% of predicted genes, revealing broad metabolic and cellular capabilities. Gene distribution across COG confirmed the prominence of fundamental biological processes such as transcription, carbohydrate metabolism, and amino acid biosynthesis which are critical for cellular growth and environmental adaptation. Additional signatures related to signal transduction pathways and iron transport emphasize mechanisms for environmental sensing and nutrient acquisition, key features that likely contribute to the yeast's ecological fitness. Interestingly, a significant fraction of genes remains uncharacterized ('function unknown') compared to other yeasts such as *M. bicuspidata* (Jiang et al. 2022), *C. oleophila* (Sui et al. 2020), *Cyberlindnera rhodanensis* (Shen et al. 2025), and *C. santaluciae* (Franco-Duarte et al. 2022), reflecting the still-incomplete understanding of yeast genomes and hinting at the presence of novel functionalities in *H. burtonii* Y951, as observed in other *H. burtonii* strains (Riley et al. 2016, Lee et al. 2021).

Analysis of CAZymes revealed a diversity of genes encoding enzymes spanning 44 families with GTs and GHs being the dominant classes as observed in other yeasts, including *Meyerozyma* sp. (Bin Rosali et al. 2025) and *Aureobasidium pullulans* (Xiao et al. 2025). A high number of genes associated with GT15, GT17, and GT65 alludes to an ability to yeast to undertake various glycosylation activities. Moreover, the diversity of GHs to broad metabolic versatility that likely facilitates efficient utilization of diverse carbohydrate substrates, including cellulose (GH5), starch (GH13), xylan (GH43), and fructan (GH32). This is not surprising as *H. burtonii* is already known for this metabolic versatility, where similar to other Serinales yeasts, the genes for utilization of substrates such as cellulose have been reported to occur in functional clusters (Riley et al. 2016). Conversely, enzymes in family GH5, GH16, GH17, and GH55 display glucan endo- and exo-β-glucosidase activities with abilities to hydrolyse β-1,3 and β-1,6 linked glucan present in fungal cell walls. In addition, GH18 represents chitinases which are involved in the degradation of chitin. Glucanases and chitinases are involved in fungal cell wall remodelling thereby supporting cell growth and hyphal extension. Beyond these core functions, both enzymes have been associated with antifungal activity and mycoparasitism in yeasts reported as biological control agents against several phytopathogens. Similarly, phenotypic characterization of *H. burtonii* Y951 revealed glucanase and chitinase production to be part of the antagonistic activity of this yeast against different strains of *Botrytis cinerea* as well as other fungi such as *Alternaria alternata* and *Aspergillus niger* (Maluleke et al. 2022). Sequence comparison using blastp revealed that among all the predicted CAZymes, considerable homology with EC:3.2.1.58, exo-1,3-β-glucanase of the *W. anomalus* and *C. oleophila* was evident within the GH5 assigned sequences of *H. burtonii* Y951. These enzymes have been shown particularly in *W. anomalus* to be directly linked with the antifungal activity of these yeasts (Friel et al. 2007, Parafati et al. 2017).

More than 200 genes in the *H. burtonii* Y951 genome were linked to the biosynthesis of secondary metabolites and AntiSMASH analysis identified two primary BGC types: the terpene and NRPS-like clusters. KEGG pathway mapping revealed genes involved in the mevalonate pathway and terpenoid backbone biosynthesis. In particular, the enzymes highlighted in these pathways, e.g. geranyl-diphosphate synthase, farnesyl-diphosphate synthase, geranylgeranyl-diphosphate synthase, and hexaprenyl-diphosphate synthase, suggest an ability to synthesize diterpenoids as well as ubiquinone and other terpenoid quinones, which comprise chemically diverse natural products involved in antimicrobial activity, membrane interactions, and intracellular signalling, often underpinning yeast attachment and defence strategies (González-Hernádez et al. 2023).

NRPS clusters represent a class of large, modular enzymes that synthesize the peptide backbones of bioactive natural products, including antimicrobials and microbial signalling molecules (Bushley and Turgeon 2010). NRPS clusters typically contain three core domains that are required for their functionality: an adenylation domain, a thiolation domain (or peptide carrier protein) and a condensation domain. In NRPs-like clusters, the condensation domain is typically missing and is replaced by a thioesterase or a reductase or the cluster might consist of single A domains (Palys et al. 2020). The NRPS-like cluster in *H. burtonii* Y951 comprises a ThiF which catalyses the ATP-dependent adenylation of the ThiS protein in thiamine biosynthesis, and a PP-binding domain which serves as the peptidyl carrier protein providing an anchoring point for 4-phosphopantetheine. Together, these domains account for adenylation and thiolation. The cluster also contained a dienelactone hydrolase (DLH) family protein domain. DLH domains are occasionally found in fungal NRPS-like clusters. In *A. niger*, DLH family proteins were reported to play a role in the biosynthesis of alkylcitric acids (Palys et al. 2020), while in *Metarhizium robertsii* , DLH was localized in an NRPS cluster driving the biosynthesis of the hybrid polyketide-peptides (NG-39x), non-ribosomal peptides such as destruxins, serinocilins, and siderophores such as ferricrocin and metachelins (Donzelli et al. 2017). This diversity of metabolites suggests that the presence of the NRPS-ilke cluster comprising the DLH domain in *H. burtonii* Y951 endows this yeast with the ability to produce diverse biologically active compounds, including those involved in antifungal activity, detoxification, and stress tolerance.

Altogether, the high-quality, highly contiguous genome assembly of *H. burtonii* Y951 provides a comprehensive resource elucidating the molecular bases for its ecological competitiveness and biocontrol potential. The detailed characterization of predicted carbohydrate-active enzymes and secondary metabolite biosynthetic clusters reveals a potential complex arsenal likely contributing to its antagonistic and adaptive traits. The predicted genes and metabolic pathways define the inferred functional potential of *H. burtonii* Y951 and generate hypotheses about its possible role within microbial communities, which may be relevant to sustainable crop protection if confirmed by targeted phenotypic and ecological studies.

Analysis of the *H. burtonii* Y951 genome shows that this yeast encodes for the potential production of various aroma compounds including higher alcohols, esters, volatile organic acids, benzyl-alcohols, and volatile thiols. The genome is evidently endowed with most of the genes recently proposed as the putative fermentome and flavourome of non-*Saccharomyces* yeasts (Franco-Duarte et al. 2025). However, compared to other strains of *H. burtonii* as well as other *Hyphopichia* species, *H. burtonii* Y951 seems

to lack several genes, including *PDC1*, *PDC5*, and *PDC6* encoding pyruvate decarboxylase, a key enzyme in directing pyruvate to alcoholic fermentation and in the decarboxylation of branched-chain 2-oxoacids-intermediates in the synthesis of higher alcohols, as well as *THI3* which contributes to biosynthesis of isoamyl alcohol. This could suggest that the role of pyruvate decarboxylase in *H. burtonii* Y951 is taken over by a decarboxylase encoded by the *ARO10* gene similar to what has been proposed for other yeasts such as species of the *Starmerella* clade that lack a yeast *PDC1* orthologue (Gonçalves et al. 2020). This lack of the classical pyruvate decarboxylase encoding genes could partly explain the low fermentation power observed in this strain in high sugar environments (Rollero et al. 2018). In addition, orthologues of genes encoding aryl alcohol dehydrogenases were not found in the genome *H. burtonii* Y951 suggesting limited capacity to reduce aryl-aldehydes to their corresponding aryl-alcohols, while other strains of *H. burtonii* as well as other *Hyphopichia* spp. were found to possess *AAD3*, *AAD4*, *AAD10*, *AAD14*, and *AAD16* (Franco-Duarte et al. 2025). Apart from *AAD4* and *AAD14*, little is known about the ecological and physiological role of these genes in ascomycetous yeasts. Overall, the data reveal that there is some intraspecific variation in the putative fermentome and flavourome of *H. burtonii* that could be due to differences in strain origins as has been observed in other species (Silva et al. 2023), Vincente et al. 2025). Mauve alignment of the genomes of strain Y951 and NRRL Y-1933 indicated that several rearrangement events have occurred on the chromosomes of *H. burtonii*, which would contribute to this variation between strains. However, this should be further tested with a bigger pool of strains of distinct originating from different geographical areas, source materials and biotechnological processes.

Conclusions

The genome assembly of *H. burtonii Y951* demonstrates high completeness, with a BUSCO score of 99.3%, and notable contiguity as indicated by assembly metrics. Annotation of the genome highlights strong conservation of core genes along with a diverse metabolic repertoire. This includes a rich set of carbohydrate-active enzymes and multiple BGCs, specifically those encoding terpenes and non-ribosomal peptides, which suggest potential functional roles in secondary metabolism and microbial interactions. Comparative genomic analysis with the wild-type NRRL Y-1933 strain reveals overall synteny, while distinct structural variations are evident, reflecting evolutionary divergence between strains. Together, these genomic insights provide a foundation for understanding the biology, ecology of *H. burtonii* Y951 and potential biotechnological applications, and for in-depth exploration into the genomic and phenotypic diversity within the species. Although the genome analysis identified genes and pathways potentially associated with antifungal activity, fermentation, aroma production, metabolic capacity, and ecological adaptation, their specific contributions to these traits remain to be experimentally established.

Acknowledgements

We thank the National Research Foundation (NRF) of South Africa for the Postdoctoral Fellowship (PSTD23042496653) awarded to Dr. Evelyn Maluleke.

Author contributions

Evelyn Maluleke analysed and interpreted all the genome sequencing data and wrote the original draft manuscript. Mathabatha Evodia Setati and Evelyn Maluleke conceived the project. Mathabatha Evodia Setati acquired funding for the research. All authors edited and approved the final version of the manuscript.

Supplementary material

Supplementary material is available at *FEMSYR Journal* online.

Conflicts of interest

None declared.

Funding

This project was supported wholly by the National Research Foundation of South Africa Grant No. 150549.

Data availability

The raw sequencing reads for *H. burtonii* strain Y951 are available in the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA1391825, BioSample accession SAMN54234055, and SRA run accession SRR36539643. The genome assembly has been submitted to GenBank under submission ID SUB16549568.

References

Andrews S. *FastQC: a quality control tool for high throughput sequence data*. 2010. https://github.com/s-andrews/FastQC (02 December 2024, date last accessed).

Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–29. https://doi.org/10.1038/75556.

Bin Rosali MI, Musa SM, Kiram SK *et al.* Whole-genome sequencing of a novel Meyerozyma sp. nov, fungi isolated from Harumanis mango, Mangifera indica L. in Malaysia. *Microbe* 2025;**7**:100354. https://doi.org/10.1016/j.microb.2025.100354

Bushley KE, Turgeon BG. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC Evol Biol* 2010;**10**:26. https://doi.org/10.1186/1471-2148-10-26.

Chamroensakchai T, Kanjanabuch T, Saikong W *et al.* The first human report of *Hyphopichia burtonii*, initially misdiagnosed as sterile peritonitis in a patient on peritoneal dialysis. *Med Mycol Case Rep* 2021;**33**:26–29. https://doi.org/10.1016/j.mmcr.2021.06.005.

Chan PP, Lin BY, Mak AJ *et al.* tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* 2021;**49**:9077–96. https://doi.org/10.1093/nar/gkab688.

Donzelli BGD, Turgeon BG, Gibson DM *et al.* Disruptions of the genes involved in lysine biosynthesis, iron acquisition, and secondary metabolisms affect virulence and fitness in *Metarhizium*

robertsii. *Fungal Genet Biol* 2017;**98**:23–34. https://doi.org/10.1016/j.fgb.2016.11.005.

Franco-Duarte R, Čadež N, Rito T *et al.* Whole-genome sequencing and annotation of the yeast *Clavispora santaluciae* reveals important insights about its adaptation to the vineyard environment. *J Fungi* 2022;**8**:52. https://doi.org/10.3390/jof8010052.

Franco-Duarte R, Fernandes T, Sousa MJ *et al.* Phylogenomics and functional annotation of 530 non-Saccharomyces yeasts from winemaking environments reveals their fermentome and flavorome. *Stud Mycol* 2025;**111**: 1–17. https://doi.org/10.3114/sim.2025.111.01

Friel D, Gomez Pessoa NM, Vandenbol M *et al.*. Separate and combined disruptions of two exo-β-1,3-glucanase genes decrease the efficiency of *Pichia anomala* (Strain K) biocontrol against *Botrytis cinerea* on apple. *Mol Plant-Microbe Interact* 2007;**20**:371–9. https://doi.org/10.1094/MPMI-20-4-0371.

Gadanho M, Sampaio JP. Occurrence and diversity of yeasts in the mid-Atlantic ridge hydrothermal fields near the Azores Archipelago. *Microb Ecol* 2005;**50**:408–17. https://doi.org/10.1007/s00248-005-0195-y.

Galperin MY, Vera Alvarez R, Karamycheva S *et al.* COG database update 2024. *Nucleic Acids Res* 2025;**53**:D356–63. https://doi.org/10.1093/nar/gkae983.

Garcia MV, Bernardi AO, Parussolo G *et al.* Spoilage fungi in a bread factory in Brazil: diversity and incidence through the bread-making process. *Food Res Int* 2019;**126**:108593. https://doi.org/10.1016/j.foodres.2019.108593.

Ghosh S, Bagheri B, Morgan HH *et al.* Assessment of wine microbial diversity using ARISA and cultivation-based methods. *Ann Microbiol* 2015;**65**:1833–40. https://doi.org/10.1007/s13213-014-1021-x.

Gonçalves P, Gonçalves C, Brito PH *et al.* The *Wickerhamiella*/*Starmerella* clade-A treasure trove for the study of the evolution of yeast metabolism. *Yeast* 2020;**37**:313–20. https://doi.org/10.1002/yea.3463.

González-Hernández RA, Valdez-Cruz NA, Macías-Rubalcava ML *et al.* Overview of fungal terpene synthases and their regulation. *World J Microbiol Biotechnol* 2023;**39**:194. https://doi.org/10.1007/s11274-023-03635-y.

Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinf* 2013;**10**:645–56. https://doi.org/10.1109/TCBB.2013.68.

Groenewald M, Smith MT. Re-examination of strains formerly assigned to *Hyphopichia burtonii*, the phylogeny of the genus *Hyphopichia*, and the description of *Hyphopichia pseudoburtonii* sp. nov. *Int J Syst Evol Microbiol* 2010;**60**:2675–80. https://doi.org/10.1099/ijs.0.018580-0.

Hon T, Mars K, Young G *et al.* Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* 2020;**7**:399. https://doi.org/10.1038/s41597-020-00743-4.

Huerta-Cepas J, Szklarczyk D, Forslund K *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 2016;**44**:D286–93. https://doi.org/10.1093/nar/gkv1248.

Indoung S, Chumtong S, Prachantasena S *et al.* Whole-genome sequencing of *Hyphopichia burtonii* from isolated yeast recovered from zebra dove droppings in Thailand. *PLoS One* 2025;**20**:e0315067. https://doi.org/10.1371/journal.pone.0315067.

Indoung S, Prachantasena S, Wiriyaprom R *et al.* Detection of zoonotic *Cryptococcus neoformans* in zebra dove (*Geopelia striata*) dropping by nested PCR. *Southeast Asian J Trop Med Public Health* 2021;**54**:780–9.

Jeong DM, Kim HJ, Jeon MS *et al.* Genomic and functional features of yeast species in Korean traditional fermented alcoholic beverage and soybean products. *FEMS Yeast Res* 2023;**23**:foac066. https://doi.org/10.1093/femsyr/foac066.

Jiang H, Bao J, Xing Y *et al.* Comparative genomic analyses provide insight into the pathogenicity of *Metschnikowia bicuspidata* LNES0119. *Front Microbiol* 2022;**13**:939141. https://doi.org/10.3389/fmicb.2022.939141.

Kim JY, Yeo SH, Baek SY *et al.* Molecular and morphological identification of fungal species isolated from bealmijang meju. *J Microbiol Biotechnol* 2011;**21**:1270–9. https://doi.org/10.4014/jmb.1105.05013.

Kurtzman CP, Fell JW, Boekhout T. *The Yeasts, a Taxonomic Study*, 5th edn. Amsterdam: Elsevier B.V., 2011, 2080.

Lee DW, Hong CP, Thak EJ *et al.* Integrated genomic and transcriptomic analysis reveals unique mechanisms for high osmotolerance and halotolerance in *Hyphopichia* yeast. *Environ Microbiol* 2021;**23**:3499–522. https://doi.org/10.1111/1462-2920.15464.

Lee JI, Kim DW, Jang GJ *et al.* Effects of different storage conditions on the metabolite and microbial profiles of white rice (*Oryza sativa* L.). *Food Sci Biotechnol* 2019;**28**:623–31. https://doi.org/10.1007/s10068-018-0520-0.

Lin Y, Yuan J, Kolmogorov M *et al.* Assembly of long error-prone reads using de Bruijn graphs. *Proc Natl Acad Sci* 2016;**113**:E8396–405. https://doi.org/10.1073/pnas.1604560113.

Lowe TM, Eddy SR tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997;**25**:955–64. https://doi.org/10.1093/nar/25.5.955.

Maluleke E, Jolly NP, Patterton HG *et al.* Antifungal activity of non-conventional yeasts against *Botrytis cinerea* and non-*Botrytis* grape bunch rot fungi. *Front Microbiol* 2022;**13**:986229. https://doi.org/10.3389/fmicb.2022.986229.

Manni M, Berkeley MR, Seppey M *et al.* BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* 2021;**38**:4647–54. https://doi.org/10.1093/molbev/msab199.

Medema MH, Blin K, Cimermancic P *et al.* AntiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 2011;**39**:W339–46. https://doi.org/10.1093/nar/gkr466.

Mikheenko A, Prjibelski A, Saveliev V *et al.* Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 2018;**34**:i142–50. https://doi.org/10.1093/bioinformatics/bty266.

Mistry J, Chuguransky S, Williams L *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**:D412–9. https://doi.org/10.1093/nar/gkaa913.

Nawrocki EP, Eddy SR Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;**29**:2933–5. https://doi.org/10.1093/bioinformatics/btt509.

Ohisa N, Suzuki N, Miura M *et al.* Halotolerant yeast participates in the cavity formation of hand-pulled dry noodles. *J Japan Soc Food Sci Technol* 2012;**59**:442–6. https://doi.org/10.3136/nskkk.59.442.

Palmer JM, Stajich J. Funannotate v1.8.1: eukaryotic genome annotation (Version v1.8.1) [Computer software]. *Zenodo* 2020. https://doi.org/10.5281/zenodo.4054262

Palys S, Pham TTM, Tsang A. Biosynthesis of alkylcitric acids in *Aspergillus niger* involves both co-localized and unlinked genes. *Front Microbiol* 2020;**11**:1378. https://doi.org/10.3389/fmicb.2020.01378.

Parafati L, Cirvilleri G, Restuccia C *et al.* Potential role of exoglucanase genes (*WaEXG1* and *WaEXG2*) in the biocontrol activity of *Wickerhamomyces anomalus*. *Microb Ecol* 2017;**73**:876–84. https://doi.org/10.1007/s00248-016-0887-5.

Pinheiro REE, Rodrigues AMD, Santos J *et al.* Occurrence and diversity of yeast species isolated from fish feed and tambatinga gut. *Lat Am J Aquat Res* 2018;**46**:837–42. https://doi.org/10.3856/vol46-issue4-fulltext-22.

Qi T, Delneri D, Timouma S. Functional analysis and transcriptional profiling of non-coding RNAs in yeast. *Biochem J* 2025;**482**:1465–87. https://doi.org/10.1042/BCJ20253069.

Riley R, Haridas S, Wolfe KH *et al.* Comparative genomics of biotechnologically important yeasts. *Proc Natl Acad Sci* 2016;**113**:9882–7. https://doi.org/10.1073/pnas.1603941113.

Rollero S, Bloem A, Ortiz-Julien A *et al.* Fermentation performances and aroma production of non-conventional wine yeasts are influenced by nitrogen preferences. *FEMS Yeast Res* 2018;**18**:foy055. https://doi.org/10.1093/femsyr/foy055.

Seco MNP, Beltran AKM, Balendres MAO. First record of *Hyphopichia burtonii* isolated from the storage pest *Sitophilus zeamais* and its bioactivity against mycotoxigenic fungi. *Hell Plant Prot J* 2024;**17**:97–110. https://doi.org/10.2478/hppj-2024-0009.

Seemann T. *barrnap 0.8: rapid ribosomal RNA prediction*. 2013.

Shen Y, Tu Z, Tang L *et al.* Genome sequencing, assembly and characterization of *Cyberlindnera rhodanensis* J52 as a non-*Saccharomyces* yeast with ester-enhancing potential. *J Fungi* 2025;**11**:135. https://doi.org/10.3390/jof11020135.

Silva M, Pontes A, Franco-Duarte R *et al.* A glimpse at an early stage of microbe domestication revealed in the variable genome of *Torulaspora delbrueckii,* an emergent industrial yeast. *Mol Ecol* 2023;**32**:2396–412. https://doi.org/10.1111/mec.16428.

Simão FA, Waterhouse RM, Ioannidis P *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**:3210–2. https://doi.org/10.1093/bioinformatics/btv351.

Simoncini N, Rotelli D, Virgili R *et al.* Dynamics and characterization of yeasts during ripening of typical Italian dry-cured ham. *Food Microbiol* 2007;**24**:577–84. https://doi.org/10.1016/j.fm.2007.01.003.

Smit A, Hubley R, Green P. *RepeatMasker Open-4.0* 2013. Avaliable from: http://www.repeatmasker.org/ (05 December 2024, date last accessed).

Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 2005;**33**:W465–7. https://doi.org/10.1093/nar/gki458.

Sui Y, Wisniewski M, Droby S *et al.* Genome sequence, assembly, and characterization of the antagonistic yeast *Candida oleophila* used as a biocontrol agent against post-harvest diseases. *Front Microbiol* 2020;**11**:295. https://doi.org/10.3389/fmicb.2020.00295.

The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* 2025;**53**:D609–D617.

Vincente J, Friedrich A, Schacherer J *et al.* Whole-genome sequencing and phenotyping reveal specific adaptations of *Lachancea thermotolerans* to the winemaking environment. *Mol Ecol* 2025;**0**:e17667. https://doi.org/10.1111/mec.17667.

Virgili R, Simoncini N, Toscani T *et al.* Biocontrol of *Penicillium nordicum* growth and ochratoxin A production by native yeasts of dry cured ham. *Toxins* 2012;**4**:68–82. https://doi.org/10.3390/toxins4020068.

Xiao D, Driller M, Stein K *et al.* Genome mining the black-yeast *Aureobasidium pullulans* NRRL 62031 for biotechnological traits. *Bmc Genom [Electron Res]* 2025;**26**:244. https://doi.org/10.1186/s12864-025-11395-2.

Zheng J, Ge Q, Yan Y *et al.* dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acid Res* 2023;**51**:W115–21. https://doi.org/10.1093/nar/gkad328.

Received: 8 April 2026. **Revised:** 14 October 2026. **Accepted:** 20 October 2026
© The Author(s) 2026. Published by Oxford University Press on behalf of FEMS. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site-for further information please contact journals.permissions@oup.com