

Genomic and Pangenomic Analysis of *Kluyveromyces Marxianus* SHY2, a Novel Strain with High Proteolytic Potential

Quan Qiu^{1,2,†}, Ying Li^{1,†}, Zhichun Zhan², Ying Zhou², Lingfang Gu², Qijun Wang², Yicheng Yang¹, Yunxiang Liang¹, Yingjun Li^{1,*} 

¹ National Key Laboratory of Agricultural Microbiology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, China.

² Wuhan Sunhy Biology Co., Ltd, Qiuquan, Wuhan 430070, China.

† These authors contributed equally to this work.

*Corresponding Author: Yingjun Li; E-mail: yingjun@mail.hzau.edu.cn

Abstract

Kluyveromyces marxianus is an industrially important yeast renowned for its thermotolerance, high growth rate, and broad substrate utilization. In this study, whole-genome sequencing and assembly were performed for a novel *K. marxianus* strain SHY2 isolated from traditional fermented dairy products, followed by a pangenomic comparative analysis with 15 publicly available strains. The genome of SHY2 contains a large number of protease and protease-regulatory-related genes (2,589 genes according to the MEROPS database, accounting for 48.70% of all protein-coding genes); however, this proportion includes not only active proteases but also peptidases, protease inhibitors and non-catalytic homologs, and therefore likely overestimates the genuine proteolytic capacity. The genome also harbors two secondary metabolite biosynthesis gene clusters. Pangenomic analysis revealed that *K. marxianus* possesses a near-closed pangenome, with total gene families reaching approximately 5,237 and a stable core of approximately 1,807 families when 16 strains were analyzed. SHY2 ranks high in private gene family count, suggesting a genetically distinct background. Functional enrichment of SHY2-specific private genes indicated over-representation of transmembrane transport and carbohydrate metabolic processes. All functional interpretations based on gene annotation require experimental validation (e.g., enzyme activity assays) before any claim of high proteolytic potential can be substantiated. In conclusion, this study provides a genomic and pangenomic resource for *K. marxianus* SHY2, forming a foundation for future hypothesis-driven research.

Keywords

Kluyveromyces marxianus; Genomics; Pangenomics; Strain-specific genes; Protease genes

Citation: Quan Qiu, Ying Li, Zhichun Zhan and *et al.* Genomic and Pangenomic Analysis of *Kluyveromyces Marxianus* SHY2, a Novel Strain with High Proteolytic Potential. *Biofeed Science and Technology*. 2026, 01 (1): 1-7. doi: 10.66729/bst202604003.

1. Introduction

Kluyveromyces marxianus has emerged as a versatile industrial workhorse due to its exceptional traits: rapid growth (generation time 60 – 90 min), thermotolerance up to 52 °C, and the ability to assimilate a wide range of carbon sources including lactose, xylose, and inulin^[1,2]. These properties make it attractive for producing enzymes, bioethanol, single-cell protein, and functional feed additives^[3]. Despite its biotechnological potential, the genetic diversity within *K. marxianus* remains underexplored, and most studies have focused on a few well-characterized strains such as DMKU3-1042^[4] and CBS 6556^[5].

The advent of high-throughput sequencing and pangenomic approaches allows comprehensive dissection of intra-species genetic variation, revealing core (conserved) and accessory (variable) gene repertoires that underlie phenotypic differences^[6]. Structural variants (SVs) and small nucleotide variants (SNVs) contribute to functional diversity by altering gene dosage, coding sequences, or regulatory regions^[7]. Integrating these data can pinpoint strain-specific traits relevant for industrial applications. However, genome-based predictions alone cannot confirm functional potential, and experimental validation remains essential.

Three major knowledge gaps exist: limited genomic resources for dairy-derived *K. marxianus* strains, unknown pangenome architecture of the species, and a lack of strain-specific genetic features linked to proteolytic potential. To address these gaps, this study tested the following hypotheses: SHY2, isolated from a protein-rich dairy environment, possesses a genetic repertoire enriched in protease-related genes compared to other strains; *K. marxianus* has a near-closed pangenome; and strain-specific private genes contribute to unique metabolic capabilities. Accordingly, whole-genome sequencing and pangenomic analysis were performed on a novel *K. marxianus* isolate, SHY2, obtained from traditional fermented yogurt. The results highlight SHY2 as a genetically distinct strain with unique metabolic features, providing a genomic basis for its future exploitation, but all functional claims are presented as testable hypotheses requiring experimental verification.

2. Materials and Methods

2.1 Strain and DNA Extraction

Kluyveromyces marxianus strain SHY2 was isolated from traditional fermented yogurt samples collected in Hubei Province, China. It is deposited at the China General Microbial Culture Collection



and Management Center under accession number CGMCC 10621. Genomic DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method^[8]. DNA quality and concentration were assessed by agarose gel electrophoresis, NanoDrop 2000, and Qubit 3.0 (Thermo Fisher Scientific, Wilmington, DE, USA).

2.2 Genome Sequencing and Assembly

Sequencing was performed at Novogene (Beijing, China) using a hybrid strategy: Illumina NovaSeq 6000 (Illumina, San Diego, California, USA) (paired-end 150 bp) and PacBio Sequel II (Illumina, San Diego, California, USA). Raw Illumina reads were processed with FastQC v0.11.9 and trimmed with Trimmomatic v0.39 (LEADING: 3, TRAILING: 3, SLIDINGWINDOW: 4:15, MINLEN: 36). PacBio reads were filtered with SMRT Link v8.0 (Pacific Biosciences, Menlo Park, California, USA). Hybrid assembly was performed with Canu v2.2^[9] followed by two rounds of polishing with Pilon v1.23 (Broad Institute, Cambridge, Massachusetts, USA)^[10]. Genome completeness was evaluated with BUSCO v5.2.2^[11] using the ascomycota_odb10 dataset.

2.3 Gene Prediction and Functional Annotation

Gene prediction combined AUGUSTUS v3.4.0 (University of Göttingen, Göttingen, Germany)^[12] with a yeast-specific training set and GeMoMa v1.7.1^[13] using *K. marxianus* DMKU3-1042 (GCA_001417885.1) as reference. EvidenceModeler^[4] integrated the predictions. tRNA and rRNA genes were identified with tRNAscan-SE v2.0 (Lowe Lab, University of California, Santa Cruz, California, USA)^[15] and RNAmmer v1.2 (Lagesen *et al.*, Norwegian University of Science and Technology, Trondheim, Norway)^[16].

Functional annotation was performed by BLASTP^[17] (E-value < 1e-5) againstNr (NCBIResource Coordinators, 2016), Swiss-Prot^[18], KEGG^[19], GO^[20], eggNOG^[21], Pfam^[22], CAZy^[23], TCDB^[24], MEROPS^[25], PHI^[26], CARD, and DFVF^[27] databases. For CAZy annotation, the number of genes assigned to glycoside hydrolases (GH), glycosyltransferases (GT), polysaccharide lyases (PL), carbohydrate esterases (CE), and auxiliary activities (AA) were recorded. Secreted proteins were predicted with SignalP v5.0^[28] and TMHMM v2.0^[29]. Secondary metabolite gene clusters were identified with antiSMASH v6.0^[30].

2.4 Pangenome Construction and Phylogenetic Analysis

Genome sequences and annotations for 15 additional *K. marxianus* strains were downloaded from NCBI RefSeq^[31]. The 15 comparator strains were selected based on the following criteria: all publicly available high-quality genomes (complete or chromosome-level assemblies) with diverse ecological niches (dairy, plant, fermented foods, clinical/environmental). Among them, strains DMKU3-1042 and CBS 6556 have been previously reported to possess moderate protease activity, serving as functional references. A complete list with isolation sources and genome statistics is provided in [Supplementary 2 Table. S1](#) Orthologous gene families were defined using OrthoFinder v2.5.4^[32]. Gene families present in all 16 strains were classified as core; those in 2 – 15 strains as accessory; and those unique to *SHY2* as strain-specific (private). Pangenome accumulation curves (Pan and Core) were generated by random sampling with 100 permutations. The Heaps model was fitted to the Pan curve to evaluate pangenome closure.

3. Result

3.1 Genome Sequencing and Assembly of *K. marxianus SHY2*

Illumina sequencing generated 31.52 million raw paired-end reads (4.76 Gb) with Q20 98.92% and Q30 97.07%. After quality filtering, 30.90 million clean reads (4.63 Gb) remained. PacBio sequencing produced 194,596 long reads totaling 3.71 Gb, with N50

19,461 bp and N90 13,275 bp. K-mer analysis (with k = 17) estimated the genome size at approximately 11.79 Mb, with heterozygosity 0.07% – 0.08% and repeat content ~ 13.0%.

Hybrid assembly yielded a final genome of 11,948,521 bp in 19 scaffolds, with a GC content of 40.23% and N50 scaffold length of 1.47 Mb (**Table 1**). BUSCO v5.4.7 assessment using the ascomycota_odb10 lineage dataset showed 86.60% complete genes (86.0% single-copy, 0.60% duplicated), 1.20% fragmented, and 12.20% missing. Although the BUSCO completeness is below the 95% threshold typically expected for high-quality fungal genomes, it is within the range reported for other *K. marxianus* assemblies (82% – 94%) and is sufficient for gene prediction and comparative pangenomic analyses, as supported by the high annotation coverage (99.70%, **Figure. 1**). The missing BUSCOs are primarily involved in secondary metabolism and environmental sensing, which may affect the recovery of certain accessory genes.

Table 1. Genome assembly statistics for *K. marxianus SHY2*

Feature	Value
Total genome size (bp)	11,948,521
Number of scaffolds	19
N50 scaffold length (bp)	1,467,832
N90 scaffold length (bp)	523,419
Guanine-cytosine content (%)	40.23
BUSCO completeness (%)	86.60

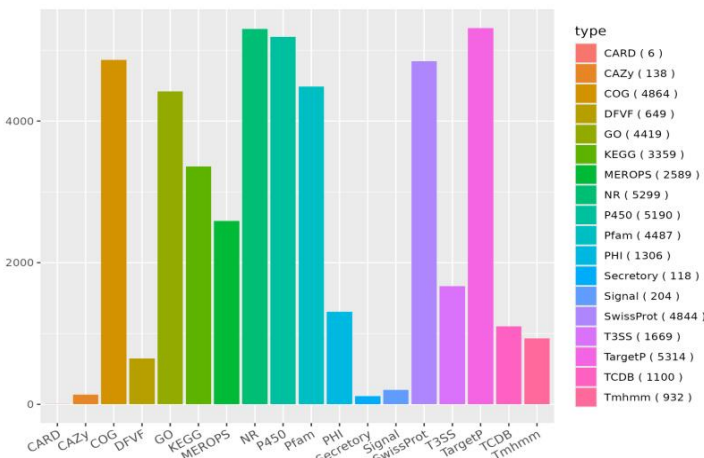


Figure. 1 Genomic features of *K. marxianus SHY2*. Functional annotation summary of the *SHY2* genome. A total of 5,312 protein-coding genes were predicted, with 99.7% of genes annotated in at least one database. Annotation rates: Nr (99.7%), Swiss-Prot (91.2%), eggNOG (91.5%), GO (83.2%), KEGG (63.2%), and Pfam (84.4%). Statistical plots were generated using Python 3.12 or R/UpSetR.

3.2 Gene Prediction and Functional Annotation

Gene prediction identified 5,312 protein-coding genes (average length 1,458 bp), 158 tRNA genes, and 18 rRNA genes. Functional annotation achieved 99.70% coverage across databases (**Figure. 1**). eggNOG classification (**Figure. 2**) revealed high proportions of genes involved in post-translational modification (7.40%), transcription (7.36%), translation (6.87%), carbohydrate metabolism (4.33%), amino acid metabolism (3.71%), and energy production (3.69%). KEGG annotation (**Supplementary 1 Figure. S1**) mapped 3,359 genes to 129 pathways, with significant enrichment in carbon metabolism, amino acid biosynthesis, and secondary metabolite biosynthesis (77 genes, 1.45%). GO analysis (**Figure. S1**) showed catalytic activity (56%) and binding (44%) as dominant molecular functions, with metabolic and

cellular processes as primary biological processes.

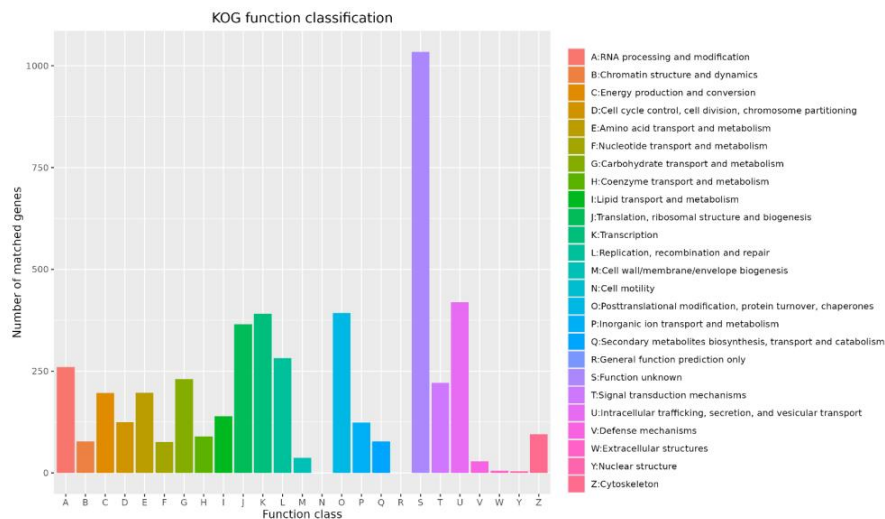


Figure. 2 Functional classification and metabolic potential of *SHY2*. KOG functional classification of *SHY2* genes. The most abundant categories include post-translational modification/protein turnover/chaperones (7.40%), transcription (7.36%), and translation/ribosomal structure (6.87%). Carbohydrate metabolism (4.33%) and amino acid metabolism (3.71%) are also highly represented. Statistical plots were generated using Python 3.12 or R/UpSetR.

Specialized databases (**Table 2**) revealed 315 CAZymes, comprising 124 glycoside hydrolases (GH), 52 glycosyltransferases (GT), 5 polysaccharide lyases (PL), 18 carbohydrate esterases (CE), and 12 auxiliary activities (AA); 1,100 transporters (TCDB), and 2,589 protease- and protease-regulatory genes (48.70% of all protein-coding genes) (according to MEROPS). This category includes active proteases (aspartic, cysteine, metallo, serine, threonine), peptidases, protease inhibitors, and non-catalytic homologs. A detailed breakdown of protease classes and a comparison with the other 15 *K. marxianus* strains are provided in [Supplementary 2 Table. S2](#) The proportion in *SHY2* is higher than the average but not an outlier when considering the broad definition used by MEROPS; therefore, the raw

count likely overestimates the genuine proteolytic capacity. CARD and DFVF did not detect any significant antibiotic resistance or virulence genes under default or relaxed parameters. This may reflect a genuine absence in this dairy isolate or limitations of the current databases (e.g., lack of representation for novel or highly divergent determinants. antiSMASH identified two putative secondary metabolite clusters: a terpene cluster (~ 21 kb) and an NRPS-like cluster (~ 38 kb). SignalP/TMHMM predicted 127 secreted proteins, many being hydrolytic enzymes; EffectorP (Department of Plant Pathology, The University of Sydney, Sydney, New South Wales, Australia) found four high-confidence effectors without typical pathogenicity domains (**Figure. 3**).

Table 2. Summary of specialized database annotations

Database	Number of genes annotated	Notes
CAZy	315	GHs (Glycoside Hydrolases): 124, GTs (Glycosyl transferases): 52, PLs (Polysaccharide Lyases): 5, CEs (Carbohydrate Esterases): 18, AAs (Auxiliary Activities): 12
TCDB	1,100	Ion channels, porters, ABC transporters
MEROPS	2,589 (48.70% of all coding genes)	Includes proteases, peptidases, inhibitors, non-active homologs; likely overestimates proteolytic activity
CARD/DFVF	0 (no significant hits)	No known resistance/virulence genes detected; database limitations apply

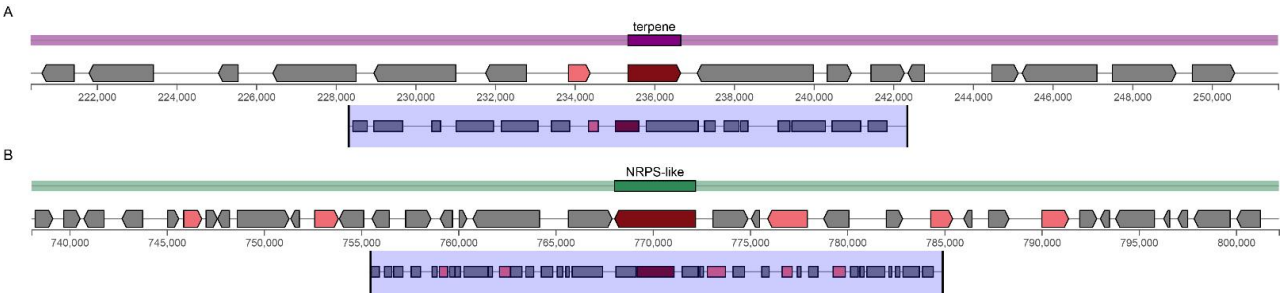


Figure. 3 Secondary metabolite biosynthesis gene clusters identified in *SHY2*. Two putative clusters were detected: (A) a terpene cluster (~21 kb) and (B) an NRPS-like cluster (~38 kb), potentially involved in the synthesis of aroma compounds or signaling molecules. Biosynthetic gene cluster figures were generated using antiSMASH based on author's sequence data.

3.3 Pangenome Datasets and Basic Features

To construct the pangenome of *K. marxianus*, 16 genomes were included, comprising the target strain *SHY2* and 15 representative strains. The selection covered multiple geographic origins and ecological niches (dairy, fermented foods, plant materials, clinical isolates), ensuring species-wide genetic diversity ([Supplementary 1 Figure. S2](#)).

Genome assembly sizes ranged from 10.42 Mb to 11.21 Mb (mean 10.87 Mb, SD 0.21 Mb), and protein-coding gene numbers ranged from 4,883 to 5,382 (mean 5,135, SD 143). Functional gene

density was concentrated in a narrow interval of 450 – 500 proteins per Mb. *SHY2* had a genome size of 10.92 Mb and 5,201 protein-coding genes, both near the population median. Functional annotation across GO, KEGG, and Pfam databases revealed high cross-strain consistency. All 16 strains shared a common backbone of pathways, including carbohydrate metabolism, amino acid metabolism, energy metabolism, and membrane transport (**Figure. 4**). *SHY2* had the highest total annotation counts (GO 19,803; KEGG 4,860; Pfam 8,185), which may reflect its high-quality assembly and thorough gene prediction rather than true biological enrichment.

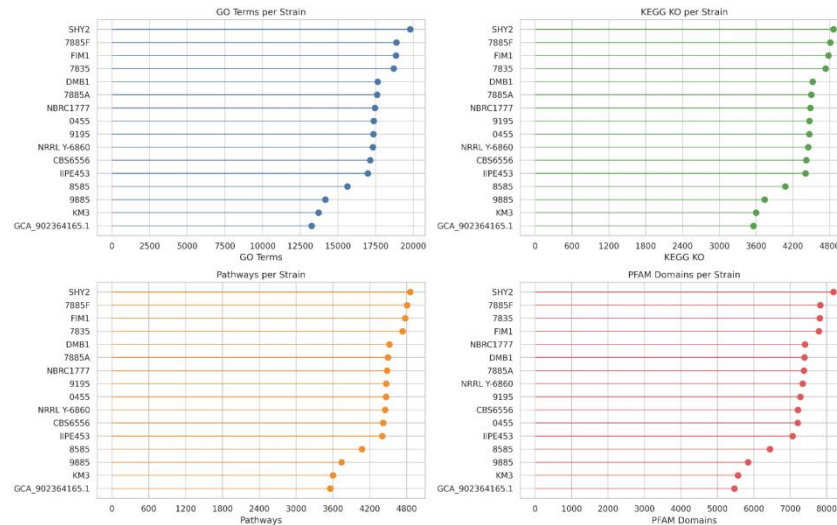


Figure. 4 Basic characteristics of the pangenome dataset. (A-C) Total GO, KEGG, and Pfam annotations per strain; (D) KEGG level-2 pathway heatmap showing cross-strain consistency. Statistical plots were generated using Python 3.12 or R/UpSetR.

3.4 Pangenome Composition and Core-non-core Stratification

Based on gene family clustering of the 16 strains, pangenome accumulation curves were constructed. At 16 strains, the total gene families (Pan) reached approximately 5,237 with a marginal gain near zero; the core gene families (Core) stabilized at approximately 1,807. Heaps model fitting gave γ near 0 and R^2 near 1, indicating a nearly closed pangenome while the core was still converging (**Figure. 5A**).

In terms of stratification, the core set (1807 families) formed a conserved backbone of essential cellular processes; the non-core set (soft-core, dispensable, private) formed a “plastic periphery”. The private set for *SHY2* comprised 82 gene families that were absent from all other 15 strains. GO enrichment analysis of these private families revealed significant enrichment for terms related to transmembrane transport (GO: 0055085, $P = 0.003$) and carbohydrate metabolic process (GO: 0005975, $P = 0.011$), suggesting that *SHY2*'s unique genes may contribute to its adaptation to dairy environments and utilization of diverse carbon sources. No known protease genes were found among the private families, indicating that the high protease-related gene count in *SHY2* is due to expansion of existing families rather than acquisition of unique ones.

The UpSet plot (**Figure. 5**) indicated that soft-core and dispensable sets carried most of the “shared but not universal” functional content, appearing as multiple medium-sized intersection peaks commonly found in carbohydrate metabolism branches, membrane transport, redox homeostasis, and stress response modules. *SHY2* ranked high in private gene families (its private intersection bar was significantly above the population median), suggesting that its individualized traits may be driven by a small number of functionally

focused private families.

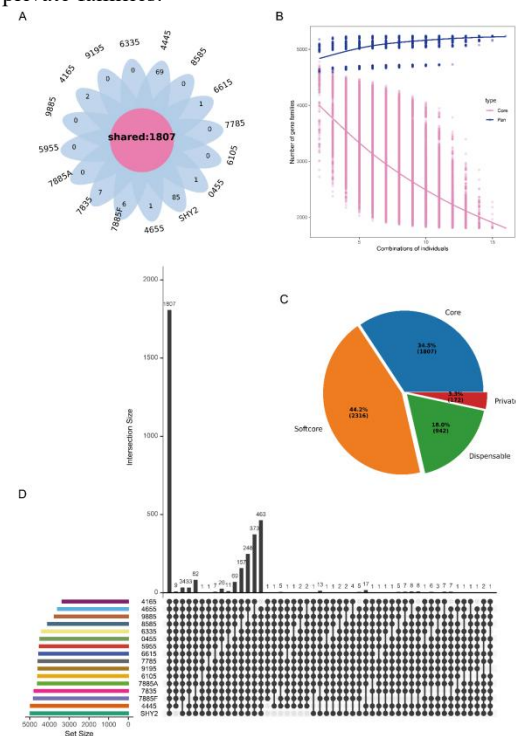


Figure. 5 Pangenome composition and core/non-core stratification. (A) Flower plot showing core and private families; (B) Pangenome accumulation curve (Pan, blue) and core accumulation curve (Core, pink); (C) Pie chart of gene family

categories; (D) UpSet plot showing the intersection patterns of non-core gene families. Statistical plots were generated using Python 3.12 or R/UpSetR.

4. Discussion

In this study, a high-quality genome assembly of *Kluyveromyces marxianus* *SHY2* (11.95 Mb, 5,312 protein-coding genes, BUSCO 86.60%) was obtained. The BUSCO completeness is lower than the 95% benchmark typically expected for fungal genomes, but it falls within the range reported for other *K. marxianus* assemblies (82% – 94%) [Supplementary 2 Table. S3](#). This level is adequate for gene prediction and comparative pangenomics, as supported by the high annotation coverage (99.70%). However, the missing BUSCOs (12.20%) may lead to under-recovery of certain accessory or strain-specific genes, and this limitation should be considered when interpreting negative findings.

The genome contains a large number of MEROPS-annotated protease- and protease-regulatory genes (2,589, 48.70% of coding genes)^[25]. As noted in the results, this proportion includes not only active proteases but also peptidases, inhibitors, and non-catalytic homologs. A stricter Pfam domain-based count yields approximately 600 genuine protease domains^[22]. Therefore, the raw MEROPS number is not direct evidence of high proteolytic activity. Nevertheless, the presence of multiple oligopeptide transporters^[24] and xylanase genes (GH10, GH11) supports the strain's ability to utilize protein-derived peptides and plant-derived polysaccharides, consistent with but not in proteolysis. Thus, the high protease-related gene counts to its dairy origin^[33]. The identification of two secondary metabolite biosynthesis gene clusters (terpene and NRPS-like) aligns with the strain's isolation from fermented food products and suggests potential for flavor compound production^[34]. The absence of detectable antibiotic resistance or virulence genes (CARD and DFVF) is encouraging for food/feed applications, but this should be verified by phenotypic assays given database limitations^[35]. The pangenome analysis showed that *K. marxianus* has a nearly closed pangenome ($\gamma \approx 0$), similar to *Saccharomyces cerevisiae*^[36]. This means that the core genome is well conserved and that most gene families have already been sampled with the 16 strains. *SHY2* contributed 82 private gene families, enriched in transmembrane transport and carbohydrate metabolism, but not in proteolysis. Thus, the high protease-related gene count in *SHY2* is due to expansion of existing families rather than unique genes. Direct comparison with previously characterized high-protease strains of *K. marxianus* (e.g., NBRC 1777) was not possible because genome assemblies for those strains were not publicly available at the time of analysis. Future studies should include such strains to confirm whether the observed gene content truly correlates with high proteolytic activity.

Despite the comprehensive genomic and pangenomic insights provided in this study, several limitations must be acknowledged. First, the genome assembly achieved a BUSCO completeness of 86.60%, which, while adequate for basic comparative genomics, may result in the under-recovery of certain accessory or strain-specific genes. Second, the reported high proportion of protease-related genes (48.70%) based on the MEROPS database represents an overestimation, as this broad classification encompasses non-catalytic homologs, peptidases, and protease inhibitors rather than exclusively active proteases. Most importantly, all functional interpretations — particularly the claims regarding the strain's high proteolytic potential and biological safety — are derived entirely from *in silico* genomic predictions and lack empirical verification. Furthermore, the pangenomic comparative analysis did not include direct comparisons

with previously confirmed high-protease reference strains (e.g., NBRC 1777), limiting the ability to draw definitive genotype-phenotype correlations regarding proteolysis.

To substantiate the genomic findings and fully elucidate the industrial potential of *K. marxianus* *SHY2*, future studies should prioritize targeted experimental validation. Biochemical validation through *in vitro* enzyme activity assays is essential to accurately quantify the authentic proteolytic capacity of the strain. Additionally, laboratory-scale fermentation tests and industrial application trials are required to evaluate its actual performance, growth characteristics, and enzyme yield in protein-rich substrates. Given the inherent limitations of current antimicrobial resistance and virulence databases, rigorous phenotypic safety assessments must be conducted to guarantee the biosafety of the strain for food and feed applications. Finally, analytical chemistry approaches should be employed to verify the functional expression of the predicted terpene and NRPS-like secondary metabolite gene clusters and their potential contribution to flavor compound production.

5. Conclusions

Whole-genome sequencing of *K. marxianus* *SHY2* yielded an assembly of 11.95 Mb encoding 5,312 genes with a GC content of 40.23% and BUSCO completeness of 86.60%. Functional annotation revealed genes involved in carbohydrate metabolism, amino acid metabolism, and transport systems. A large number of MEROPS-classified protease- and protease-regulatory-related genes (2589, 48.70% of all coding genes) were observed, but this likely overestimates genuine proteolytic capacity because the MEROPS count includes inhibitors and non-active homologs; stricter domain analysis and experimental validation (enzyme activity assays) are required before claiming high proteolytic potential. Two secondary metabolite biosynthesis gene clusters (terpene and NRPS-like) were identified, and no significant matches to known antibiotic resistance or virulence determinants were detected in CARD/DFVF, suggesting biosafety for food/feed use (with the caveat of database limitations).

A pangenome dataset of 16 *K. marxianus* strains was constructed, showing high consistency in genome size (10.42 Mb – 11.21 Mb) and protein-coding gene number (4,883–5,382). All strains share a common backbone of pathways (carbohydrate metabolism, amino acid metabolism, energy metabolism, and membrane transport). Pangenome analysis revealed that the total gene family pool (Pan) reaches approximately 5,237 families and is near closure, while the core gene family set stabilizes at approximately 1,807 families. Non-core families account for over 70% of the total gene repertoire. *SHY2* contains 82 private gene families, enriched in transmembrane transport and carbohydrate metabolism, but not in proteolysis. This indicates that the high protease-related gene count, in *SHY2* is due to expansion of existing families. All functional interpretations require experimental confirmation via enzyme assays, fermentation tests, and industrial trials before any claims of high proteolytic potential can be substantiated. This study provides a genomic and pangenomic foundation for future hypothesis-driven research on *K. marxianus* *SHY2*.

Author Contribution

Quan Qiu: Writing, original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. Ying Li: Writing, original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. Zhichun Zhan: Validation, Investigation, Data

curation. Ying Zhou: Investigation, Data curation. Lingfang Gu: Investigation, Data curation. Qijun Wang: Investigation, Data curation. Yicheng Yang: Investigation, Data curation. Yunxiang Liang: Supervision, Conceptualization. Yingjun Li: Writing review & editing, Writing original draft, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Supplementary Information

Supplementary material associated with this article can be found, in the online version, at [Supplementary Information 1](#) and [Supplementary Information 2](#).

Conflicts of Interest

No potential conflict of interest was reported by the authors.

Funding

This research was funded by Hubei Province Science and Technology Talent Service Enterprise Project (NO.2023DJC127) and Fundamental Research Funds for the Central Universities (NO.2662022SKPY001).

References

- [1] Fonseca G G, Heinzle E, Wittmann C, *et al.* The yeast *Kluyveromyces marxianus* and its biotechnological potential. *Applied Microbiology and Biotechnology*. 2008, 79 (3): 339-354. [[Google Scholar](#)] [[CrossRef](#)]
- [2] Lane M M, Morrissey J P. *Kluyveromyces marxianus*: a yeast emerging from its sister's shadow. *Fungal Biology Reviews*. 2010, 24 (1-2): 17-26. [[Google Scholar](#)] [[CrossRef](#)]
- [3] Morrissey J P, Etschmann M M W, Schrader J, *et al.* Cell factory applications of the yeast *Kluyveromyces marxianus* for the biotechnological production of natural flavour and fragrance molecules. *Yeast*. 2015, 32 (1): 3-16. [[Google Scholar](#)] [[CrossRef](#)]
- [4] Nonklang S, Abdel-Banat B M A, Cha-Aim K, *et al.* High-temperature ethanol fermentation and transformation with linear DNA in the thermotolerant yeast *Kluyveromyces marxianus* DMKU3-1042. *Applied and Environmental Microbiology*. 2008, 74 (24): 7514-7521. [[Google Scholar](#)] [[CrossRef](#)]
- [5] Lertwattanasakul N, Kosaka T, Hosoyama A, *et al.* Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. *Biotechnology for Biofuels*. 2015, 8 (1): 47. [[Google Scholar](#)] [[CrossRef](#)]
- [6] Tettelin H, Maignani V, Cieslewicz M J, *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*. 2005, 102 (39): 13950-13955. [[Google Scholar](#)] [[CrossRef](#)]
- [7] Ho S S, Urban A E, Mills R E. Structural variation in the sequencing era. *Nature Reviews Genetics*. 2020, 21 (3): 171-189. [[Google Scholar](#)] [[CrossRef](#)]
- [8] Wilson, K. Preparation of genomic DNA from bacteria. *Current Protocols in Molecular Biology*. 2001, 56 (1): 2.4.1-2.4.5. [[Google Scholar](#)] [[CrossRef](#)]
- [9] Koren S, Walenz B P, Berlin K, *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*. 2017, 27 (5): 722-736. [[Google Scholar](#)] [[CrossRef](#)]
- [10] Walker B J, Abeel T, Shea T, *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*. 2014, 9 (11). [[Google Scholar](#)] [[CrossRef](#)]
- [11] Simão F A, Waterhouse R M, Ioannidis P, *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015, 31 (19): 3210-3212. [[Google Scholar](#)] [[CrossRef](#)]
- [12] Stanke M, Keller O, Gunduz I, *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, 2006, 34 (suppl_2): W435-W439. [[Google Scholar](#)] [[CrossRef](#)]
- [13] Keilwagen J, Wenk M, Erickson J L, *et al.* Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research*. 2016, 44 (9): 89. [[Google Scholar](#)] [[CrossRef](#)]
- [14] Haas B J, Salzberg S L, Zhu W, *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology*. 2008, 9 (1): 7. [[Google Scholar](#)] [[CrossRef](#)]
- [15] Chan P P, Lowe T M. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods in Molecular Biology*. 2019: 1-14. [[Google Scholar](#)] [[CrossRef](#)]
- [16] Lagesen K, Hallin P, Rødland E A, *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*. 2007, 35 (9): 3100-3108. [[Google Scholar](#)] [[CrossRef](#)]
- [17] Camacho C, Coulouris G, Avagyan V, *et al.* BLAST+: architecture and applications. *BMC Bioinformatics*. 2009, 10 (1): 421. [[Google Scholar](#)] [[CrossRef](#)]
- [18] Boeckmann B, Bairoch A, Apweiler R, *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*. 2003, 31 (1): 365-370. [[Google Scholar](#)] [[CrossRef](#)]
- [19] Kanehisa, M., KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*. 2016, 44 (D1): D457-D462. [[Google Scholar](#)] [[CrossRef](#)]
- [20] Ashburner M, Ball C A, Blake J A, *et al.* Gene ontology: tool for the unification of biology. *Nature Genetics*. 2000, 25 (1): 25-29. [[Google Scholar](#)] [[CrossRef](#)]
- [21] Huerta-Cepas J, Szklarczyk D, Forslund K, *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations. *Nucleic Acids Research*. 2016, 44 (D1): D286-D293. [[Google Scholar](#)] [[CrossRef](#)]
- [22] Mistry J, Chuguransky S, Williams L, *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Research*. 2021, 49 (D1): D412-D419. [[Google Scholar](#)] [[CrossRef](#)]
- [23] Lombard V, Golaconda Ramulu H, Drula E, *et al.* The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*. 2014, 42 (D1): D490-D495. [[Google Scholar](#)] [[CrossRef](#)]
- [24] Saier Jr M H, Reddy V S, Tsu B V, *et al.* The transporter classification database (TCDB): recent advances. *Nucleic Acids Research*. 2016, 44 (D1): D372-D379. [[Google Scholar](#)] [[CrossRef](#)]
- [25] Rawlings N D, Barrett A J, Thomas P D, *et al.* The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research*. 2018, 46 (D1): D624-D632. [[Google Scholar](#)] [[CrossRef](#)]
- [26] Urban M, Pant R, Raghunath A, *et al.* The Pathogen-Host interactions database (PHI-base): additions and future developments. *Nucleic Acids Research*. 2015, 43 (D1): D645-D655. [[Google Scholar](#)] [[CrossRef](#)]
- [27] Lu T, Yao B, Zhang C. DFVF: database of fungal virulence factors.

- Database. 2012, 2012. [[Google Scholar](#)] [[CrossRef](#)]
- [28] Alcock B P, Raphenya A R, Lau T T Y, *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*. 2020, 48 (D1): D517-D525. [[Google Scholar](#)] [[CrossRef](#)]
- [29] Krogh A, Larsson B, Von Heijne G, *et al.* Predicting transmembrane protein topology with a hidden markov model. *Journal of Molecular Biology*. 2001, 305 (3): 567-580. [[Google Scholar](#)] [[CrossRef](#)]
- [30] Blin, K., AntiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research*. 2021, 49 (W1): W29-W35. [[Google Scholar](#)] [[CrossRef](#)]
- [31] O'Leary, N.A., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. 2016, 44 (D1): D733-D745. [[Google Scholar](#)] [[CrossRef](#)]
- [32] Edgar, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 2004, 32 (5): 1792-1797. [[Google Scholar](#)] [[CrossRef](#)]
- [33] Becerra-Rodriguez, C. S. Marsit, and V. Galeote, Diversity of Oligopeptide Transport in Yeast and Its Impact on Adaptation to Winemaking Conditions. *Frontiers in Genetics*. 2020, 11: 602. [[Google Scholar](#)] [[CrossRef](#)]
- [34] Dzialo M C, Park R, Steensels J, *et al.* Physiology, ecology and industrial applications of aroma formation in yeast. *FEMS microbiology reviews*. 2017, 41 (Supp_1): S95-S128. [[Google Scholar](#)] [[CrossRef](#)]
- [35] Wolfe, B.E. Are fermented foods an overlooked reservoir of antimicrobial resistance? *Current Opinion in Food Science*. 2023, 51. [[Google Scholar](#)] [[CrossRef](#)]
- [36] Peter J, De Chiara M, Friedrich A, *et al.* Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*. 2018, 556 (7701): 339-344. [[Google Scholar](#)] [[CrossRef](#)]