Received: 21 May 2023 | Accepted: 24 September 2023

DOI: 10.1111/1751-7915.14354

MICROBIAL BIOTECHNOLOGY Applied Microbiology International

RESEARCH ARTICLE

Integrated genomics and phenotype microarray analysis of *Saccharomyces cerevisiae* industrial strains for rice wine fermentation and recombinant protein production

Ye Ji Son[1] | **Min-Seung Jeon**[1] | **Hye Yun Moon**[1] | **Jiwon Kang**[1] | **Da Min Jeong**[1] | **Dong Wook Lee**[1] | **Jae Ho Kim**[2] | **Jae Yun Lim**[3] | **Jeong-Ah Seo**[3] | **Jae-Hyung Jin**[4] | **Yong-Sun Bahn**[4] | **Seong-il Eyun**[1] | **Hyun Ah Kang**[1]

[1]Department of Life Science, Chung-Ang University, Seoul, Korea

[2]Korea Food Research Institute, Wanju-Gun, Jeollabukdo, Korea

[3]School of Systems Biomedical Science, Soongsil University, Seoul, Korea

[4]Department of Biotechnology, Yonsei University, Seoul, Korea

Correspondence
Seong-il Eyun and Hyun Ah Kang, Department of Life Science, Chung-Ang University, Seoul, 06974, Korea.
Email: eyun@cau.ac.kr and hyunkang@cau.ac.kr

Funding information
Chung-Ang University Graduate Research Scholarship Grants in 2021; National Research Foundation of Korea, Grant/Award Number: NRF2018R1A5A1025077; Rural Development Administration, Republic of Korea, Cooperative Research Program for Agriculture Science & Technology Development, Grant/Award Number: PJ01710102

Abstract

The industrial potential of *Saccharomyces cerevisiae* has extended beyond its traditional use in fermentation to various applications, including recombinant protein production. Herein, comparative genomics was performed with three industrial *S. cerevisiae* strains and revealed a heterozygous diploid genome for the 98-5 and KSD-YC strains (exploited for rice wine fermentation) and a haploid genome for strain Y2805 (used for recombinant protein production). Phylogenomic analysis indicated that Y2805 was closely associated with the reference strain S288C, whereas KSD-YC and 98-5 were grouped with Asian and European wine strains, respectively. Particularly, a single nucleotide polymorphism (SNP) in *FDC1*, involved in the biosynthesis of 4-vinylguaiacol (4-VG, a phenolic compound with a clove-like aroma), was found in KSD-YC, consistent with its lack of 4-VG production. Phenotype microarray (PM) analysis showed that KSD-YC and 98-5 displayed broader substrate utilization than S288C and Y2805. The SNPs detected by genome comparison were mapped to the genes responsible for the observed phenotypic differences. In addition, detailed information on the structural organization of Y2805 selection markers was validated by Sanger sequencing. Integrated genomics and PM analysis elucidated the evolutionary history and genetic diversity of industrial *S. cerevisiae* strains, providing a platform to improve fermentation processes and genetic manipulation.

INTRODUCTION

Yeasts have been used for thousands of years in food and fermentation processes to produce alcoholic beverages and breads (Copetti, 2019). In addition, yeasts have been used for producing a great variety of biomolecules applied to chemicals, fuels, food and pharmaceuticals and are currently one of the most used hosts for producing recombinant proteins and metabolites (Kavščeк et al., 2015; Kim et al., 2015). Among several yeast species, the traditional yeast *Saccharomyces cerevisiae*, has been most frequently used not only for traditional fermentation, such as baking, brewing and winemaking, but also to produce bioethanol, recombinant proteins

Ye Ji Son and Min-Seung Jeon contributed equally to this work.

and metabolites (Parapouli et al., 2020). Biochemical and genomic studies on *S. cerevisiae*, a eukaryotic model organism, have greatly contributed to much of our understanding of eukaryotic biology. The *S. cerevisiae* S288C strain, which is the ancestor to many commonly used yeast laboratory strains (Engel et al., 2013), was the first eukaryotic genome to be completely sequenced (Goffeau et al., 1996). Since then, many functional genomic studies, such as transcriptomics using the S288C genome as a reference sequence, have greatly enriched our knowledge of how yeast cells respond to and resist various environmental stresses (Capaldi et al., 2008; Gasch et al., 2000). However, in many ways, the information that has been gathered from the S288C strain cannot always be extrapolated to other *S. cerevisiae* strains because of their diverse genomes and phenotypes (Kvitek et al., 2008).

Today, many different *S. cerevisiae* strains are exploited for specific fermentation and industrial processes. At present, genomes of more than 1000 *S. cerevisiae* strains of different origins, including natural and human-related isolates, have been sequenced (Borneman & Pretorius, 2015; Liti et al., 2009; Peter et al., 2018). Comparing a wide variety of genomes helps to clarify the natural history of yeast populations and allows the identification of genomic elements that play important roles in their metabolic activities and physiological characteristics essential for biotechnical applications (Borneman et al., 2008, 2013; Strope et al., 2015). With more information on genome sequences, there is an increasing number of studies focusing on the genotype–phenotype relationship to explore the phenotypic diversity at the genome level (Gallone et al., 2016).

Several omics-based technologies allow for global analysis of the important macromolecules of cells that convey the information flow from DNA to RNA to protein. The information initially encoded in the genome is ultimately displayed at the cellular level as cellular traits or phenotypes. As a tool for live cell analysis (phenomics), phenotype microarray (PM) techniques that can continuously monitor and record cell responses in all array wells were developed as a semi-high throughput assay for the characterization and monitoring of microbial cellular phenotypes (Bochner, 2003). PM techniques have gained increased attention as complementary tools to next generation sequencing (NGS) for the characterization of various microorganisms, including *S. cerevisiae* (Kang et al., 2019; Wimalasena et al., 2014). A PM technique was also applied for screening novel yeast strains with the ability to metabolize compounds present in pyrolysis bio-oil (Kostas et al., 2019). Recently, the PM was used to investigate the industrial potential of the cold-tolerance *S. cerevisiae* Cheongdo strain, an isolate from frozen peach samples, using 192 different carbon sources (Jung et al., 2021).

In this study, to investigate the genetic basis of phenotypic variation in industrial strains of *S. cerevisiae*, we carried out the complete whole-genome sequencing of *S. cerevisiae* strains KSD-YC and 98-5, which are used for fermentation of Korean traditional rice wine, 'Makgeolli' (Kim et al., 2014; Shin et al., 2019) and *S. cerevisiae* strain Y2805, which is widely used in Korea as a host strain for the production of recombinant proteins and metabolites (Table S1). In addition to the comparative genomics analysis using the laboratory strain S288C as a reference genome, we performed PM analysis and mapped in the single nucleotide polymorphisms (SNPs) responsible for the observed differences in the PM data. The integrated genomics and PM analysis data elucidated the evolutionary history and genotype–phenotype relationship to explain the phenotypic diversity at the genome level.

EXPERIMENTAL PROCEDURES

Yeast strains, culture conditions and primers

The yeast strains used in this study are listed in Table 1. Yeast cells were generally cultured in a YPD medium (1% yeast extract, 2% bacto peptone and 2% glucose) at 28°C. The primers used for PCR amplification and sequencing of genetic markers in this study are listed in Table S2.

Whole-genome (WG) sequencing, assembly and annotation

To obtain high-quality genomic DNA for WG sequencing of *S. cerevisiae* strains, genomic DNA was extracted from spheroplasts and harvested by spooling, as previously described (Jeong et al., 2022). WG sequencing was performed using different sequencing techniques, depending on the service provided by Theragen Bio (Korea), which had initially carried out genome sequencing analysis using PacBio RS II and Illumina HiSeq 2500, but recently updated the sequencing platform with PacBio Sequel and Illumina NovaSeq 6000. For de novo WG sequencing of the 98-5 and KSD-YC strains, long, short and long-mated pair reads were produced using PacBio RS II and Illumina HiSeq 2500 sequencing technologies, respectively, according to the manufacturer's instructions. The genomic raw data of 98-5 and KSD-YC were assembled through the PacBio Corrected Reads (PBcR) assembly pipeline (ver. Wgs-8.3) and Hierarchical Genome Assembly Process (HGAP) (ver. 3.0) with the estimated genome size (25 Mbp) (Berlin et al., 2015). The draft assemblies of 98-5 and KSD-YC were further polished with Illumina data by Pilon (ver. 1.2.2) (Walker et al., 2014). For de novo WG sequencing of the Y2805 strain, long and short pair reads were produced using the PacBio Sequel

TABLE 1 *Saccharomyces cerevisiae* strains analysed in this study.

S. cerevisiae strains	Genotypes	Characteristics	Sources /References
S288C	*MATα SUC2 gal2 mal2 mel flo1 flo8-1 hap1 ho bio1 bio6*	Haploid laboratory strain, Reference genome	ATCC 204508
BY4741	*Matta; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0*	Haploid laboratory strain	ATCC 4040002
BY4743	*MATa/MATα; his3Δ1/his3Δ1; leu2Δ0 / leu2Δ0; met15Δ0/MET15; LYS2/ lys2Δ0; ura3Δ0/ura3Δ0*	Diploid laboratory strain	ATCC 4040005
CEN.PK2-1C	*MATa ura3-52 leu2-3112 trp1-289 his3Δ MAL2-8c SUC2*	Haploid laboratory strain	EUROSCARF:30000A
KSD-YC	Not determined	Diploid strain used for fermentation of commercial Korean traditional rice wine	Kooksoondang Brewery Co., Ltd /KACC 93276P
98-5	Not determined	Diploid strain used for fermentation of commercial Korean traditional rice wine	KFRI /PRJNA348390
Y2805	*MATa pep:HIS3 prb1-Δ1.6R can1 his3-Δ200 ura3-52*	Haploid industrial strain used for production of recombinant proteins	KRIBB/KCTC 37201

Abbreviations: ATCC, American Type Culture Collection; EUROSCARF, European Saccharomyces Cerevisiae Archive for Functional Analysis; KACC, Korean Agricultural Culture Collection; KCTC, Korean Collection for Type Cultures; KFRI, Korea Food Research Institute; KRIBB, Korea Research Institute of Bioscience and Biotechnology.

and Illumina NovaSeq 6000. The genomic reads from the PacBio Sequel were assembled with Canu (ver. 2.2) (Koren et al., 2017), and paired-end Illumina reads with high accuracy were used in the polishing process for increasing the quality of the draft assembly (Jung, Jeon, et al., 2020; Jung, Ventura, et al., 2020). The Illumina reads were mapped into the erroneous draft assembly, which generated the binary alignment map (BAM) file, and the resulting alignment information was subjected to processing using Pilon (ver. 1.24), yielding the final assembly. The quality metrics of the WG reconstructions generated from each assembly pipeline were measured, and the chromosomal structure underwent an assembly evaluation (Jeon et al., 2023). The ratios of completeness at the gene level were subsequently scored using Benchmarking Universal Single-Copy Orthologue (BUSCO) (ver. 5.3.0) with reference to the ascomycota_odb10 data set, generating results regarding the number of single or duplicated complete genes, fragmented genes and missing genes (Seppey et al., 2019). For the functional annotation for predicting protein-encoding genes, the soft-masked assembly files were submitted to the standalone gene prediction process using Augustus (ver. 3.3.3) and InterProScan (ver. 5.52–86.0) (Jones et al., 2014; Stanke et al., 2008) and analysed through the Funannotate (ver. 1.8.9) pipeline (Huerta-Cepas et al., 2018). The additional RNA predictions were conducted by using Infernal (ver. 1.1.4) (Nawrocki & Eddy, 2013).

Comparative genomics

Comparative genomics analysis was carried out with the alignment process using Bowtie2 (ver. 2.4.1) (Langmead & Salzberg, 2012). To compare SNPs and insertions and deletions (INDELs), the Illumina reads from three strains (Y2805, KSD-YC and 98-5) were mapped to the *S. cerevisiae* reference genome (S288C). FreeBayes (ver. 1.3.6) was used to examine the nucleotide variations with variant calling depth and degree of zygosity (Garrison & Marth, 2012). The gene copy numbers were analysed using Funannotate (ver. 1.8.9) and the SNPs and INDELs search was processed by MUMmer (ver. 4.0.0rc1) (Kurtz et al., 2004). Our assembly files of Y2805, KSD-YC and 98-5 genomes were used in searching INDELs, heterozygous SNPs and homozygous SNPs using S288C genome as reference and for the genome comparison between KSD-YC and K7. Read mapping depth was measured by samtools (ver. 1.10) and shown through the visualization process by Circos (ver. 0.69-8) with the information of SNPs and INDELs (Krzywinski et al., 2009; Li et al., 2009).

Phylogenetic tree and genome structure analysis

The phylogenetic tree analysis was conducted using 13 concatenated genes, which were reported as a gene set of the strain-level classification in *S. cerevisiae* (Ramazzotti et al., 2012). Orthologous sequences from the 66 total *Saccharomyces* strains were aligned with using MAFFT (ver. 7.475) (Katoh & Standley, 2013) and sequence alignments were concatenated by custom Perl script (Eyun, 2017). Phylogenetic relationships were reconstructed using maximum-likelihood with the JTT + F + I model (JTT matrix with gamma variation and invariable sites) using raxml-ng (ver. 1.2.0). The best evolutionary

model for the tree construction was inferred by IQ-TREE (ver. 2.1.4). Non-parametric bootstrapping with 1000 pseudo-replicates was used to estimate the confidence of branching topology for the maximum-likelihood (Nguyen et al., 2014). The phylogenetic tree was visualized by FigTree (ver. 1.4.4) (http://tree.bio.ed.ac.uk/software/figtree). The WG structure comparisons were conducted by starting with the chromosome-level multiple sequence alignment, NUCleotide MUMmer (NUCmer) (ver. 3.1). The pairwise alignments and multiple sequence alignment of the WG were subsequently visualized using MUMmerplot (ver. 3.5) and Integrative Genomics Viewer (IGV) (ver. 2.8.0) (Thorvaldsdottir et al., 2012).

PM analysis

The Biolog Phenotype Microarray plates (MicroPlate™), the preconfigured 96 well plates containing different substrates, were purchased from Biolog. Inc (Hayward): PM1 and PM2 (carbon sources), PM3 (nitrogen sources), PM4 (phosphorus and sulphur sources), PM5 (biosynthesis pathway end products and nutrient supplements) and PM9 (osmotic stress). The various substrates can be accessed through the Biolog website (http://www.biolog.com/products/metabolic-characterization-microplates/microbial-phenotype/). Yeast cells were incubated overnight in 2 mL of YPD medium and washed with distilled water. The preparation of the inoculating fluids was performed as specified in Biolog's instructions. Afterward, the cells were suspended in 15 mL of NS solution (0.05 mM adenine HCl, 0.01 mM L-histidine HCl monohydrate, 0.1 mM L-leucine, 0.05 mM L-lysine HCl, 0.025 mM L-methionine, 0.025 mM L-tryptophan, 0.03 mM uracil) at an initial cell optical density (OD) of 0.02 at 600 nm. Next, 0.25 mL of the cell suspension were added to 11.75 mL of PM inoculating fluids. The PM plates were then inoculated with 100 μL of the cell suspension per well. The inoculated microplates were incubated in an OmniLog reader (Biolog, Hayward) for 7 days and colour changes were automatically recorded every 15 min using a charge-coupled device camera and converted into OmniLog units. Grouping the growth signals was implemented in R (ver. 4.1.2) (*Bird Hippie*, https://www.r-project.org/) with the pipeline proposed by Vehkala et al. (2015) built upon the opm package (ver. 1.3.77) (Vaas et al., 2013). Each experiment was performed in duplicate.

4-vinylguaiacol (4-VG) production analysis using solid-phase microextraction with gas chromatography–mass spectrometry (SPME/GC–MS)

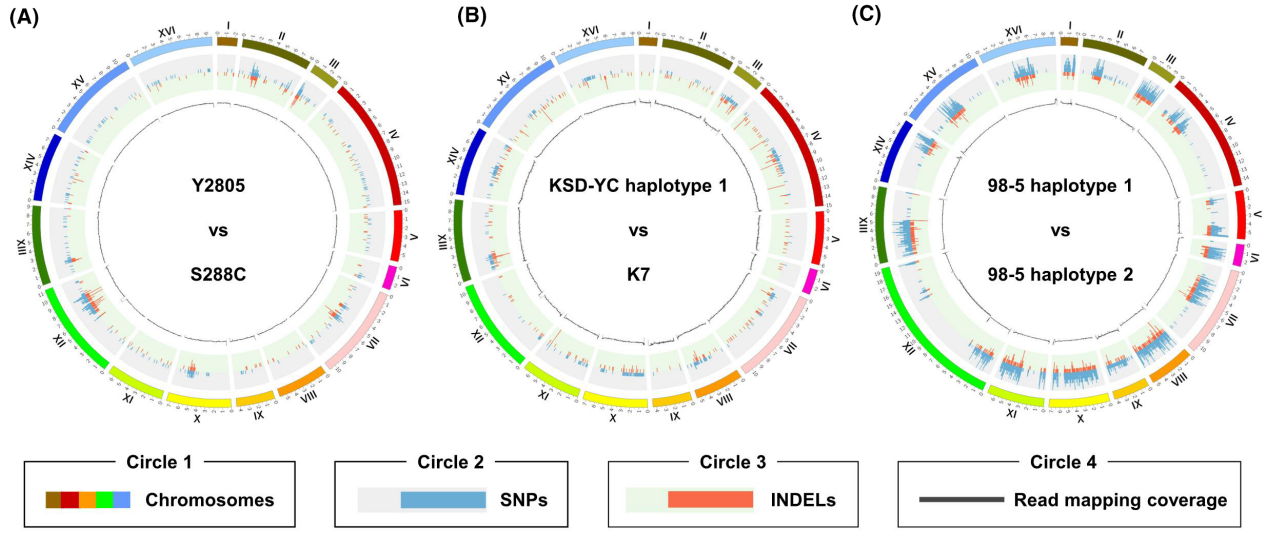
The production capability of 4-VG was analysed by cultivating the yeast cells in YPD in the presence of 50 ppm ferulic acid (Suezawa & Suzuki, 2007). The culture supernatants were transferred into glass vials with polytetrafluoroethylene (PTFE)/silicone septa (Supelco) and analysed as described previously (Jeong et al., 2022) using a Prep And Load (PAL) automated GC sampler (Agilent Technologies). Briefly, a gas chromatograph-5977E quadrupole mass selective detector (Agilent Technologies; 7820A series) was combined with the HP-INNOWax GC column (Agilent Technologies; 19,091 N-133; 30 m length × 250 μm i.d. × 0.25 μm) using helium as the carrier gas at a flow rate of 1.0 mL/min. The mass spectra of the volatile compounds were acquired from *m/z* 33–250 at a fragment voltage of 70 eV and identified using a library search (National Institute of Standards and Technology, NIST ver. 11).

RESULTS

WG analysis of *S. cerevisiae* 98-5, KSD-YC and Y2805 strains

The high-quality WG sequence information of the three industrial *S. cerevisiae* strains, 98-5, KSD-YC and Y2805, were generated by Illumina and PacBio, and the obtained contigs were assembled at the chromosome level (Figure 1; Table S3). The 12.14 Mb Y2805 genome was assembled into 16 supercontigs, consistent with the 16 chromosomes of the reference strain S288C genome (Figure 1A). The haploid genome of Y2805 showed highly conserved synteny without any chromosomal rearrangement when compared to the S288C genome (Figure 1D), thus displaying a very low level of SNPs in the annotated genes between the two strains (Table S5). The functional annotation of the Y2805 genome by Funannotate identified a similar number of genes between Y2805 and S288C (5769 and 5735, respectively), supporting the high quality of genome assembly (Tables S3 and S4).

The genomes of the KSD-YC and 98-5 strains were diploid due to their total length (25.6 and 24.04 Mb, respectively) and numbers of finally assembled supercontigs (32) (Figure 1B,C), which was about 2-fold compared to the haploid genomes of Y2805 and S288C and is thus consistent with the ploidy analysis data using flow cytometry (Figure S1). The diploid KSD-YC genome consists of two nearly identical genome copies with low heterozygosity between each haplotype (Table S5). Interestingly, the 98-5 diploid genome showed a significantly high heterozygosity with an uneven distribution between each haplotype (Figure 1C; Table S5). Our de novo WG data analysis indicated that the rice wine strains, KSD-YC and 98-5, are heterozygous diploids, which resulted from hybridization between two inter-strains that diverged from a common ancestor. In the chromosome synteny analysis with the reference S288C genome (Figure 1D), the genomes of all three industrial strains exhibited highly conserved synteny, except a large inversion (383 ~ 471 kb) in



- **Information on the mapping and sequencing coverages of Y2805, KSD-YC, and 98-5 genomes**

	Y2805 vs S288C	KSD-YC hap1 vs K7	98-5 hap1 vs hap2
SNPs	3,993	1,036	19,967
INDELs	1,290	965	2,171
Average read mapping coverage (Illumina)	204.559	118.413	80.4841
Sequencing coverage (Illumina)	235.39	169.55	110.25

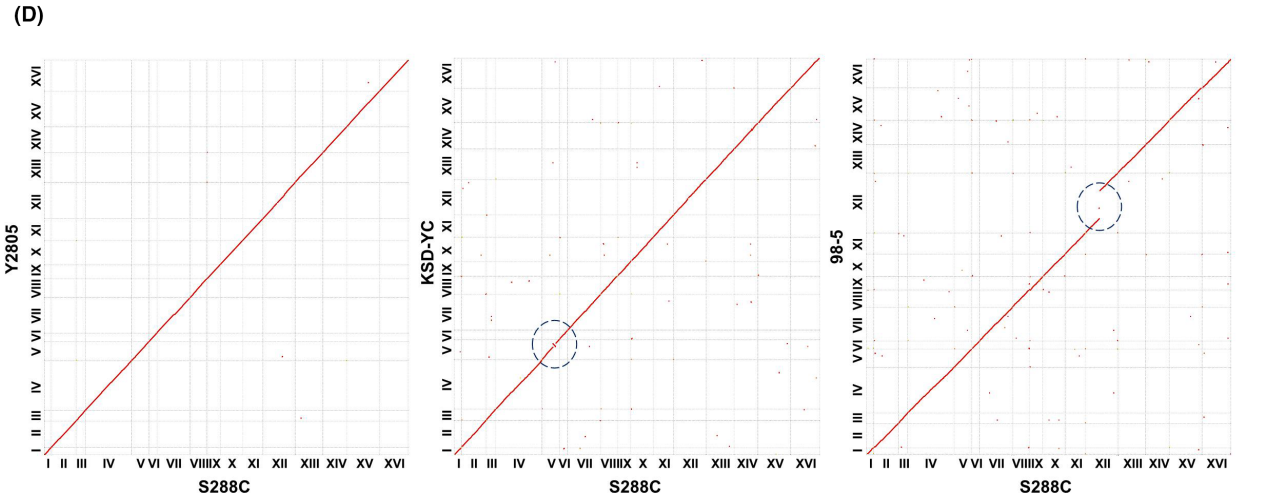


FIGURE 1 Comparative single nucleotide polymorphism (SNP) and insertion or deletion (INDEL) detection in the whole genomes of *S. cerevisiae* KSD-YC, 98-5 and Y2805 strains. Each structural SNP and INDEL chromosome comparison was visualized in (A) *S. cerevisiae* Y2805 and S288C. (B) *S. cerevisiae* KSD-YC and the sake yeast K7. (C) Each haplotype of *S. cerevisiae* 98-5. Read mapping coverages, along with the numbers of SNPs and INDELs, were noted as a table. (D) Synteny analysis of the de novo assemblies of *S. cerevisiae* KSD-YC, 98-5 and Y2805 genomes with the reference S288C genome.

chromosome V of KSD-YC and a large gap in chromosome XII of 98-5.

Phylogenomic analysis of *S. cerevisiae* 98-5, KSD-YC and Y2805 strains

In the phylogenomic analysis including the 66 total *S. cerevisiae* strains with different locations (Europe, Australia, Africa, Middle East, Asia, Malaysia and North America) and source backgrounds (baking, wine, sake/ragi, huangjiu, bio-EtOH, clinical), the Y2805 strain was closely located to the reference strain S288C (Figure 2) as expected in the WG SNP analysis. The KSD-YC strain was evolutionarily positioned very close to the Japanese sake strain, K7, in the phylogenetic tree analysis and thus belongs to the group of phylogenetic niches representing the sake/ragi-derived strains. The large inversion in chromosome V, which was detected in the KSD-YC genome in Figure 1D, was also reported in the sake strain

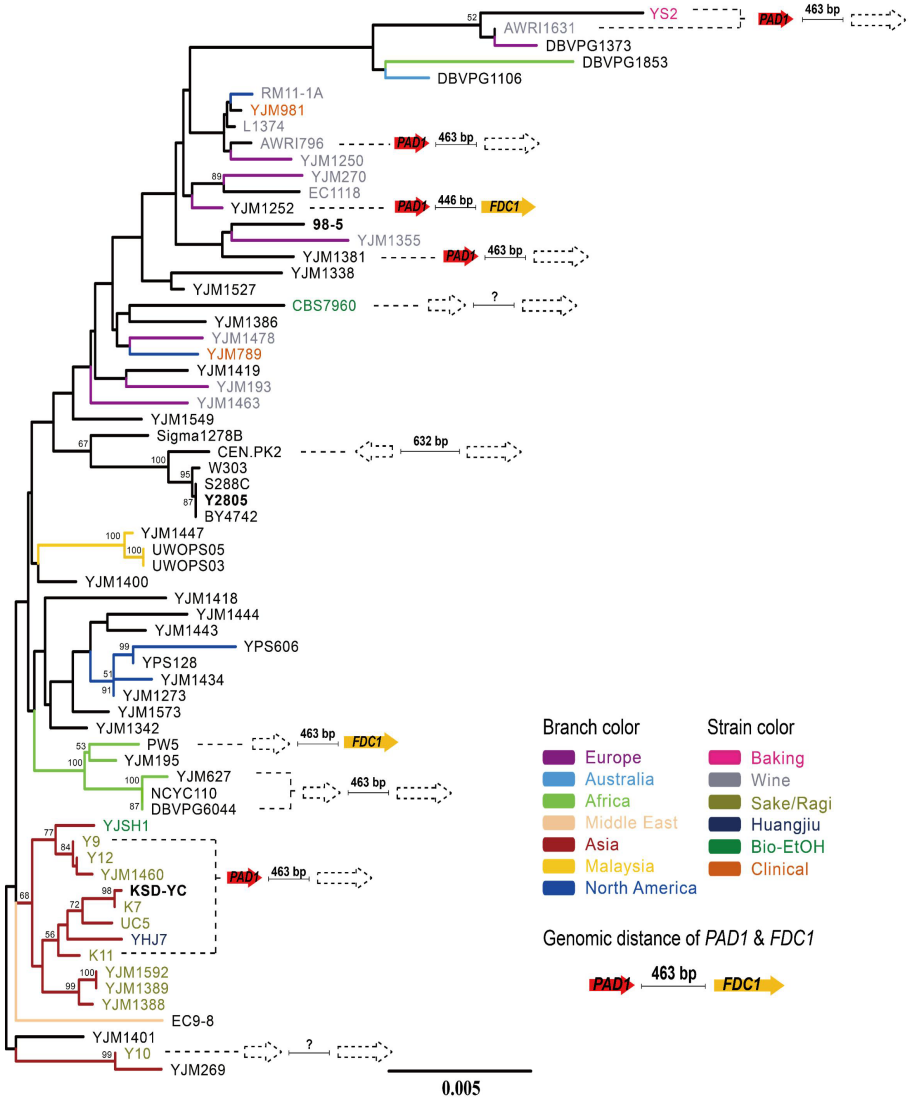


FIGURE 2 Phylogenetic tree construction with maximum likelihood. The 13 concatenated genes, including YPR152C, YJL099W, YJL057C, YJL051W, YKL068W, YML080W, YML056C, YNL161W, YNL125C, YOR133W, YAR042W, YBL052C and YBR163W, were used in the phylogenetic tree construction to classify *S. cerevisiae* at the strain level. The bootstrap values greater than 50% are shown at the branches. The groups of yeast strains are presented as a single model (JTT + F + I) based on the tree topology. The seven region types (Europe, Australia, Africa, Middle East, Asia, Malaysia and North America) and six source types (baking, wine, sake/ragi, huangjiu, bio-EtOH and clinical) are represented as different coloured branches and strains, respectively. The structural features with the genomic distance between *PAD1* and *FDC1* are indicated on the right side. Non-functional genes are represented by dotted arrows.

K7 (Akao et al., 2011), indicating the Asian wine strains KSD-YC and K7 share the same parent, which evolved independently from the ancient *S. cerevisiae* parent strain of S288C. In contrast, the 98-5 strain was closely grouped with JM1355, a European wine strain (Figure 2), indicating that the origin of 98-5 is distinctive from that of KSD-YC, despite both strains being currently used for industrial brewing of Korean traditional rice wines.

Distinctive genome features of *S. cerevisiae* 98-5, KSD-YC and Y2805 strains

Multiple sequence alignment at the chromosomal level featured detailed differences in the ribosomal (rDNA) clusters in chromosome XII (Figure 3A). The genomic organization of 98-5 was shown to be most similar to that of the YJM1381 strain used for rum fermentation in Trinidad, Cuba (Gallone et al., 2016). Although the size of the assembled chromosome XII in S288C and KSD-YC is about 1 Mb, the size of chromosome XII of 98-5 and YJM1381 is 1.9 Mb. This chromosomal size difference resulted from the different copy numbers of the rDNA genes, which are highly likely generated during the assembly process due to the highly repeated features of rDNA clusters (considering that a contig extension with short sequencing reads can result in inaccurate repeat counts) (Jeon et al., 2023). However, it is noticeable that different gene compositions were detected near rDNA clusters even between the closely

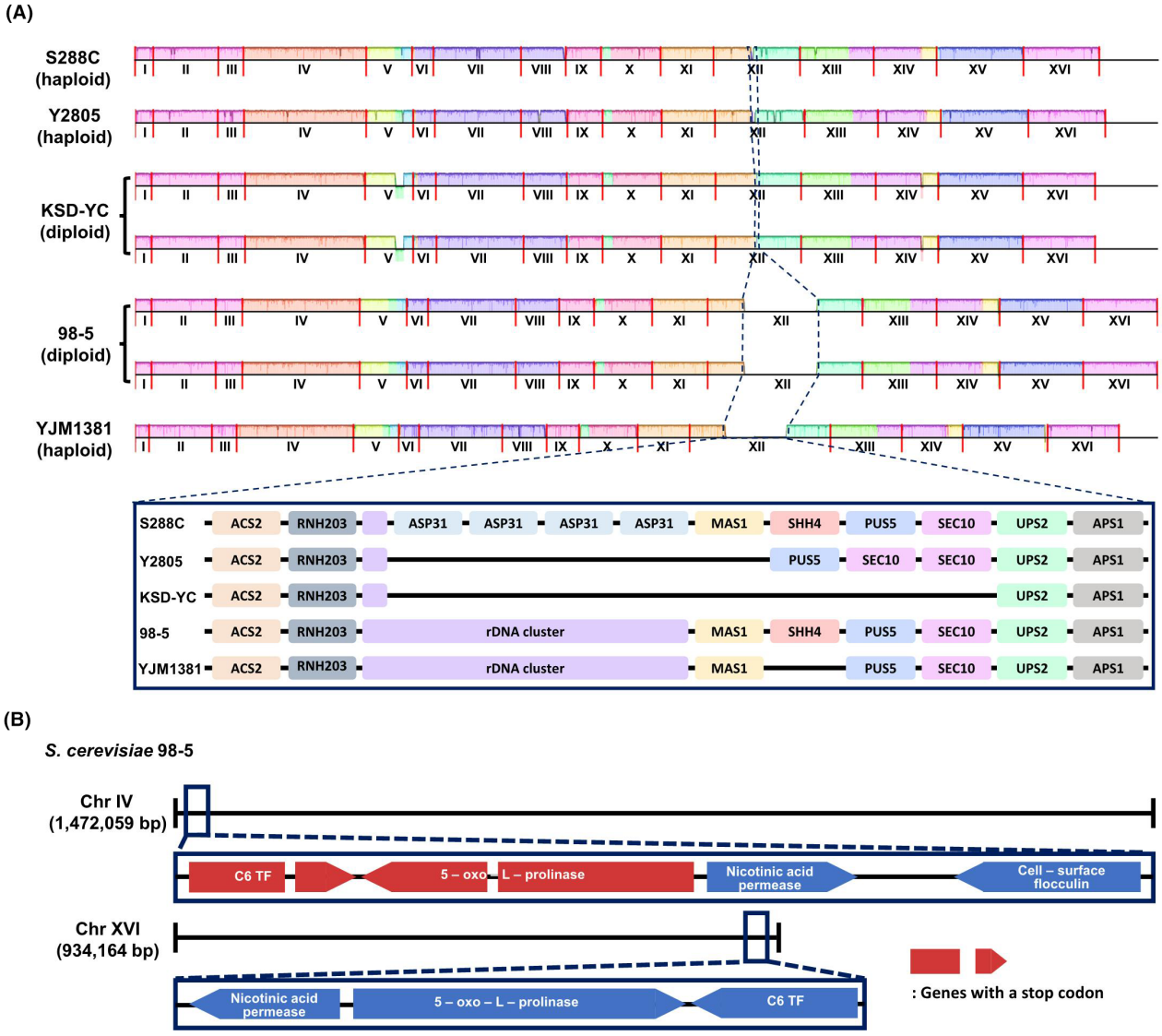


FIGURE 3 Chromosome structure analysis of *S. cerevisiae* KSD-YC, 98-5 and Y2805 strains. (A) Multiple comparative analysis of the whole genomes of the *S. cerevisiae* strains. The synteny blocks of chromosome XII of the *S. cerevisiae* strains were extended to compare the gene components around the rDNA cluster region. The space marked with a dotted line describes the absolute length or the interval of the gene, with the relative position of each gene block. (B) The location of the cluster of five genes conserved in the *S. cerevisiae* wine strains. Among the four *S. cerevisiae* strains, 98-5 only showed the partial cluster of those genes in chromosomes IV and XVI. The genes with a premature stop codon and intact genes are displayed in red and blue, respectively.

related S288C and Y2805 strains. *ASP3* and its paralogues, which encode the cell wall L-asparaginase II, were found only in the S288C genome. The genomes of Y2805 and KSD-YC lack either a few or all four genes (*MAS1*, *SHH4*, *PUS5* and *SEC10*) positioned adjacent to the rDNA locus, which is different from the genomes of the S288C and 98-5 strains. The genes *MAS1* (encoding the mitochondrial-processing peptidase subunit beta) and *SHH4* (encoding a putative alternate subunit of a mitochondrial succinate dehydrogenase) were absent in the Y2805 genome. Additionally, the duplicated *SEC10* genes, encoding a subunit of the exocyst complex, were present following the *PUS5* gene, implying a probable increase in exocytosis activity in the Y2805 strain. Notably, the four genes (*MAS1*, *SHH4*, *PUS5* and *SEC10*) around the rDNA were not detected in the KSD-YC genome. The mutation of *MSA1* was previously reported to cause the increased heat sensitivity (Yaffe & Schatz, 1984), and *SHH4* was identified as one of genes involved in survival of heat shock (Jarolim et al., 2013). The KSD-YC and Y2805 strains, lacking both *MSA1* and *SHH4*, cannot survive at 40°C, whereas the S288C and 98-5 strains can grow robustly at 40°C (Figure S2), reflecting the thermotolerance-associated function of *MSA1* and *SHH4*. Considering the previous report that the deletion of *PUS5*, encoding a pseudouridine synthase for pseudouridine formation of 21S mitochondrial ribosomal RNA, did not generate any defective growth phenotype at various temperatures and media

conditions (Ansmant et al., 2000), it is speculated that the absence of *PUS5* in the genome might not apparently affect the cell growth of KSD-YC.

The characteristic cluster of five genes, including two potential transcription factors (one zinc cluster and one C6 type), a cell surface flocculin, a nicotinic acid permease and a 5-oxo-L-prolinase, have been reported in the genomes of all *S. cerevisiae* wine strains (Borneman et al., 2011) and the probiotic strain, *S. boulardii* (Khatri et al., 2017). The five-gene cluster is thought to have been horizontally acquired by *S. cerevisiae* from *Zygosaccharomyces* spp. and was further subjected to duplication (Novo et al., 2009). Although the individual genes within this cluster are highly conserved between strains, the cluster itself shows high diversity with respect to copy number, genomic location and overall gene order, possibly via the resolution of a circular DNA intermediate. Intriguingly, this cluster was only observed in the 98-5 genome, which displays the incompletely duplicated clusters localized separately on two chromosomes, IV and XVI (Figure 3B). This particular cluster was not present in the genomes of the KSD-YC, S288C and Y2805 strains. As indicated in our phylogenetic tree analysis data (Figure 2), several aspects of the 98-5 strain genome structure strongly support a closer relationship with the European wine strains rather than with the Asian wine strains.

Genomic structures of *FDC1* and *PAD1*, associated with 4-VG production, in *S. cerevisiae* industrial strains

Production of 4-VG, a phenolic compound with a smoke-like flavour, is made by yeast via the decarboxylation of ferulic acid, an abundant phenolic compound found in many plant cell walls, by the *FDC1*-encoded decarboxylase. This decarboxylase requires a flavin-derived cofactor encoded by *PAD1*. The biological role of Pad1p and Fdc1p is to detoxify phenylacrylic acids (Mukai et al., 2014), contributing to survival and proliferation in natural habitats. However, 4-VG is an undesirable trait for the production of most beers and thus many industrial yeasts for beer brewing acquired loss-of-function mutations in *PAD1* and/or *FDC1*, resulting in a loss of the ability to produce 4-VG (Gallone et al., 2018). Among the 66 total *S. cerevisiae* strains analysed for the phylogenetic tree (Figure 2), we observed different genomic structures in *FDC1* and *PAD1* due to three aspects: the mutations in each gene, distance between two genes and direction of each gene. In most cases, the genetic distance between *FDC1* and *PAD1* was 463 bp, but the distance was 632 bp in the genome of CEN.PK2, one of the representative laboratory strains. The mutations detected in *FDC1* and *PAD1* were mostly early stop codons with a few cases of deletions and amino acid substitutions (Table S6).

Comparison of the *PAD1* and *FDC1* gene sequences, which are clustered in the subtelomeric region of the right arm of chromosome IV in the *S. cerevisiae* strains, revealed notable SNPs and deletion mutations in the S288C, CEN.PK2-1C, KSD-YC, 98-5 and Y2805 strains (Figure 4A; Figure S3A). While the S288C, Y2805 and 98-5 strains retain the wild-type *PAD1* and *FDC1* genes, a premature stop codon mutation was detected in the *FDC1* gene of KSD-YC, indicating its 4-VG production ability loss. As previously reported, *S. cerevisiae* CEN.PK2-1C has a mutation in *PAD1*, resulting in a stop codon instead of tyrosine at the 98th amino acid position (Richard et al., 2015). Intriguingly, we further detected additional mutations in CEN.PK2-1C, including the extended space length between the *PAD1* and *FDC1* genes, the inverted orientation of *PAD1*, and the deletion of *FDC1* corresponding to the 267–198th amino acids. By analysing the bioconversion activity of ferulic acid to 4-VG in the *S. cerevisiae* strains (Figure 4B; Figure S3B), the loss of 4-VG production activity was confirmed in the CEN.PK2-1C and KSD-YC strains, while the 4-VG production phenotype (4-VG^{+}) was observed in the S288C, Y2805 and 98-5 strains. The results support the positive relationship between the sequence variation of *PAD1* and/or *FDC1* and the ferulic acid decarboxylation ability of industrial yeast strains.

Phenotypic profiles of *S. cerevisiae* industrial strains analysed by phenotype microarray

To identify the phenotypic profile of the industrial *S. cerevisiae* strains under different nutrient conditions, the growth of each strain was monitored during a 7-day cultivation in PM microplates with carbon sources (PM1 and PM2), nitrogen sources (PM3), phosphorus and sulphur sources (PM4), nutrient supplements (PM5) and osmolytes (PM9). Notably, the rice wine yeast KSD-YC and 98-5 strains grew better in various carbon (C) sources, including maltose, sucrose, maltotriose and turanose, compared to the S288C and Y2805 strains (Figure 5A). Particularly, 98-5 utilized galactose most efficiently and Y2805 exhibited slightly delayed growth, while KSD-YC and S288C exhibited Gal^{-} phenotypes. Only KSD-YC showed moderate growth when utilizing α-methyl-D-glucoside as a sole C-source. For palatinose utilization, the *S. cerevisiae* strains were distinguished by their growth at different degrees, with the least growth seen with strain 98-5.

Unlike other strains, nitrogen (N) sources, such as histidine and pyroglutamic acid, were preferable for the growth of 98-5, and only S288C was able to grow using asparagine as the sole N-source (Figure 5B). In the plates containing various phosphorus and sulphur sources, the Y2805 strain showed a defect in utilizing

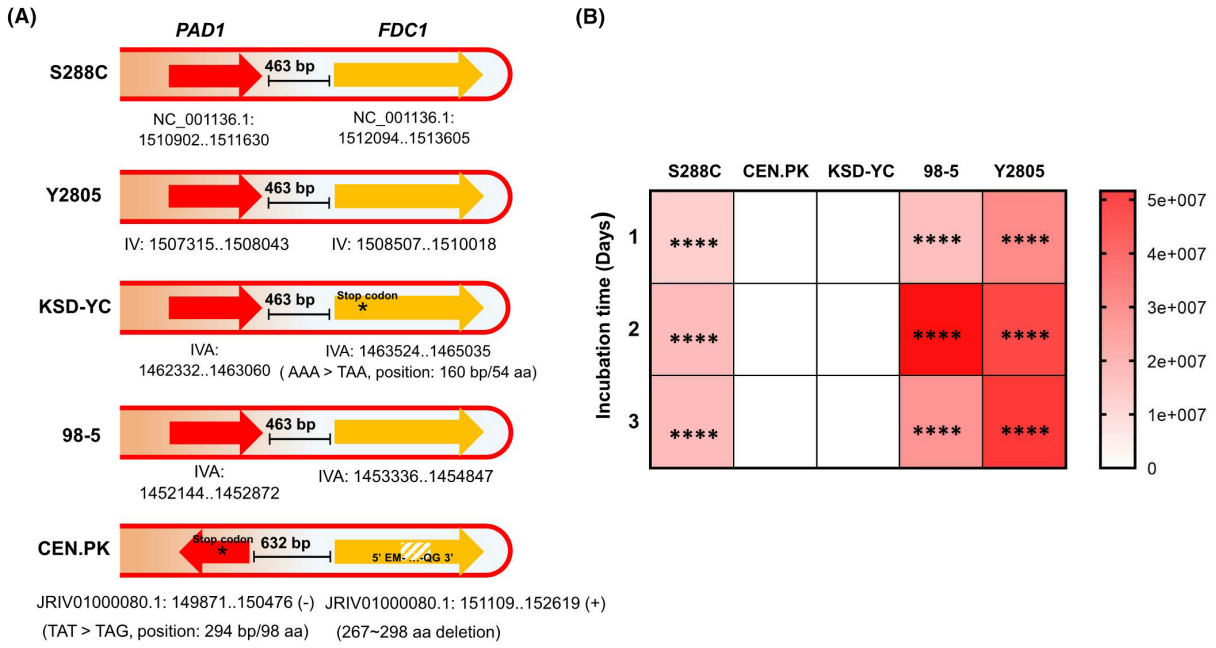


FIGURE 4 Comparative analysis of the 4-vinylguaiacol (4-VG) bioconversion activity of *S. cerevisiae* strains. (A) Schematic representation of the *PAD1* and *FDC1* genes required for 4-VG bioconversion in *S. cerevisiae* S288C, CEN.PK2-1C, KSD-YC, 98-5 and Y2805 strains. The red lines, represents chromosomes, indicating the subteolomere location of the *PAD1* and *FDC1* genes on the right arm of chromosome IV. The information on the accession numbers of the *PAD1* and *FDC1* genes with the detected SNP/deletion is provided in Table S6. (B) Heatmap of 4-VG production in *S. cerevisiae* strains through headspace-solid-phase microextraction with gas chromatography/mass spectrometry (HS-SPME GS/MS). To test 4-VG production capability, yeast cells were grown in YPD medium (1% yeast extract, 2% bacto peptone and 2% glucose) in the presence of 50 ppm ferulic acid, and samples were collected after 1, 2 and 3 days of incubation.

methylene diphosphonic acid and dithiophosphate and also displayed retarded growth in inositol hexaphosphate compared to the other strains (Figure 5C). In the nutrient supplement plate, only the KSD-YC strain grew well with pantothenic acid supplementation, while S288C grew faster than the other strains with biotin supplementation (Figure 5D). In addition, under various osmotic stress conditions (generated by supplementation with different concentrations of NaCl, urea and sodium nitrite), KSD-YC and 98-5 generally showed more sensitivity compared to S288C (Figure 5E). While Y2805 showed relatively strong resistance to the osmotic shock caused by high concentrations of NaCl, this *S. cerevisiae* strain did not grow at higher concentrations of urea and sodium nitrite.

Integration of PM data with SNPs in the carbon metabolism pathway

To investigate the genetic differences associated with the diverse growth patterns observed for each *S. cerevisiae* strain in the PM analysis, the genes involved in the metabolism of each substrate source were identified and their amino acid sequences were aligned and compared among the four *S. cerevisiae* strains (Table S7 and Figure S4). Interestingly, the different growth patterns were mapped with mutations in key genes responsible for the corresponding nutrient metabolism (Table 2). The alignment of these *GAL* genes strongly indicated that the inability of KSD-YC to utilize galactose as a C-source is evidently determined by mutations in *GAL3* and *GAL4*, which contain many SNPs, deletions and truncations (Figure S4A). Compared to 98-5, which showed the most active galactose metabolism, Y2805 showed a few non-synonymous SNPs in the *GAL1*, *GAL2*, *GAL7* and *GAL10* genes, which might explain its reduced galactose metabolism activity. The inability of S288C to metabolize galactose (Gal^-) has been correlated with the significant number of non-synonymous SNPs observed in its *GAL1*, *GAL10* and *GAL2* genes (Otero et al., 2010). However, when compared to the *GAL* genes of the 98-5 strain, S288C showed SNPs only in *GAL2* and *GAL10* (Figure S4A), implying that the *GAL1* gene might be functional in S288C.

In the genes involved in the maltose metabolism pathway (Table S7), *MAL31*, which encodes a maltose permease involved in maltose metabolism, was found to have multiple non-synonymous SNPs between the Mal^+ strains (KSD-YC and 98-5) and the Mal^- strains (S288C and Y2805) (Figure S4B). The *SUC* gene family of *S. cerevisiae*, encoding an invertase that catalyses the hydrolysis of sucrose and inulin, includes six structural genes for invertase (*SUC1*, *SUC2*, *SUC3*, *SUC4*, *SUC5* and *SUC7*) found at unlinked chromosomal loci

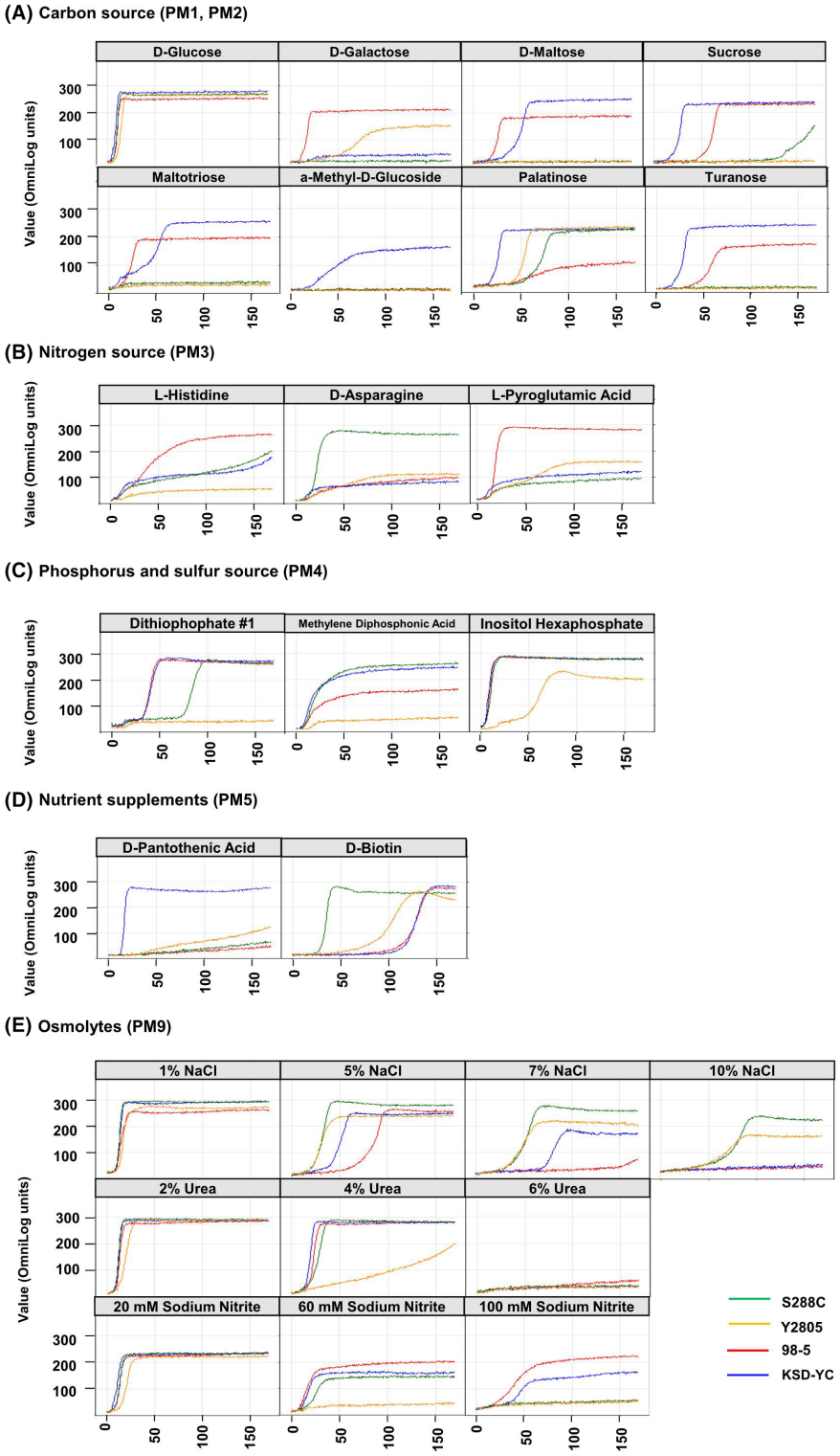


FIGURE 5 Phenotype microarray analysis of *S. cerevisiae* strains. Representative growth patterns of *S. cerevisiae* 98-5 (red), KSD-YC (blue), S288C (green) and Y2805 (yellow) strains on Biolog microplates. (A) Carbon source plate (PM1 and PM2), (B) Nitrogen source (PM3), (C) Phosphorus and sulphur source (PM4), (D) Nutrient supplements (PM5) and (E) Osmolytes (PM9), where the x- and y-axis represent time in hours and Omnilog units, respectively. The Omnilog unit is a standard representation of respiration rate.

(Carlson & Botstein, 1983). We detected only the presence of *SUC2*, without other members of *SUC* family, in the genomes of S288C, CEN.PK2-1C, KSD-YC, 98-5 and Y2805 strains, indicating that Suc2p is the only invertase responsible for sucrose utilization. In the case of Suc2p, two SNPs at the 84th and 88th amino acid positions were detected between the robustly growing KSD-YC and 98-5 strains and the poorly growing S288C and Y2805 strains, suggesting that the amino acid changes from histidine to asparagine (H84N) and from glutamate to glutamine (E88Q) might result in a dramatic decrease in the function of invertase (Figure S4B). The identical amino acid changes at 84th and 88th positions were also previously reported in Suc2 proteins between the strong- and the weak-inulin-degrading strains (Wang & Li, 2013). However, the amino acid sequence changes

TABLE 2 Mapping of single nucleotide polymorphisms (SNPs) to the *S. cerevisiae* genes involved in carbon metabolism based on phenotype microarray (PM) analysis data[a].

Genes	Function	SNP positions	Strains	Amino acids
Galactose				
GAL1	Galactokinase	48, 415	S288C, KSD-YC, 98-5 / Y2805	A/S, S/G
		297	S288C, 98-5 / Y2805, KSD-YC	L/P
GAL2	Galactose permease	50, 392	S288C, Y2805 / KSD-YC, 98-5	S/P, H/R
		369	S288C, KSD-YC, 98-5 / Y2805	S/Y
		463	S288C, Y2805, KSD-YC / 98-5	P/R
GAL3	Transcriptional regulator	70, 202, 246	S288C, Y2805, 98-5/ KSD-YC	D/G, G/E, S/deletion
		352	S288C, Y2805, KSD-YC / 98-5	H/D
GAL4	DNA-binding transcription factor	154, 370, 451, 508, 573, 607, 630, 639, 640, 647, 788, 798, 815	S288C, Y2805, 98-5 / KSD-YC	Multi-SNPs
GAL7	Galactose-1-phosphate uridyl-transferase	211	S288C, Y2805 / KSD-YC, 98-5	T/I
		258	S288C, Y2805, KSD-YC / 98-5	A/T
		267, 345	S288C, KSD-YC, 98-5 / Y2805	V/A, T/I
GAL10	UDP-glucose-4-epimerase	263	S288C, Y2805, 98-5 / KSD-YC	Q/K
		359, 621	S288C, KSD-YC, 98-5 / Y2805	G/D, T/A
		518	S288C / KSD-YC, 98-5, Y2805	M/I
GAL80	Transcriptional regulator	92	S288C, Y2805, KSD-YC / 98-5	I/M
		101	S288C, 98-5 / Y2805, KSD-YC	E/D
Maltose				
MAL31	Maltose permease	31, 58, 122, 265, 268, 415, 506, 609	S288C, Y2805, 98-5 / KSD-YC	Multi-SNPs
		607	S288C, Y2805 / KSD-YC, 98-5	N/S
Sucrose				
SUC2	Sucrose hydrolysing enzyme	14, 138, 409	S288C, Y2805, KSD-YC / 98-5	A/T, S/I, A/P
		84,88	S288C, Y2805 / KSD-YC, 98-5	N/H, Q/E
α-Methyl-D-Glucoside				
IMA4	Alpha-glucosidase	54	S288C, Y2805 / KSD-YC, 98-5	T/A
		240, 279	S288c, Y2805, 98-5/ KSD-YC	L/P, R/Q
Turanose				
IMA3	Alpha-glucosidase	54	S288C, Y2805 / KSD-YC, 98-5	T/A
		183	S288C, Y2805, 98-5/ KSD-YC	P/L
Palatinose				
IMA5	Alpha-glucosidase	123, 175, 254, 255, 261, 275, 279, 283, 306, 308, 364, 405, 433, 447, 448, 450, 455, 549, 550, 562, 566, 579, 580	S288C, Y2805, KSD-YC / 98-5	Multi-SNPs
		182, 477	S288C, Y2805, 98-5/ KSD-YC	W/deletion, A/S
		480, 483	S288C, Y2805 / KSD-YC, 98-5	K/N, N/D

[a]The table contains information on the position and altered amino acids of SNPs identified in the genes present in all the four *S. cerevisiae* strains, S288C, KSD-YC, 98-5 and Y2805, using S288C as reference.

in Suc2p were proven not to be the main reason for the discrepancy in enzyme activity. The subsequent study showed that the sequence variation in *SUC2* promoters affected the expression level of *SUC2* in *S. cerevisiae* strains, leading to different enzyme activity (Yang et al., 2015). It is notable that the previously reported sequence variations in *SUC2* promoter sequences, including the change in transcription activator Msn2p/Msn4p-, repressor Sko1p-, activator Gcr1p-, repressor Mig1p- and RNA-Pol II-binding sites, were also observed between the slow-growing strains (S288C, Y2805) and the most vigorously growing strain KSD-YC analysed in the present study (Figure S5), indicating that the different *SUC2* expression at the transcription level would generate the different metabolic activity. Regarding the 98-5 strain, having additional SNPs at 14th, 138th, 409th positions in ORF (Table 2) and unique variations in the promoter sequence of *SUC2* (Figure S5), we might analyse the mRNA level and enzyme activity to identify what causes the different growth rate on sucrose.

The metabolism of the sucrose isomers, palatinose and turanose, require the *MAL* genes encoding maltases and the closely related *IMA* gene family encoding isomaltases (Table S7). *IMA3*, encoding an α-glucosidase specific for α-1,3 linkage turanose, was found to have one SNP at 183th amino acid position (P183L) that is specific to KSD-YC (Figure S4B). In the case of *IMA4*, an α-glucosidase with a broad substrate specificity for α-1,4- and α-1,6-glucosides, two SNPs at the 240th and 279th amino acids (L240P and R279Q) were detected between the other three strains and the KSD-YC strain (Figure S4B). These KSD-YC-specific SNPs detected in *IMA3* and *IMA4* might account for the highest growth of this strain when using α-methylglucoside and turanose as the sole C-sources. Particularly, *IMA5*, encoding an α-glucosidase with a specificity for isomaltose, maltose and palatinose, showed many SNPs specific to the 98-5 strain with a poor palatinose utilization (Figure S4B). Besides *MAL11* (*AGT1*) and *MAL31*, the genes of the maltose permease homologues, such as *MPH2* and *MPH3*, encode α-glucoside permease, which is involved in transporting maltose, maltotriose, α-methylglucoside and turanose in some beer yeast strains (Vidgren et al., 2005). However, the KSD-YC and 98-5 genomes do not contain *MPH2* and *MPH3* (Table S7), implying that the presence of functional Mal31p is sufficient for the transport of maltose, maltotriose, α-methylglucoside and turanose.

Integration of the PM data with SNPs with other metabolisms and osmotic stress

It is also noticeable that the 98-5 strain, which can use histidine as an N-source more efficiently than the other strains, has a strain-specific SNP in the *HIP1* gene coding for a high-affinity histidine permease (Table 3; Figure S4C). Regarding the growth on asparagine, all four *S. cerevisiae* strains contain *ASP1*, encoding a cytosolic L-asparaginase, whereas only the S288C strain contains at least four copies of *ASP3*, *ASP3-1*, *ASP3-2, ASP3-3 and ASP3-4*, which encode the cell wall L-asparaginase II involved in asparagine catabolism (Kim et al., 1988), adjacent to the rDNA repeats

TABLE 3 Mapping of single nucleotide polymorphisms (SNPs) to the *S. cerevisiae* genes associated with nitrogen and nutrient supplements based on phenotype microarray (PM) analysis data[a].

Gene	Function	Position	Strains	Amino acids
L-histidine				
HIP1	High-affinity histidine permease	560, 595	S288C, Y2805, KSD-YC / 98-5	K/N, V/I
L-Pyroglutamic acid				
OXP1	5-Oxoprolinase	296, 436, 455, 734, 1107	S288C, Y2805, KSD-YC / 98-5	Multi-SNPs
D-Pantothenic acid				
CAB1	Pantothenate kinase	2	S288C, Y2805, 98-5/ KSD-YC	P/S
		111	S288C, Y2805, KSD-YC / 98-5	H/Q
		246, 366	S288C, Y2805 / KSD-YC, 98-5	M/I, S/N
Biotin				
VHT1	High-affinity plasma membrane H+-biotin symporter	24	S288C, Y2805 / KSD-YC, 98-5	Y/S
		119	S288C, Y2805, KSD-YC / 98-5	R/G
		306, 579	S288C, Y2805, 98-5/ KSD-YC	I/L, N/D
NaCl				
ENA1	P-type ATPase sodium pump	38, 83, 101, 102, 106, 129, 191, 214, 244, 249, 252, 266, 308, 312, 313, 348, 402, 407, 409, 414, 496, 505, 511, 513, 517, 518, 524, 525, 529, 557, 579, 618, 619, 620, 622, 627, 633, 752, 843, 882, 902, 903, 905, 921, 923, 926, 927, 930, 931, 934, 937, 984, 1024, 1045, 1085	S288C, Y2805 / KSD-YC	Multi-SNPs
ENA2	P-type ATPase sodium pump	753	S288C, Y2805_1, Y2805_2 / Y2805_3	A/T
ENA5	Protein with similarity to P-type ATPase sodium pumps	2, 6, 12, 22, 42, 43, 44, 48, 214, 444, 497, 556, 1038	S288C, Y2805 / 98-5	Multi-SNPs

[a]The table contains information on the position and altered amino acids of SNPs identified in the genes present in all the four *S. cerevisiae* strains, S288C, KSD-YC, 98-5 and Y2805, using S288C as reference.

in chromosome XII (Figure 3A). The presence of four copies of *ASP3* only in S288C might attribute to its unique ability to grow rigorously using asparagine as the N-source. The 98-5 strain had four strain-specific SNPs in Oxp1p, an ATP-dependent 5-oxoprolinase, compared to those of the other three *S. cerevisiae* strains (Table 3; Figure S4C), thus explaining why 98-5 could utilize pyroglutamic acid as the N-source most efficiently.

For nutrient supplementation, the presence of pantothenate (also called vitamin B5) supported the robust growth of the KSD-YC strain only. It was revealed that the KSD-YC strain has a strain-specific SNP at the second amino acid residue (P2S) in the pantothenate kinase Cab1p, which is responsible for the catalysis of the first step in the metabolism of pantothenic acid for CoA biosynthesis in the budding yeast *S. cerevisiae* (Table 3; Figure S4D). It can be speculated that the amino acid change from serine to proline detected in the other *S. cerevisiae* strains might generate a non-functional Cab1p. Intriguingly, the presence of biotin apparently stimulated the growth of the biotin auxotrophic S288C strain, which lacks *BIO1* and *BIO6* required for biotin biosynthesis (Wronska et al., 2020). Our WG sequencing data revealed that the other three *S. cerevisiae* strains also lacked *BIO1* and *BIO6* (Table S7). However, the growth of the KSD-YC and 98-5 strains was not recovered by biotin supplementation, indicating that they might have defects in biotin uptake. As expected, the alignment of the *VHT1* genes, coding for a high-affinity plasma membrane H^+-biotin (vitamin H) symporter, shows an amino acid change at the 24th position, commonly observed in the KSD-YC and 98-5 strains (Table 3; Figure S4D).

Regarding osmolyte resistance, the S288C strain exhibited the strongest resistance to high NaCl concentrations, and the Y2805 strain showed a stronger resistance compared to the KSD-YC and 98-5 rice wine strains. It was reported that the strong NaCl tolerance of S288C is due to the tandemly triplicated *ENA1*/*ENA2*/*ENA5*, encoding P-type ATPases located on the *PMR2* locus of chromosome IV (Wieland et al., 1995). This locus has been described as highly variable among *S. cerevisiae* strains and can contain one to five highly conserved copies of the *ENA* genes. Our genome analysis data revealed that Y2805 has one copy of *ENA1* and three copies of *ENA2*, but lacking *ENA5*, on the same chromosome. In contrast, the KSD-YC and 98-5 strains have only one copy of either *ENA1* or *ENA5*, respectively (Table S7). The *ENA1* of the KSD-YC strain and *ENA5* of the 98-5 strain showed multiple SNPs compared to the corresponding genes of the S288C and Y2805 (Table 3; Figure S4E), strongly indicating that the differences in the amino acid sequence and the copy number of *ENA* genes might account for the low resistance of the KSD-YC and 98-5 strains to NaCl (Figure S2).

Structural features of the auxotrophic marker genes of Y2805 validated by Sanger sequencing

The *S. cerevisiae* Y2805 strain has been widely used as a host strain to produce recombinant proteins with the potential for medical and industrial applications (Table S1). This *S. cerevisiae* strain has three auxotrophic markers (*can1, his3-200 and ura3-52*), which are useful for transformant selection during genetic manipulation, and two mutations in its vacuolar proteases (*pep:HIS3* and *prb-Δ1.6R*), which are beneficial for blocking protein degradation. To obtain detailed information on the genotype markers of Y2805, the DNA sequences of each marker gene in Y2805 genome were initially generated by Illumina sequencing data (more than 200 coverages, Figure 1). For verification of the sequence information from massive NGS data, the DNA fragments containing the marker genes were amplified by PCR and subjected to Sanger sequencing analysis.

It was revealed that the *PEP4* gene, encoding vacuolar protease A, is disrupted by the insertion of the 1771 bp *HIS3* in a reverse direction at the 411 bp position of the *PEP4* open reading frame (ORF), causing the *pep4::HIS3* mutation in *S. cerevisiae* Y2805 (Figure 6A). The inserted *HIS3* gene fragment contains its native 469 bp promoter and 639 bp terminator, respectively. The *PRB1* gene, encoding vacuolar protease B, has a partial deletion from the 251 bp to 1840 bp position of the ORF, generating *prb-Δ1.6R* (Figure 6B). When compared to *S. cerevisiae* S288C, which is known as the *GAL2* auxotroph, the *GAL2* gene, coding for a galactose permease, shows one SNP at the 369th amino acid position in Y2805, compared to that of S288C. The amino acid at the 369th position of Gal2p is serine (TCC) in S288C, while that of Gal2p is tyrosine (TAC) in Y2805 (Figure 6C). The *CAN1* gene, encoding an arginine amino acid transporter localized to the plasma membrane, has a frameshift mutation in which there is a deletion of the single 'C' base at the 1,002 bp position in the coding region in Y2805 (Figure 6D). The *HIS3* gene, encoding imidazoleglycerol-phosphate dehydratase required for histidine biosynthesis, was completely lost in Y2805 by deleting a 1,039 bp DNA fragment including the promoter and terminator of *HIS3*, thus generating the *his3-Δ200* mutation (Figure 6E). Since the *HIS3* gene shares a common promoter with *MRM1* (previously named *PET56*), encoding a mitochondrial rRNA methyltransferase, the *his3-Δ200* mutation was reported to show decreased *MRM1* expression (Zhong et al., 2004). Notably, the *URA3* gene, coding for an orotidine 5-phosphate decarboxylase involved in the biosynthesis of uracil, is mutated by the integration of a Ty element at the 119th bp position of the *URA3* ORF, as previously reported in the *ura3-52* mutation (Rose & Winston, 1984). The integrated Ty element is 5,919 bp in length and consists of a

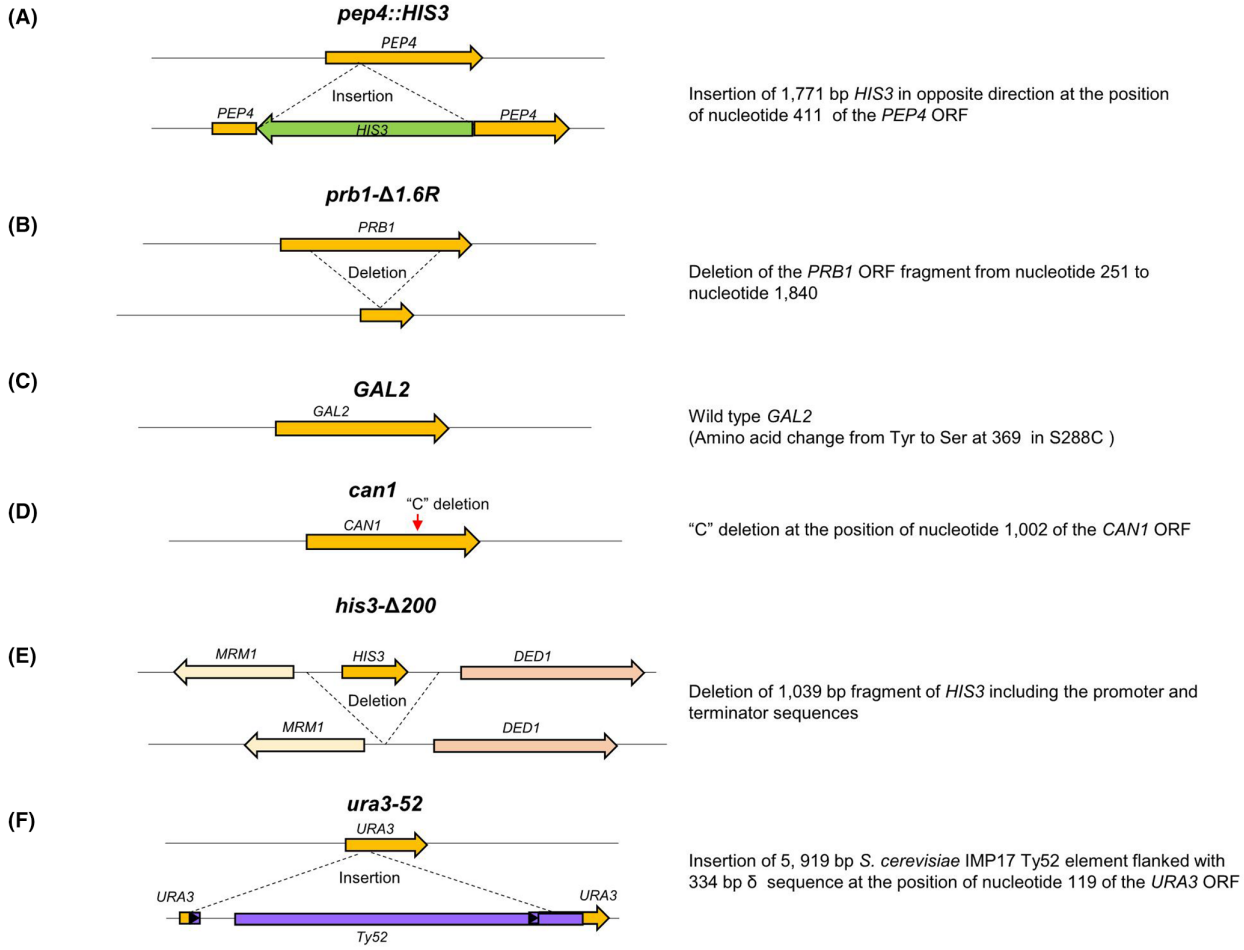


FIGURE 6 Schematic representation of structural and sequence features of *S. cerevisiae* Y2805 genetic markers validated by Sanger sequencing. (A) *pep4::HIS3*, (B) *prb1-Δ1.6R*, (C) *GAL2*, (D) *can1*, (E) *his3-Δ200*, (F) *ura3-52*. The DNA fragments of the genotype marker genes were amplified by PCR from the total chromosomal DNAs, which were prepared by lysing the yeast cells with glass beads, using the gene-specific primers (Table S2). The PCR products were directly subjected to sequencing or subcloned into a T vector (T-Blunt™ PCR Cloning kit; SolGent, Daejeon, South Korea) before DNA sequencing by the dye-terminator sequencing method.

5,250 bp region flanked by a 334 bp perfect direct repeat, called delta in Y2805 (Figure 6F).

When we compared the growth of the *S. cerevisiae* S288C, 98-5, KSD-YC and Y2805 strains under different salt- /osmotic conditions and at different culture temperatures by spotting analysis, they exhibited distinctive physiological characteristics (Figure S2). Although the 98-5 and KSD-YC strains display low NaCl tolerance, as indicated by PM analysis (Figure 5), they show a relatively high tolerance to sorbitol, indicating that high tolerance to osmotic stress is the requisite physiological features of starter strains for rice wine fermentation. It is notable that Y2805 showed the least tolerance to both KCl and sorbitol, along with the least resistance to both cold and high temperatures. This high temperature sensitivity phenotype of Y2805 might be associated with the auxotrophic marker *his3-Δ200* mutation, which results in the loss of respiratory competence at 37°C due to the defective expression of *MRM1* (Young & Court, 2008). The absence of other mitochondrial associated genes, *MSA1* and *SHH4*, in the genome of Y2805 strain (Figure 3A) also indicated the decreased thermotolerance of this strain. Along with WG sequence information to provide holistic genetic features, detailed structural organization and sequence information of the Y2805 selection marker genes would deepen our understanding of host cell physiology, which should be practically considered in designing large-scale cultivation processes for the industrial production of recombinant proteins and metabolites.

DISCUSSION

Due to enormous progress in NGS techniques, comparative genomics has become a powerful instrument to study the origin, diversity, population structure and natural history of industrial microorganisms. In the present study, we report the genomic features and physiological characteristics of three *S. cerevisiae* industrial strains, 98-5, KSD-YC and Y2805, by performing a complete WG sequencing and chromosome-level assembly.

Comparative analysis at the whole-genome level reveals the structural variants, including SNPs, INDELs and copy number (CN) variation, in the three industrial *S. cerevisiae* strains (Table S8). The obtained genome information serves as a good basis for comparative genomics to elucidate the evolutionary consequences of *S. cerevisiae* strains in fermentation environments. Combined with PM analysis, the integrated genomic-phenotype data allows the identification of key genetic variations responsible for the phenotype diversity at the genome level, providing insight into understanding the specific genotype–phenotype relationship in three *S. cerevisiae* strains.

Traditional Korean rice wine is made by fermenting nuruk, the starch materials of rice containing diverse airborne microorganisms, including bacteria, fungi and yeast (Song, 2013). These microorganisms provide several hydrolytic enzyme sources required for starch degradation during saccharification to produce glucose and other organic acids. Glucose is mainly fermented by *S. cerevisiae* to produce alcohol and volatile components important for wine aroma. It is notable that the genomes of the 98-5 and KSD-YC strains, which are currently used for brewing commercial rice wine in Korea by several local brewery companies (Kim et al., 2014; Shin et al., 2019), were revealed to contain heterozygous diploid genomes, whereas that of Y2805, which is mostly used as a host strain for the production of recombinant proteins (Table S1), was haploid (Figure 1). A heterothallic diploid could be generated by the out-crossing of two different haploid strains and the subsequent loss of heterozygosity (LOH), thus resulting in the observed pattern. Interestingly, the 98-5 strain showed a significantly high heterozygosity with uneven distribution (Figure 1C), suggesting that sequential LOH events have resulted in the uneven heterozygosity distribution. Conversely, it is reasonable to presume that isolated heterozygosities were introduced by point mutations independent of LOH events. Besides the *S. cerevisiae*-related species, our previous work on the comparative genomics of yeast species isolated from Korean traditional food indicated that the interspecies hybridization within the yeast species occurs frequently as an evolutional strategy in the fermentation environment (Choo et al., 2016; Jeong et al., 2022).

The diploid KSD-YC genome showed two nearly identical genome copies, which are evolutionarily very close to the Japanese sake strain K7 in the phylogenetic tree analysis (Figure 2). When compared to the genome of the sake yeast K7, several noticeable differences are observed between it and that of KSD-YC. Several K7 genes have been demonstrated to be involved in the characteristic features of sake yeast, including *AWA1*, encoding a glycosylphosphatidylinositol anchor protein that is required for cell surface hydrophobicity (Shimizu et al., 2005), *BIO1* and a paralogous set of *BIO6* genes (*BIO6-1*/*BIO6-2a*/*BIO6-2b*/*BIO6-3*/*BIO6-4a*/*BIO6-4b*), which are required for biotin biosynthesis (Akao et al., 2011). However, KSD-YC does not contain *BIO1* and *BIO6* (Table S7) and has only the partial sequence of *AWA1*. The sequential comparison between *AWA1* of K7 and KSD-YC revealed five INDEL regions even in putative functional domains, suggesting there is an inactivated *AWA1* in KSD-YC (Figure S4F). Whereas the tandemly triplicated *ENA1*/*ENA2*/*ENA5* array is present in KSD-YC, only one copy was identified in K7 (K7_ENA1/K7_01190). Such detectable differences indicate that KSD-YC and K7 have recently diverged from the same linage and their genomes have evolved differentially under distinctive conditions for brewing Korean rice wine and sake, respectively. While a total of 1,347 heterozygous sites between two homologous chromosomes were reported for the K7 genome (Shimizu et al., 2005), a total of 1,827 heterozygous sites were counted in the KSD-YC genome. Considering that LOH decreased heterozygosity, it can be speculated that KSD-YC more recently diverged compared to K7, as indicated in our phylogenetic tree analysis (Figure 2).

The genetic variation is comprised of SNPs and large-scale INDELs, with the latter often being associated with the heterogeneity of ORFs between strains. The selection against 4-VG production is a well-documented domestication trait, favouring the spread of domesticated beer yeasts unable to produce this specific off-flavour. Conversely, it is a selected trait for wheat and Belgian beers, contributing markedly to their characteristic clove-like flavour (Gonçalves et al., 2016). In our SNP analysis, we found that while the 4-VG$^+$ phenotype was observed in the S288C, Y2805 and 98-5 strains, the KSD-YC strain was 4-VG$^-$ due to a premature stop codon mutation in *FDC1* (Figure 4). Moreover, it was revealed that the CEN.PK2-1C strain has not only a nonsense mutation in *PAD1* but also an additional deletion mutation in *FDC1*. To validate the phenotype–genotype correlations, we introduced the functional wild-type genes of *FDC1* and *PAD1* into the CEN.PK2-1C strain, using *CEN* vectors (Table S9). When transformed separately with either a *CEN* vector expressing *FDC1* (YCpUT-ScFDC1-HA) or *PAD1* (YCpTT-ScPAD1-6HIS) only, the 4-VG production activity of the recombinant CEN.PK2-1C strains was not recovered (Figure S3C). In contrast, the co-transformation of CEN.PK2-1C with both vectors, YCpUT-ScFDC1-HA and YCpTT-ScPAD1-6HIS, successfully converted the 4-VG$^-$ phenotype of CEN.PK2-1C to the 4-VG$^+$ phenotype. Furthermore, we also confirmed the recovery of 4-VG production ability of the KSD-YC strain by introducing a *CEN* vector, YCpNT-ScFDC1-HA (Table S9), which expresses the WT functional *FDC1* under the control of *TEF1* promoter and carries a nourseothricin N-acetyl transferase (*NAT*) as a selection marker, into the KSD-YC strain (Figure S3D). The results strongly demonstrate

that the genetic variations detected in *FDC1* and *PAD1* are responsible for the loss of 4-VG production activity in the CEN.PK2-1C and KSD-YC strains. Considering the clear and simple aromatic characteristics of the rice wine-brewed KSD-YC strain, commercialized by Kooksoondang Brewery Co. Ltd., it is highly likely that KSD-YC was selected for 4-VG-free fermentations.

By integrating the genetic variation, detected by comparative genomics, with the growth patterns of the PM analysis, SNPs were mapped within the genes responsible for each phenotype difference, suggesting this genetic variation could be the reason for the observed differences in the PM data (Table 2). For example, several SNPs were detected in the structural and regulatory genes associated with galactose metabolism in the *S. cerevisiae* industrial strains analysed in this study. The KSD-YC strain showed a Gal^- phenotype, which is ascribed due to the start codon deletion of *GAL3* and to the premature termination mutation of *GAL4*, respectively. The ability of the 98-5 strain to metabolize galactose could be an advantage in the culture medium containing coffee and cocoa beans, which are high in galactose, when using a starter culture of *S. cerevisiae* (Redgwell et al., 2003). On the other hand, the *S. cerevisiae* Y2805 strain can grow robustly using galactose as the sole C-source, although the growth is slower than that of the 98-5 strain, which is advantageous as a host strain to produce recombinant proteins and metabolites under the control of inducible *GAL* promoters. Many foreign genes have been expressed using the *GAL1* and *GAL10* promoters to produce recombinant therapeutic proteins in Y2805, including the recombinant human papillomavirus (HPV) types 16, 18 and 58 virus-like particle (VLP) vaccines, which are currently in clinical Phase I in Korea (POSVAX Co., Ltd).

In recent years, CN variation has emerged as a new and significant source of genetic polymorphisms contributing to the phenotypic diversity of populations. A recent intensive study on the genomic variation of 132 wine yeast strains reported that these strains harbour low levels of genetic diversity in the form of SNPs and suggested genomic structural variants, such as CN variants, are substantial contributors to the genomic diversity of the wine yeast strains (Steenwyk & Rokas, 2017). It was found that the gene families involved in fermentation-related processes, such as copper resistance (*CUP*), flocculation (*FLO*) and glucose metabolism (*HXT*), as well as the *SNO* gene family, whose members are expressed before or during the diauxic shift, showed substantial CN diversity across the *S. cerevisiae* wine strains examined. The results of our study showed that none of the four strains contained the key gene for copper resistance, *CUP1*. However, traces of *CUP2*, the transcription factor of *CUP1*, were detected in all four strains (Table S10), indicating a loss of copper resistance function in each of these strains. Our findings also revealed that *FLO1* and *FLO11*, the members of the *FLO* gene family responsible for flocculation, were absent in KSD-YC and 98-5. Furthermore, *FLO10* was also absent in 98-5, suggesting a loss or weak function of flocculation in 98-5 and KSD-YC. Additionally, our analysis showed that *SNO1*, a putative protein for pyridoxine metabolism, was commonly found with a CN variation of 2 in the diploid genomes of KSD-YC and 98-5, as compared to the haploid genomes of S288C and Y2805. However, four copies of *SNO2* were present in 98-5 while lacking *SNO3*, while six copies of *SNO3* were present in KSD-YC while lacking *SNO2* (Kang et al., 2019). The hexose transporter family is large and comprised of the *HXT1-17* gene in *S. cerevisiae*, and duplication can allow the number of hexose transport molecules to be increased in response to selection in a glucose-limited environment (Brown et al., 1998). The reference S288C strain contains the full set of *HXT* genes, while Y2805 lacks *HXT6* and *HXT16* but has two copies of *HTX15*. The diploid genomes of KSD-YC and 98-5 show diverse CN variation from 0 to 8 for all *HXT* genes (Table S10). These findings provide insights into the genetic basis for the observed phenotypic differences between the *S. cerevisiae* strains and highlight the importance of understanding the underlying genetic mechanisms governing these physiological traits at genomic levels.

In conclusion, by integrating high-quality WG sequencing and PM analysis, we were able to map specific SNPs to major phenotypes of three *S. cerevisiae* industrial strains, which provides not only direct correlations between observed phenotypes and genotypes, but also offers a high probability of identifying metabolic targets for further improving the functions of the yeast strains. However, there are still some phenotypes that cannot be fully explained by SNPs, suggesting that genotype to phenotype correlations might be manifested post-transcriptionally or posttranslationally through modulation of protein concentration and/or function. Clearly, future work is needed to validate these correlations through genetic engineering of the identified SNPs to elucidate whether the desired phenotypes, such as improved C or N metabolism, are observed. The intensive examination of CN variation throughout WGs combined with comparative transcriptomics analysis might provide additional information on the phenotypic diversity of *S. cerevisiae* industrial strains.

AUTHOR CONTRIBUTIONS

Ye Ji Son: Data curation (lead); investigation (lead); writing – original draft (supporting). **Min-Seung Jeon:** Data curation (lead); formal analysis (lead); writing – original draft (supporting). **Hye Yun Moon:** Investigation (supporting); writing – original draft (supporting). **Jiwon Kang:** Formal analysis (supporting). **Da Min Jeong:** Investigation (supporting); writing – original draft (supporting). **Dong Wook Lee:** Formal analysis (supporting). **Jae Ho Kim:** Resources (equal). **Jae Yun Lim:** Formal analysis

(supporting). **Jeong-Ah Seo:** Resources (equal). **Jae-Hyung Jin:** Formal analysis (supporting). **Yong-Sun Bahn:** Resources (equal). **Seong-il Eyun:** Conceptualization (equal); supervision (equal); writing – original draft (supporting); writing – review and editing (equal). **Hyun Ah Kang:** Conceptualization (equal); project administration (lead); supervision (equal); writing – original draft (lead); writing – review and editing (equal).

ACKNOWLEDGEMENTS

This research was supported by Rural Development Administration, Republic of Korea. (Cooperative Research Program for Agriculture Science & Technology Development, Project No. PJ01710102) and the National Research Foundation of Korea (Advanced Research Center Program, Grant no. NRF2018R1A5A1025077). This research was also supported by the Chung-Ang University Graduate Research Scholarship Grants in 2021.

CONFLICT OF INTEREST STATEMENT

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

DATA AVAILABILITY STATEMENT

The whole-genome data of *S. cerevisiae* 98-5, KSD-YC and Y2805 have been deposited in the NCBI database under the accession numbers CP025097-CP25112, CP023995-CP24010 and CP093858-CP093873, respectively.

ORCID

Ye Ji Son https://orcid.org/0009-0006-4324-5800
Min-Seung Jeon https://orcid.org/0000-0002-9273-222X
Hye Yun Moon https://orcid.org/0000-0002-0988-8064
Jiwon Kang https://orcid.org/0000-0002-1621-8042
Da Min Jeong https://orcid.org/0000-0002-8531-1152
Dong Wook Lee https://orcid.org/0009-0002-8423-9184
Jae Ho Kim https://orcid.org/0000-0002-6037-427X
Jae Yun Lim https://orcid.org/0000-0002-7321-9068
Jeong-Ah Seo https://orcid.org/0000-0002-0566-1217
Jae-Hyung Jin https://orcid.org/0000-0002-0589-9444
Yong-Sun Bahn https://orcid.org/0000-0001-9573-5752
Seong-il Eyun https://orcid.org/0000-0003-4687-1066
Hyun Ah Kang https://orcid.org/0000-0002-3722-525X

REFERENCES

Akao, T., Yashiro, I., Hosoyama, A., Kitagaki, H., Horikawa, H., Watanabe, D. et al. (2011) Whole-genome sequencing of sake yeast *Saccharomyces cerevisiae* Kyokai no. 7. *DNA Research*, 18(6), 423–434. Available from: https://doi.org/10.1093/dnares/dsr029

Ansmant, I., Massenet, S., Grosjean, H., Motorin, Y. & Branlant, C. (2000) Identification of the *Saccharomyces cerevisiae* RNA:pseudouridine synthase responsible for formation of psi(2819) in 21S mitochondrial ribosomal RNA. *Nucleic Acids Research*, 28(9), 1941–1946. Available from: https://doi.org/10.1093/nar/28.9.1941

Berlin, K., Koren, S., Chin, C.-S., Drake, J.P., Landolin, J.M. & Phillippy, A.M. (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6), 623–630. Available from: https://doi.org/10.1038/nbt.3238

Bochner, B.R. (2003) New technologies to assess genotype–phenotype relationships. *Nature Reviews Genetics*, 4(4), 309–314. Available from: https://doi.org/10.1038/nrg1046

Borneman, A.R., Desany, B.A., Riches, D., Affourtit, J.P., Forgan, A.H., Pretorius, I.S. et al. (2011) Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*. *PLoS Genetics*, 7(2), e1001287. Available from: https://doi.org/10.1371/journal.pgen.1001287

Borneman, A.R., Forgan, A.H., Pretorius, I.S. & Chambers, P.J. (2008) Comparative genome analysis of a *Saccharomyces cerevisiae* wine strain. *FEMS Yeast Research*, 8(7), 1185–1195. Available from: https://doi.org/10.1111/j.1567-1364.2008.00434.x

Borneman, A.R. & Pretorius, I.S. (2015) Genomic insights into the *Saccharomyces sensu stricto* complex. *Genetics*, 199(2), 281–291. Available from: https://doi.org/10.1534/genetics.114.173633

Borneman, A.R., Pretorius, I.S. & Chambers, P.J. (2013) Comparative genomics: a revolutionary tool for wine yeast strain development. *Current Opinion in Biotechnology*, 24(2), 192–199. Available from: https://doi.org/10.1016/j.copbio.2012.08.006

Brown, J.L., Mucci, D., Whiteley, M., Dirksen, M.-L. & Kassis, J.A. (1998) The *Drosophila* polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1. *Molecular Cell*, 1(7), 1057–1064. Available from: https://doi.org/10.1016/s1097-2765(00)80106-9

Capaldi, A.P., Kaplan, T., Liu, Y., Habib, N., Regev, A., Friedman, N. et al. (2008) Structure and function of a transcriptional network activated by the MAPK Hog1. *Nature Genetics*, 40(11), 1300–1306. Available from: https://doi.org/10.1038/ng.235

Carlson, M. & Botstein, D. (1983) Organization of the *SUC* gene family in *Saccharomyces*. *Molecular and Cellular Biology*, 3(3), 351–359. Available from: https://doi.org/10.1128/mcb.3.3.351-359.1983

Choo, J.H., Hong, C.P., Lim, J.Y., Seo, J.-A., Kim, Y.-S., Lee, D.W. et al. (2016) Whole-genome de novo sequencing, combined with RNA-seq analysis, reveals unique genome and physiological features of the amylolytic yeast *Saccharomycopsis fibuligera* and its interspecies hybrid. *Biotechnology for Biofuels*, 9(1), 246. Available from: https://doi.org/10.1186/s13068-016-0653-4

Copetti, M.V. (2019) Yeasts and molds in fermented food production: an ancient bioprocess. *Current Opinion in Food Science*, 25, 57–61. Available from: https://doi.org/10.1016/j.cofs.2019.02.014

Engel, S.R., Dietrich, F.S., Fisk, D.G., Binkley, G., Balakrishnan, R., Costanzo, M.C. et al. (2013) The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3: Genes|Genomes|Genetics*, 4(3), 389–398. Available from: https://doi.org/10.1534/g3.113.008995

Eyun, S. (2017) Phylogenomic analysis of Copepoda (Arthropoda, Crustacea) reveals unexpected similarities with earlier proposed morphological phylogenies. *BMC Evolutionary Biology*, 17(1), 23. Available from: https://doi.org/10.1186/s12862-017-0883-5

Gallone, B., Mertens, S., Gordon, J.L., Maere, S., Verstrepen, K.J. & Steensels, J. (2018) Origins, evolution, domestication and diversity of *Saccharomyces* beer yeasts. *Current Opinion in Biotechnology*, 49, 148–155. Available from: https://doi.org/10.1016/j.copbio.2017.08.005

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B. et al. (2016) Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell*, 166(6), 1397–1410. e16. Available from: https://doi.org/10.1016/j.cell.2016.08.020

Garrison, E. & Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint* arXiv:1207.3907 [q-bio.GN] https://doi.org/10.48550/arXiv.1207.3907

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G. et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11(12), 4241–4257. Available from: https://doi.org/10.1091/mbc.11.12.4241

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H. et al. (1996) Life with 6000 genes. *Science*, 274(5287), 546–567. Available from: https://doi.org/10.1126/science.274.5287.546

Gonçalves, M., Pontes, A., Almeida, P., Barbosa, R., Serra, M., Libkind, D. et al. (2016) Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Current Biology*, 26(20), 2750–2761. Available from: https://doi.org/10.1016/j.cub.2016.08.040

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H. et al. (2018) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1), D309–D314. Available from: https://doi.org/10.1093/nar/gky1085

Jarolim, S., Ayer, A., Pillay, B., Gee, A.C., Phrakaysone, A., Perrone, G.G. et al. (2013) *Saccharomyces cerevisiae* genes involved in survival of heat shock. *G3 (Bethesda)*., 3(12), 2321–2333. Available from: https://doi.org/10.1534/g3.113.007971

Jeon, M.-S., Jeong, D.M., Doh, H., Kang, H.A., Jung, H. & Eyun, S. (2023) A practical comparison of the next-generation sequencing platform and assemblers using yeast genome. *Life Science Alliance*, 6(4), e202201744. Available from: https://doi.org/10.26508/lsa.202201744

Jeong, D.M., Yoo, S.J., Jeon, M.-S., Chun, B.H., Han, D.M., Jeon, C.O. et al. (2022) Genomic features, aroma profiles, and probiotic potential of the *Debaryomyces hansenii* species complex strains isolated from Korean soybean fermented food. *Food Microbiology*, 105, 104011. Available from: https://doi.org/10.1016/j.fm.2022.104011

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C. et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9), 1236–1240. Available from: https://doi.org/10.1093/bioinformatics/btu031

Jung, H., Jeon, M.-S., Hodgett, M., Waterhouse, P. & Eyun, S. (2020) Comparative evaluation of genome assemblers from long-read sequencing for plants and crops. *Journal of Agricultural and Food Chemistry*, 68(29), 7670–7677. Available from: https://doi.org/10.1021/acs.jafc.0c01647

Jung, H., Ventura, T., Chung, J.S., Kim, W.-J., Nam, B.-H., Kong, H.J. et al. (2020) Twelve quick steps for genome assembly and annotation in the classroom. *PLoS Computational Biology*, 16(11), e1008325. Available from: https://doi.org/10.1371/journal.pcbi.1008325

Jung, K.-M., Park, J., Jang, J., Jung, S.-H., Lee, S.H. & Kim, S.R. (2021) Characterization of cold-tolerant *Saccharomyces cerevisiae* Cheongdo using phenotype microarray. *Microorganisms*, 9(5), 982. Available from: https://doi.org/10.3390/microorganisms9050982

Kang, K., Bergdahl, B., Machado, D., Dato, L., Han, T.-L., Li, J. et al. (2019) Linking genetic, metabolic, and phenotypic diversity among *Saccharomyces cerevisiae* strains using multi-omics associations. *Gigascience*, 8(4), giz015. Available from: https://doi.org/10.1093/gigascience/giz015

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. Available from: https://doi.org/10.1093/molbev/mst010

Kavšček, M., Stražar, M., Curk, T., Natter, K. & Petrovič, U. (2015) Yeast as a cell factory: current state and perspectives. *Microbial Cell Factories*, 14(1), 94. Available from: https://doi.org/10.1186/s12934-015-0281-x

Khatri, I., Tomar, R., Ganesan, K., Prasad, G.S. & Subramanian, S. (2017) Complete genome sequence and comparative genomics of the probiotic yeast *Saccharomyces boulardii*. *Scientific Reports*, 7(1), 371. Available from: https://doi.org/10.1038/s41598-017-00414-2

Kim, H., Yoo, S.J. & Kang, H.A. (2015) Yeast synthetic biology for the production of recombinant therapeutic proteins. *FEMS Yeast Research*, 15(1), 1–16. Available from: https://doi.org/10.1111/1567-1364.12195

Kim, H.R., Kim, J.-H., Ahn, B.H. & Bai, D.-H. (2014) Metabolite profiling during fermentation of makgeolli by the wild yeast strain *Saccharomyces cerevisiae* Y98-5. *Mycobiology*, 42(4), 353–360. Available from: https://doi.org/10.5941/myco.2014.42.4.353

Kim, K.W., Kamerud, J.Q., Livingston, D.M. & Roon, R.J. (1988) Asparaginase II of *Saccharomyces cerevisiae*. *The Journal of Biological Chemistry*, 263, 11948–11953.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. & Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5), 722–736. Available from: https://doi.org/10.1101/gr.215087.116

Kostas, E.T., Cooper, M., Shepherd, B.J. & Robinson, J.P. (2019) Identification of bio-oil compound utilizing yeasts through phenotypic microarray screening. *Waste and Biomass Valorization*, 11(6), 2507–2519. Available from: https://doi.org/10.1007/s12649-019-00636-7

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D. et al. (2009) Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9), 1639–1645. Available from: https://doi.org/10.1101/gr.092759.109

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. et al. (2004) Versatile and open software for comparing large genomes. *Genome Biology*, 5(2), R12. Available from: https://doi.org/10.1186/gb-2004-5-2-r12

Kvitek, D.J., Will, J.L. & Gasch, A.P. (2008) Variations in stress sensitivity and genomic expression in diverse *S. cerevisiae* isolates. *PLoS Genetics*, 4(10), e1000223. Available from: https://doi.org/10.1371/journal.pgen.1000223

Langmead, B. & Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359. Available from: https://doi.org/10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. Available from: https://doi.org/10.1093/bioinformatics/btp352

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A. et al. (2009) Population genomics of domestic and wild yeasts. *Nature*, 458(7236), 337–341. Available from: https://doi.org/10.1038/nature07743

Mukai, N., Masaki, K., Fujii, T. & Iefuji, H. (2014) Single nucleotide polymorphisms of PAD1 and FDC1 show a positive relationship with ferulic acid decarboxylation ability among industrial yeasts used in alcoholic beverage production. *Journal of Bioscience and Bioengineering*, 118(1), 50–55. Available from: https://doi.org/10.1016/j.jbiosc.2013.12.017

Nawrocki, E.P. & Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22), 2933–2935. Available from: https://doi.org/10.1093/bioinformatics/btt509

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. (2014) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1), 268–274. Available from: https://doi.org/10.1093/molbev/msu300

Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S. et al. (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. *Proceedings of the National Academy of Sciences*, 106(38), 16333–16338. Available from: https://doi.org/10.1073/pnas.0904673106

Otero, J.M., Vongsangnak, W., Asadollahi, M.A., Olivares-Hernandes, R., Maury, J., Farinelli, L. et al. (2010) Whole genome sequencing of *Saccharomyces cerevisiae*: from genotype to phenotype for improved metabolic engineering applications. *BMC Genomics*, 11(1), 723. Available from: https://doi.org/10.1186/1471-2164-11-723

Parapouli, M., Vasileiadi, A., Afendra, A.-S. & Hatziloukas, E. (2020) *Saccharomyces cerevisiae* and its industrial applications. *AIMS Microbiology*, 6(1), 1–32. Available from: https://doi.org/10.3934/microbiol.2020001

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A. et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*, 556(7701), 339–344. Available from: https://doi.org/10.1038/s41586-018-0030-5

Ramazzotti, M., Berná, L., Stefanini, I. & Cavalieri, D. (2012) A computational pipeline to discover highly phylogenetically informative genes in sequenced genomes: application to *Saccharomyces cerevisiae* natural strains. *Nucleic Acids Research*, 40(9), 3834–3848. Available from: https://doi.org/10.1093/nar/gks005

Redgwell, R.J., Curti, D., Rogers, J., Nicolas, P. & Fischer, M. (2003) Changes to the galactose/mannose ratio in galactomannans during coffee bean (Coffea arabica L.) development: implications for in vivo modification of galactomannan synthesis. *Planta*, 217(2), 316–326. Available from: https://doi.org/10.1007/s00425-003-1003-x

Richard, P., Viljanen, K. & Penttilä, M. (2015) Overexpression of PAD1 and FDC1 results in significant cinnamic acid decarboxylase activity in *Saccharomyces cerevisiae*. *AMB Express*, 5(1), 12. Available from: https://doi.org/10.1186/s13568-015-0103-x

Rose, M. & Winston, F. (1984) Identification of a Ty insertion within the coding sequence of the *S. cerevisiae* URA3 gene. *Molecular and General Genetics MGG*, 193(3), 557–560. Available from: https://doi.org/10.1007/bf00382100

Seppey, M., Manni, M. & Zdobnov, E.M. (2019) BUSCO: assessing genome assembly and annotation completeness. *Methods in Molecular Biology*, 1962, 227–245. Available from: https://doi.org/10.1007/978-1-4939-9173-0_14

Shimizu, M., Miyashita, K., Kitagaki, H., Ito, K. & Shimoi, H. (2005) Amplified fragment length polymorphism of the AWA1 gene of sake yeasts for identification of sake yeast strains. *Journal of Bioscience and Bioengineering*, 100(6), 678–680. Available from: https://doi.org/10.1263/jbb.100.678

Shin, W.C., Park, S.Y. & Back, S.H. (2019) Novel yeast *Saccharomyces cerevisiae* strain KSD-YC. 2019. KR 10-2018536.

Song, S.H. (2013) Analysis of microflora profile in korean traditional nuruk. *Journal of Microbiology and Biotechnology*, 23(1), 40–46. Available from: https://doi.org/10.4014/jmb.1210.10001

Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24(5), 637–644. Available from: https://doi.org/10.1093/bioinformatics/btn013

Steenwyk, J. & Rokas, A. (2017) Extensive copy number variation in fermentation-related genes among *Saccharomyces cerevisiae* wine strains. *G3: Genes|Genomes|Genetics*, 7(5), 1475–1485. Available from: https://doi.org/10.1534/g3.117.040105

Strope, P.K., Skelly, D.A., Kozmin, S.G., Mahadevan, G., Stone, E.A., Magwene, P.M. et al. (2015) The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Research*, 25(5), 762–774. Available from: https://doi.org/10.1101/gr.185538.114

Suezawa, Y. & Suzuki, M. (2007) Bioconversion of ferulic acid to 4-vinylguaiacol and 4-ethylguaiacol and of 4-vinylguaiacol to 4-ethylguaiacol by halotolerant yeasts belonging to the genus *Candida*. *Bioscience, Biotechnology, and Biochemistry*, 71(4), 1058–1062. Available from: https://doi.org/10.1271/bbb.60486

Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. (2012) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), 178–192. Available from: https://doi.org/10.1093/bib/bbs017

Vaas, L.A., Sikorski, J., Hofner, B., Fiebig, A., Buddruhs, N., Klenk, H.P. et al. (2013) Opm: an R package for analysing OmniLog(R) phenotype microarray data. *Bioinformatics*, 29(14), 1823–1824. Available from: https://doi.org/10.1093/bioinformatics/btt291

Vehkala, M., Shubin, M., Connor, T.R., Thomson, N.R. & Corander, J. (2015) Novel R pipeline for analyzing biolog phenotypic microarray data. *PloS One*, 10(3), e0118392. Available from: https://doi.org/10.1371/journal.pone.0118392

Vidgren, V., Ruohonen, L. & Londesborough, J. (2005) Characterization and functional analysis of the MAL and MPH loci for maltose utilization in some ale and lager yeast strains. *Applied and Environmental Microbiology*, 71(12), 7846–7857. Available from: https://doi.org/10.1128/aem.71.12.7846-7857.2005

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S. et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11), e112963. Available from: https://doi.org/10.1371/journal.pone.0112963

Wang, S.A. & Li, F.L. (2013) Invertase SUC2 is the key hydrolase for inulin degradation in *Saccharomyces cerevisiae*. *Applied Environmental Microbiology*, 79(1), 403–406. Available from: https://doi.org/10.1128/AEM.02658-12

Wieland, J., Nitsche, A.M., Strayle, J., Steiner, H. & Rudolph, H.K. (1995) The PMR2 gene cluster encodes functionally distinct isoforms of a putative Na+ pump in the yeast plasma membrane. *The EMBO Journal*, 14(16), 3870–3882. Available from: https://doi.org/10.1002/j.1460-2075.1995.tb00059.x

Wimalasena, T.T., Greetham, D., Marvin, M.E., Liti, G., Chandelia, Y., Hart, A. et al. (2014) Phenotypic characterization of *Saccharomyces* spp. yeast for tolerance to stresses encountered during fermentation of lignocellulosic residues to produce bioethanol. *Microbial Cell Factories*, 13(1), 47. Available from: https://doi.org/10.1186/1475-2859-13-47

Wronska, A.K., Haak, M.P., Geraats, E., Bruins Slot, E., van den Broek, M., Pronk, J.T. et al. (2020) Exploiting the diversity of Saccharomycotina yeasts to engineer biotin-independent growth of *Saccharomyces cerevisiae*. *Applied and Environmental Microbiology*, 86(12), e00270-20. Available from: https://doi.org/10.1128/aem.00270-20

Yaffe, M.P. & Schatz, G. (1984) Two nuclear mutations that block mitochondrial import in yeast. *Proceedings of the National Academy of Sciences USA*, 81(15), 4819–4823. Available from: https://doi.org/10.1073/pnas.81.15.4819

Yang, F., Liu, Z.C., Wang, X., Li, L.L., Yang, L., Tang, W.Z. et al. (2015) Invertase Suc2-mediated inulin catabolism is regulated at the transcript level in *Saccharomyces cerevisiae*. *Microbial*

Cell Factory, 14, 59. Available from: https://doi.org/10.1186/s12934-015-0243-3

Young, M.J. & Court, D.A. (2008) Effects of the S288c genetic background and common auxotrophic markers on mitochondrial DNA function in *Saccharomyces cerevisiae*. *Yeast*, 25(12), 903–912. Available from: https://doi.org/10.1002/yea.1644

Zhong, Q., Gohil, V.M., Ma, L. & Greenberg, M.L. (2004) Absence of cardiolipin results in temperature sensitivity, respiratory defects, and mitochondrial DNA instability independent of pet56. *Journal of Biological Chemistry*, 279(31), 32294–32300. Available from: https://doi.org/10.1074/jbc.m403275200

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Son, Y.J., Jeon, M.-S., Moon, H.Y., Kang, J., Jeong, D.M., Lee, D.W. et al. (2023) Integrated genomics and phenotype microarray analysis of *Saccharomyces cerevisiae* industrial strains for rice wine fermentation and recombinant protein production. *Microbial Biotechnology*, 16, 2161–2180. Available from: https://doi.org/10.1111/1751-7915.14354