Whole-Genome Sequencing of Sake Yeast *Saccharomyces cerevisiae* Kyokai no. 7

TAKESHI Akao1, ISAO Yashiro2, AKIRA Hosoyama2, HIROSHI Kitagaki1, HIROSHI Horikawa2, DAISUKE Watanabe1, RINJI Akada3, YOSHINORI Ando4, SATOSHI Harashima5, TOYOHISA Inoue6, YOSHIHARU Inoue7, SUSUMU Kajiwara8, KATSUHIKO Kitamoto9, NORIYUKI Kitamoto10, OSAMU Kobayashi11, SATORU Kuhara12, TAKASHI Masubuchi13, HARUHIKO Mizoguchi14, YOSHIHITO Nakao15, AMI Nakazato16, MASAJIRO Namise17, TAKAHIRO Oba18, TOMO Ogata19, AKINORI Ohta9, MASAHIRO Sato20, SEIJI Shibasaka21, YOSHIFUMI Takatsume7, SHOTA Tanimoto22, HIROKAZU Tsuboi23, AKIRA Nishimura24, KOJI Yoda9, TAKEAKI Ishikawa25, KAZUHIKORO Iwashita1, NOBUYUKI Fujita2, and HITOSHI Shimoi1,*

National Research Institute of Brewing, Higashi-hiroshima, Hiroshima 739-0046, Japan1; National Institute of Technology and Evaluation, Shibuya-ku, Tokyo 151-0066, Japan2; Department of Applied Molecular Bioscience, Yamaguchi University Graduate School of Medicine, Ube, Yamaguchi 755-8611, Japan3; Kagoshima Prefectural Institute of Industrial Technology, Kagoshima 899-5105, Japan4; Department of Biotechnology, Graduate School of Engineering, Osaka University, Suita, Osaka 565-0871, Japan5; Nihonsakari Co., Ltd, Nishinomiya, Hyogo 662-8521, Japan6; Division of Applied Life Sciences, Graduate School of Agriculture, Kyushu University, Uji, Kyoto 611-0011, Japan7; Department of Life Science, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Yokohama, Kanagawa 226-8501, Japan8; Department of Biotechnology, University of Tokyo, Bunkyo-ku, Tokyo 113-8657, Japan9; Food Research Center, Aichi Industrial Technology Institute, Nagoya, Aichi 451-0083, Japan10; Central Laboratories for Frontier Technology, Kirin Holdings Co., Ltd, Yokohama, Kanagawa 236-0004, Japan11; Department of Bioscience and Biotechnology, Faculty of Agriculture, Kyushu University, Fukuoka 812-8581, Japan12; Gunma Industrial Technology Center, Maebashi, Gunma 379-2147, Japan13; General Research Laboratory, Kiku-Masamune Sake Brewing Co., Ltd, Kobe, Hyogo 658-0026, Japan14; R&D Planning Division, Products Development Center, Suntory Holdings Ltd, Kawasaki, Kanagawa 211-0067, Japan15; Department of Fermentation, Faculty of Applied Bio-Science, Tokyo University of Agriculture, Setagaya-ku, Tokyo 156-8502, Japan16; Research Institute, Gekkeikan Sake Brewing Co., Ltd, Kyoto 612-8385, Japan17; Biotechnology and Food Research Institute, Fukuoka Industrial Technology Center, Kurume, Fukuoka 839-0861, Japan18; Research Laboratories of Brewing Technology, Asahi Breweries, Ltd, Moriya, Ibaraki 302-0106, Japan19; Frontier Laboratories of Value Creation, Sapporo Breweries Ltd, Yatsu, Shizuoka 425-0013, Japan20; School of Pharmacy, Hyogo University of Health Sciences, Kobe, Hyogo 650-8530, Japan21; Food Technology Research Center, Hiroshima Prefectural Technology Research Institute, Hiroshima 732-0816, Japan22; General Research Laboratory, Ozeki Corporation, Nishinomiya, Hyogo 663-8227, Japan23; Research and Development Department, Hakutsuru Sake Brewing Co., Ltd, Kobe, Hyogo 658-0041, Japan24 and Brewing Society of Japan, Kita-ku, Tokyo 114-0023, Japan25

*To whom correspondence should be addressed. Tel. +81 82-420-8016. Fax. +81 82-420-0809.
E-mail: simoi@nrib.go.jp

Edited by Katsumi Isono
(Received 25 January 2011; accepted 26 July 2011)

**Abstract**

The term ‘sake yeast’ is generally used to indicate the *Saccharomyces cerevisiae* strains that possess characteristics distinct from others including the laboratory strain S288C and are well suited for sake brewery. Here, we report the draft whole-genome shotgun sequence of a commonly used diploid sake yeast strain, Kyokai no. 7 (K7). The assembled sequence of K7 was nearly identical to that of the S288C, except for several subtelomeric polymorphisms and two large inversions in K7. A survey of
heterozygous bases between the homologous chromosomes revealed the presence of mosaic-like uneven distribution of heterozygosity in K7. The distribution patterns appeared to have resulted from repeated losses of heterozygosity in the ancestral lineage of K7. Analysis of genes revealed the presence of both K7-acquired and K7-lost genes, in addition to numerous others with segmentations and terminal discrepancies in comparison with those of S288C. The distribution of Ty element also largely differed in the two strains. Interestingly, two regions in chromosomes I and VII of S288C have apparently been replaced by Ty elements in K7. Sequence comparisons suggest that these gene conversions were caused by cDNA-mediated recombination of Ty elements. The present study advances our understanding of the functional and evolutionary genomics of the sake yeast.

**Key words:** *Saccharomyces cerevisiae*; sake yeast; genome sequence; diploid; loss of heterozygosity

1. Introduction

Sake is a traditional Japanese alcoholic beverage that is fermented from steamed rice by the concerted action of two types of microorganisms, filamentous fungi and yeast. In the production of sake, enzymes secreted by the fungus *Aspergillus oryzae*, which is grown on steamed rice, convert rice starch into glucose. Yeast cells in the sake mash then produce ethanol, higher alcohols and their esters, organic acids and amino acids, which are important components that contribute to sake aroma and taste. Hence, the choice of a yeast strain is one of the most critical factors in determining the resulting aroma and taste characteristics of sake products. Yeast strains that were originally isolated in sake breweries have been identified as *Saccharomyces cerevisiae* and are now commercially distributed as sake yeast.1

Phylogenetic studies conducted using DNA markers have indicated that sake yeast strains are closely related, forming a sake strain cluster that belongs to a lineage distinct from other industrial and laboratory strains in the phylogenetic tree of *S. cerevisiae*.2–5 Comprehensive genome-wide studies of diverse *S. cerevisiae* strains have also indicated the existence of a unique sake cluster that is distinct from wine and laboratory strains.6,7 Consistent with their unique phylogenetic position, sake yeast strains possess characteristic traits that differ from other *S. cerevisiae* strains and are ideal for sake brewing, including ability for high ethanol productivity (reaching 20%)1,8,9 efficient growth and fermentation at low temperatures (below 15°C).1 In addition, nearly all sake yeast strains generate a foam on the mash surface during the brewing process, which results from yeast cells fermenting sugars into CO2 bubbles10 and possess biotin biosynthetic ability.11

Several sake yeast-specific genes that are responsible for the desirable features of sake yeast and affect the brewing process have been identified. For example, *AWA1*, which encodes a cell-surface hydrophobic protein with a GPI anchor, was identified from a sake yeast genome and is responsible for foam formation in sake mash.10 Many *S. cerevisiae* strains are biotin auxotrophs due to the lack of certain biotin biosynthetic pathway genes. *BIO6*, a homolog of a bacterial biotin biosynthetic pathway gene, was identified in sake yeast strains and is essential in them for the production of biotin.11 In addition to these studies, several genes involved in yeast’s quantitative phenotypes, including fermentability and aroma production, have been suggested.12,13 However, because the genetic basis for the superiority of sake yeast in sake brewing is largely unknown, genome-wide genetic approaches are required to understand these complex traits. Since the first complete genome sequencing of strain S288C in 1996,14 the genomes of several other *S. cerevisiae* strains have been sequenced.15–18 Strains such as K11, Y9 and Y12 were also subjected to genome sequencing and the sake cluster was consequently proposed. However, in each case, total length of sequence reads reached <0.9-fold of the haploid genome, not enough to get whole picture of the genome.6 In addition, these sequenced strains are not the typical industrial sake yeast strains. Accordingly, yeast strains used in sake brewing have not been subjected to whole-genome analyses, despite their importance in industry and yeast phylogenetic systematics.

In the present study, we performed the whole-genome sequencing of the sake yeast strain K7 (kyokai is the Japanese word for ‘society’) as the first step for subsequent functional, phylogenetic and evolutionary genomic studies. K7 has been one of the most extensively used industrial sake yeast strains over the past several decades and has also been employed in numerous genetic and biochemical studies as a model sake yeast and a parent strain for breeding.8,10–13,19–24 Here, we report the overall chromosome structure and remarkable features of the K7 genome.

2. Materials and methods

2.1. Strain

*Saccharomyces cerevisiae* K7 was distributed to Japanese sake breweries by the Brewing Society of Japan in 2004 and used as a DNA donor.
2.2. Sequencing, assembly and validation

The nucleotide sequence of the K7 genome was determined using the whole-genome shotgun sequencing approach. Genomic DNA was isolated from cells according to the method described by Hereford et al. Plasmid libraries with average insert sizes of 1.6 and 5.0 kb were constructed in pUC118 (Takara Bio Inc.), while a fosmid library with an average insert size of 35 kb was constructed in pCC1FOS (Epicentre Biotechnologies), as described previously. Raw sequence reads corresponded in pCC1FOS (Epicentre Biotechnologies), as described previously.26 Raw sequence reads corresponding to a 9.1-fold coverage of the haploid genome (42 842, 88 830 and 15 227 reads from libraries with 1.6, 5.0 and 35 kb inserts, respectively) were first obtained by sequencing from both ends of the inserts on an ABI 3730xl DNA Analyzer (Applied Biosystems). Sequence reads were trimmed at a threshold quality value (Phred) of 20 and assembled using the Phrap assembler.27,28 We obtained a total of 712 contigs that were then put in order based on paired-end information from the constructed fosmid library to obtain supercontigs. The overall assembly was then validated and refined using Optical Mapping (OpGen, Inc.). A number of short contigs were incorporated into supercontigs with the assistance of optical maps and transposon-mediated random sequencing from fosmid clones (2112 reads). Following this analysis, the final contig number was reduced to 706.

2.3. Sequence comparison

An overall comparison of the K7 chromosomes with S288C (NC_001133–NC_001148), EC1118 (FN393058–FN393060, FN393062–FN393087, FN394216 and FN394217) and YJM789 (A0AF 2000000) strain chromosomes was performed using MUMmer 3.0 software.29 Similarity-based searches of individual genes were performed using BLASTN searches of the K7 genome with the S288C genome supercontigs. When direct comparison was difficult, ORFs were predicted using the software programs CRITICA,33 Glimmer2,34 GlimmerHMM35 and SIM4.36 Finally, all K7 ORFs were manually validated by expert annotators. When one or more incomplete ORFs, such as those truncated by a sequence gap and lacking a start or a stop codon, were mapped to a single S288C ORF, each incomplete K7 ORF was annotated as a single ORF. Functional annotation was based primarily on the Saccharomyces Genome Database (SGD; http://www.yeastgenome.org/), secondarily on the Saccharomyces species database (yeast comparative genomics: http://www.broadinstitute.org/annotation/fungi/comp_yeasts/) and also on COG/KOG (http://www.ncbi.nlm.nih.gov/COG/) and DBJ/EMBL/GENBANK non-redundant databases. Orthology with the S288C ORF was evaluated using the BLASTP similarity and calculated as the percent of matched amino acid residues versus the total covered region between a K7 ORF and the best-hit S288C ORF (Supplementary Table S4) as truncated by a sequence gap. Similarity was calculated by the number of matching residues in only the corresponding regions of the S288C ORF. Dubious ORFs, ORFs in Ty elements and ORFs in telomeric regions were excluded as possible protein-coding genes and were not annotated. Prediction and annotation of RNA genes, Ty elements including solo long terminal repeats (LTRs) and telomeric elements were manually performed based on the results of BLASTN searches of the K7 genome with the S288C sequences of these genes and elements as queries.

All annotated ORFs and genetic elements were given individual numbers (Supplementary Table S4). Nomenclature of the K7 genes was based on the following rules: (i) each protein-encoding or RNA gene was named according to the orthologous S288C gene name (with >80% similarity) and the systematic name (with >50% similarity) given in SGD; (ii) K7 identification numbers or K7 original gene names, such as AWA1, were given to genes that were non-orthologous or of low similarity to S288C genes (with ≤50% similarity); (iii) each name of a gene truncated by a sequence gap or segmented by point mutations was followed by a lower case ‘a’, ‘b’ or ‘c’, such as ‘XXX1a’ and ‘XXX1b’, to show its correspondence to a partial region of the ortholog; and (iv) Ty elements and LTRs were independently termed according to the identical nomenclature used for S288C.

2.4. Extraction of heterozygous positions

To determine the positions of heterozygosity in the K7 genome, nucleotide positions containing one or more bases different from the consensus base with the Phred quality values >20 were automatically extracted as candidates for heterozygous positions. The identified positions were then manually validated on the electropherogram to validate the sequence. Non-supercontig contigs, which failed to be assembled into supercontigs, were not included in the analysis because their chromosomal positions were unknown.

2.5. Gene prediction and annotation

For predicting protein-encoding genes, ORFs larger than 90 bp were comprehensively included as candidates. ORF prediction was then carried out based on a direct comparison of S288C ORFs with the K7 genome supercontigs. When direct comparison was difficult, ORFs were predicted using the software programs CRITICA,33 Glimmer2,34 GlimmerHMM35 and SIM4.36 Finally, all K7 ORFs were manually validated by expert annotators. When one or more incomplete ORFs, such as those truncated by a sequence gap and lacking a start or a stop codon, were mapped to a single S288C ORF, each incomplete K7 ORF was annotated as a single ORF. Functional annotation was based primarily on the Saccharomyces Genome Database (SGD; http://www.yeastgenome.org/), secondarily on the Saccharomyces species database (yeast comparative genomics: http://www.broadinstitute.org/annotation/fungi/comp_yeasts/) and also on COG/KOG (http://www.ncbi.nlm.nih.gov/COG/) and DBJ/EMBL/GENBANK non-redundant databases. Orthology with the S288C ORF was evaluated using the BLASTP similarity and calculated as the percent of matched amino acid residues versus the total covered region between a K7 ORF and the best-hit S288C ORF (Supplementary Table S4) as truncated by a sequence gap. Similarity was calculated by the number of matching residues in only the corresponding regions of the S288C ORF. Dubious ORFs, ORFs in Ty elements and ORFs in telomeric regions were excluded as possible protein-coding genes and were not annotated. Prediction and annotation of RNA genes, Ty elements including solo long terminal repeats (LTRs) and telomeric elements were manually performed based on the results of BLASTN searches of the K7 genome with the S288C sequences of these genes and elements as queries.

All annotated ORFs and genetic elements were given individual numbers (Supplementary Table S4). Nomenclature of the K7 genes was based on the following rules: (i) each protein-encoding or RNA gene was named according to the orthologous S288C gene name (with >80% similarity) and the systematic name (with >50% similarity) given in SGD; (ii) K7 identification numbers or K7 original gene names, such as AWA1, were given to genes that were non-orthologous or of low similarity to S288C genes (with ≤50% similarity); (iii) each name of a gene truncated by a sequence gap or segmented by point mutations was followed by a lower case ‘a’, ‘b’ or ‘c’, such as ‘XXX1a’ and ‘XXX1b’, to show its correspondence to a partial region of the ortholog; and (iv) Ty elements and LTRs were independently termed according to the identical nomenclature used for S288C.

3. Results and discussion

3.1. Sequencing and assembly of the K7 genome

Genome sequencing of the diploid sake yeast strain K7 was performed by a whole-genome shotgun
method that yielded 1.49 × 10^5 sequence reads with an estimated 9.24-fold redundancy of the haploid genome. Following de novo sequence assembly, 17 supercontigs were generated. Since homologous chromosome pairs of K7 were almost indistinguishable from each other during the assembly process, nearly all the reads from each homologous chromosome pair were assembled together into a single supercontig. Consequently, the resulting consensus chromosomal sequences represented a diploid genome, although they seemed to be that of a haploid genome. The total length of the supercontig corresponded to 98.1% of the estimated K7 genome size. The sequencing results and the assembled supercontigs are summarized in Supplementary Tables S1 and S2.

Comparison with the S288C genome revealed that the 17 supercontigs corresponded to the set of S288C chromosomes and mitochondrial DNA (Fig. 1; Supplementary Table S2). No DNA sequence for the yeast 2-mg plasmid was detected from any reads, which was consistent with a previous study.

In the assembly process, only 200 of 706 contigs were used for generating the supercontigs because the remaining 506 contigs failed to align with any of the supercontigs. The total length of these non-supercontig contigs was ~604 kb. Of the 506 non-supercontig contigs, 360 (71.1%) were singletons and 395 (78.1%) contained <1000 bases (Supplementary Fig. S1A and B). The majority of these contigs seem to be excluded from the supercontigs due to the repetition of telomeric and Ty-related type sequences. The remainder seemed to be excluded as a result of considerable heterozygosity between the two homologous chromosomes due to base substitution, in-del, Ty insertion or other such events (Supplementary Fig. S1C and D). Several non-supercontig contigs contained ORF-like sequences that were not found in the supercontigs. For example, the nucleotide sequence of MATa was found in a non-supercontig contig due to its considerable heterozygosity, while the supercontig corresponding to the K7 chromosome III only contained the MATalpha sequence, even though K7 possesses both mating-type loci, MATalpha and MATa. Contigs including the sequence corresponding to VTH1/VTH2, which are paralogous and nearly indistinguishable from each other, were not assembled in the supercontigs, presumably due to adjacent repetitive sequences. Similarly, the sequence for ARR3 was also present only in the non-supercontig contigs (data not shown). The non-supercontig contigs were excluded from subsequent analyses such as comparison with related strains, gene prediction and the survey of heterozygosity and Ty elements.

To date, the genome sequences of several *S. cerevisiae* strains have been reported; therefore, we compared the K7 genome with available genomes. Pairwise nucleotide polymorphisms among four strains (K7, S288C, YJM789 and EC1118) were analyzed by sequence alignment using MUMmer 3.0 software. The number of substitutions and small indels between K7 and the three other strains ranged from ~67 900 (5.6/kb) to 78 000 (6.5/kb) and ~19 300 (1.6/kb) to 23 500 (2.0/kb), respectively, while those among the remaining three non-K7 strains ranged from ~46 100 (3.8/kb) to 56 700 (4.7/kb) and ~14 800 (1.2/kb) to 16 700 (1.4/kb), respectively (Supplementary Fig. S2). These results indicate

Figure 1. Dot-plot alignments of homologous chromosomes of *S. cerevisiae* strains K7 and S288C. The nucleotide sequence of each K7 supercontig was compared with that of the homologous S288C chromosome (indicated above each alignment) using MUMmer 3.0.
that the phylogenetic position of K7 is relatively distant from that of S288C, YJM789 and EC1118, as expected from previous studies.2,3,5–7,17

3.2. Inversions in chromosomes V and XIV

We compared chromosomal structures between K7 and S288C using MUMmer 3.0 software,29 as shown in Fig. 1. Although the overall genome structure of K7 closely resembled that of S288C, we identified two types of chromosomal rearrangements. One type involved several complicated subtelomeric rearrangements that are also observed or suggested by previous genome-wide studies of various S. cerevisiae strains, indicating that such rearrangements were not infrequent events.14,17,18,38–40 The other type of rearrangement was characterized by two large internal inverted regions. We confirmed that these inversions were homozygous by PCR analysis (data not shown). One inversion of the ∼100-kb region on the right arm of chromosome V had not previously been described. Both boundary regions of this inversion on the K7 chromosome V were flanked by two Ty2 elements (K7_YERCTy2-3 and K7_YERWTy2-4) that were inverted in relation to each other and were absent in the corresponding region of S288C. Therefore, these Ty2 elements have been proposed to mediate this reciprocal inversion (Supplementary Fig. S3).15 Since Ty insertions differ by strain or lineage, this inversion would be unique to K7 or related lineages (Supplementary Fig. S3). The second identified inversion was a ∼30-kb region on the left arm of chromosome XIV. This inversion was also observed in other strains.15,17 Inverted homologous regions located close to the breaking points on chromosome XIV (YNL018C–YNL019C and YNL033W–YNL034W) may have mediated this reciprocal inversion.

3.3. Uneven distribution of heterozygosity

The sequence obtained for the K7 genome represents the consensus haploid sequence derived from the two homologous chromosomes, although K7 is a diploid. Accordingly, we surveyed the positions of heterozygosity by carefully examining the sequence reads. A total of 1347 heterozygous sites between the homologous chromosomes were detected, and their positions were subsequently mapped (Fig. 2). Interestingly, the windows containing multiple heterozygous sites were unevenly distributed as clusters in several chromosomal regions. Simple evolutionary accumulation of point mutations is insufficient to explain such uneven distribution of heterozygosity.41 It is more likely that a heterozygous diploid was generated by the out-crossing of two different haploid strains and subsequent loss of heterozygosity (LOH), resulting in the observed pattern. Alternatively, a complex history of out- and/or back-crossings of the ancestral strain in natural environments could have also caused this pattern. However, repeated backcrossing is unlikely to have occurred in nature since the sporulation efficiency of sake yeast, including K7, is markedly low. Thus, we speculate that sequential LOH events have resulted in the uneven distribution of heterozygosity in K7. Conversely, it is reasonable to presume that isolated heterozygosities were introduced by point mutations independent of LOH events. Genome-wide LOH was
also proposed for the diploid strain YJM128.\textsuperscript{41} However, LOH in K7 was far more extensive than that found in YJM128, resulting in 82.7% of the entire genome being almost homologous. As LOH is mainly caused by mitotic recombination, the probability of LOH events is dependent on the number of clonal generations during asexual proliferation. Therefore, it is likely that K7 passed through more mitotic generations than YJM128 after the out-crossing event. A sporulation defect of K7 may have contributed to a long-term clonal proliferation that allowed extensive LOH events without meiosis. LOH events also result in the selection of one haplotype between two homologous chromosomes whose haplotypes differ from each other. Consequently, LOH can be a major driving force in the diversification and microevolution of diploid strains, such as K7, which lost a meiotic life cycle, as observed in Candida albicans.\textsuperscript{42}

3.4. Gene predictions

Following the sequencing and assembly of the K7 genome, we predicted and annotated 5815 ORFs on 16 nuclear chromosomes and mitochondrial DNA (Supplementary Table S4). We observed many incomplete ORFs interrupted by a sequence gap between contigs. 39 ORFs had a truncated terminal at one end and 124 ORFs (62 pairs) had internal gaps. When compared with the S288C genome, frame shifts caused by small indels and single-nucleotide changes at the start or the stop codons resulted in many ORF polymorphisms, including terminal disagreement, such as extension or truncation (132 genes), segmentation of a single S288C ORF into multiple K7 ORFs (89 genes corresponding to 43 orthologs in S288C) and the fusion of ORFs (13 genes corresponding to 26 orthologs in S288C; Supplementary Table S4). The influence of these polymorphisms on each respective gene function is unclear and remains to be elucidated, although for several cases, such as the K7 ortholog of MSN4, polymorphism appears to influence the characteristic features of K7.\textsuperscript{43}

The average BLAST similarity between K7 ORFs and the most similar S288C ORFs was greater than 95% (95.7 and 95.8% at the nucleotide and amino acid levels, respectively). More than 90% of the K7 ORFs displayed a similarity exceeding 97%, with the corresponding S288C ortholog at the amino acid level (Fig. 3). Genes for tRNAs and other non-coding RNA were also annotated (Supplementary Table S3), which revealed that they corresponded to an almost complete set of S288C RNA genes. However, K7 lost three tRNA genes and one copy of two RUF5 ncRNA genes (data not shown). Their absence would not be expected to have an effect on the cellular functions of K7 because tRNA genes corresponding to a specific codon are highly redundant and one copy of RUF5 was still present in the K7 genome.

3.5. Differentially present genes between K7 and S288C

The comparison of ORFs between K7 and S288C genomes disclosed 97 differentially present genes: 48 ORFs unique to K7 and 49 ORFs unique to S288C (Supplementary Tables S5 and S6). In this analysis, we excluded differences in subtelomeric multicopy gene families such as PAU, COS, DAN, SNO and MAL. Many of the unique genes were located at subtelomeric plastic regions of K7 and S288C, indicating that they may have been acquired or lost by chromosomal rearrangement events. Most differentially present genes located in internal chromosomal regions resulted from small mutations such as frame shifts or gene duplication events in a single strain.

3.5.1. K7 genes absent in S288C

The genes predicted as present in K7 that are absent in S288C are listed in Supplementary Table S5. Several K7 genes already demonstrated involvement in the characteristic features of sake yeast, including K7_AWA1 (K07\_06182) and a paralogous set of K7_BIO6 genes (K7\_BIO6-1/BIO6-2a/BIO6-2b/BIO6-3/BIO6-4a/BIO6-4b: K07\_11198/11203/11204/03384/11206/11207), which are unique to sake yeast strains.\textsuperscript{10,11} BIO1, which is absent in S288C, is also required for biotin biosynthesis in yeast.\textsuperscript{44} As expected from the biotin prototrophy displayed by K7, BIO1 orthologs were also found in K7 (K7_BIO1-1/BIO1-2/BIO1-3: K07\_00624/03376/03381).\textsuperscript{11,45}

Three paralogous genes that are not found in S288C, K07\_00009/11194/04100 (named K7_EHL1/EHL2/EHL3 in this study), were predicted to encode proteins similar to bacterial epoxide hydrolase. Numerous
bacterial orthologs to K7_EHL1/EHL2/EHL3 have been identified in the DDBJ/EMBL/GenBank non-redundant database, whereas eukaryotic orthologs have only been found in S. paradoxus to date (Supplementary Fig. S4). Thus, these epoxide hydrolase genes may have been horizontally transferred from bacteria to a common ancestor of these yeasts. K7_EHL1/EHL2/EHL3 should be involved in the detoxification of harmful epoxide compounds. However, the actual substrates are unknown, and to date, no epoxide compounds have been identified from sake or fermenting sake mash.

K7_02354, which was located in a subtelomeric region, is orthologous to YJM-GNAT of YJM789, which encodes a gene similar to bacterial GCN5-related N-acetyltransferase, suggesting that a common ancestor of K7 and YJM789 acquired their ancestral gene by horizontal transfer. Sequences of K7_02354 and YJM-GNAT were almost identical (over 99%) at both the nucleotide and the amino acid levels (Supplementary Fig. S5).

K7_KHR1 (K07_03550), which encodes a previously identified heat-resistant killer toxin,45 was located in an internal region of chromosome IX and wedged between two solo LTRs. Although a similar structure was observed in the EC1118 genome,17 S288C does not possess KHR1, and only a solo LTR (YILCdelta3) is located in the corresponding locus. This suggests that the loss of KHR1 in S288C was caused by LTR-mediated recombination, as predicted in a previous study.17 A large proportion of genes unique to K7 (Supplementary Table S5) have not been characterized, and their involvement and function in the characteristic features of sake yeast remain to be explored.

3.5.2. S288C genes absent in K7
Forty-nine genes predicted as present in S288C but absent in K7 are listed in Supplementary Table S6. We confirmed that these genes are not present in K7 non-supercontigs. Notably, two subtelomeric paralogous blocks in the S288C genome, containing HXT15–SOR2–MPH2 on chromosome IV and HXT16–SOR1–MPH3 on chromosome X, were not identified in the K7 genome. The paralogous pairs of HXT15 and HXT16, SOR1 and SOR2 and MPH2 and MPH3 encode nearly identical hexose transporters, sorbitol dehydrogenases and maltose transporters, respectively.46–48 It is likely that non-reciprocal chromosomal recombinations in subtelomeric regions caused duplication of these sequences in S288C, but resulted in their loss in the K7 lineage. Another subtelomeric gene, AIF1, which is located on S288C chromosome XIV and encodes a mitochondrial cell death factor,49 was also lost in K7. CWP1, encoding a cell-wall protein linked to glucan chains,50 was disrupted by a frame-shift mutation in K7. Although the effects of the loss of this protein are unclear, it is possible that the cell-wall properties of K7 are affected.

PPT1, which encodes a protein phosphatase,51 is located at an internal region of chromosome VII in S288C, whereas the corresponding 2.6-kb region was lost and replaced with a Ty element (K7_YGRCTy2-2) in the K7 genome (Fig. 4A). The effect of the loss of PPT1 on cellular function or the sake brewing character of K7 is unclear. A tRNA gene tr(UUC)G3 on the left side of PPT1 was also absent in K7. However, the t(AAU)G gene located on the right side of PPT1 in S288C was present in K7. Sequences located several hundred bases upstream of a tRNA gene can serve as potential target sites for Ty integration. Consequently, multiple Ty insertion and subsequent excision events may have resulted in the two solo LTRs on both sides of PPT1 in a K7 ancestral genome. Recombination, including insertion or gene conversion, between a Ty cDNA and yeast chromosome with LTRs, was reported in laboratory experiments.52 Therefore, we speculate that gene conversion between a double-stranded Ty cDNA and two solo LTRs pre-existing on the chromosome VII may have occurred in the K7 ancestor, resulting in the replacement of PPT1 with Ty2 (Fig. 4, Supplementary Fig. S6). Alternatively, double reciprocal crossing-over may have resulted in the observed structure; however, this is unlikely, as PPT1 was not found in either the supercontigs or the non-supercontig contigs of K7. A similar structure was also observed in the K7 right arm of chromosome I (Fig. 4B), involving the replacement of two tRNA genes tL(CAA)A and tS(AGA)A with Ty2, which we speculate may have involved a similar gene conversion mechanism. In previous studies, Ty cDNA-mediated gene conversions were only observed in genetically modified strains and were thought to be an artificially induced phenomena.52 The structures identified in the K7 genome would represent the first example of the spontaneous direct Ty-mediated gene conversion in a wild-type strain.

Two tandemly-duplicated acid phosphatase genes, PHO3 and PHO5, are located on chromosome II of the S288C genome.53 However, K7 possessed only PHO5 (K07_00381), consistent with the observation that K7 shows repressive, but not constitutive, acid phosphatase activity.54 These tandemly arranged acid phosphatase genes are also present in YJM789, EC1118 and several other Saccharomyces species.15,17,55 The prevalence of these genes suggests that PHO3 was looped out of the K7 genome by homologous recombination. In the K7 genome, no copies of ASP3 and its neighboring ORFs, which encode cell-wall l-asparaginase II and putative proteins of unknown function, were found, even though S288C
chromosome XII contains at least four copies of ASP3 adjacent to rDNA repeats. These sequences are also absent in numerous other S. cerevisiae strains. Although tandemly triplicated ENA1/ENA2/ENA5, encoding P-type ATPases are located on chromosome IV in S288C, only one copy was identified in K7 (K7_ENA1/K7_01190), as observed in many other strains. The triplication is a relatively specific feature of S288C and its related strains. The K7 genome also contained one copy of CUP1 (K7_CUP1/K07_03116), encoding a metallothionein, whereas S288C contains two copies in tandem, consistent with the lower copper resistance of K7 than that of X2180, an isogenic diploid of S288C (data not shown).

3.6. Hexose transporter and GAL genes
Transport of sugars across the membrane is one of the key steps in ethanol fermentation. The S288C genome contains 20 genes that encode hexose transporter family proteins (HXT1–HXT17 and GAL2) and glucose sensors (SNF3 and RGT2) as shown in Fig. 5. The products of these genes show different glucose affinities and are differentially expressed in order to coordinately control glucose uptake in environments with a broad range of glucose concentrations. Many of these genes, including two glucose sensors, were highly conserved among S288C and K7. In particular, both the low-affinity glucose transporter genes HXT1/HXT3 were conserved, implying that they may be responsible for glucose uptake during the sake brewing process, as reported in wine yeast. HXT5/HXT6/HXT7 were located at contig ends, and the DNA sequences of these regions in K7 were not completely analyzed. In S288C, HXT6 and HXT7 encode nearly identical proteins and are arranged in tandem; however, they may have been combined into a chimeric gene (HXT6/7) in K7, as observed in other strains.

We revealed that two distinct classes of S288C HXT genes displayed almost altered gene structures in K7. The first contains HXT9/HXT11 and HXT12, which is annotated as a possible pseudogene due to a frameshift mutation in SGD. In the K7 genome, although one of the duplicated HXT9 orthologs was conserved (K7_HXT9-2/K07_02205: 98% identical to HXT9 at the amino acid level), all other HXT9/HXT11 orthologs were divided into two ORFs due to frame shifts.
The disrupted structure of HXT12 was also conserved in K7 (K7_HXT12a/HXT12b: K07_03394/03395) as well as in S288C (Supplementary Table S4). Four genes HXT13/HXT15/HXT16/HXT17 are classified into the second HXT class. In the K7 genome, the HXT13/HXT17 orthologs (K7_HXT13a/HXT13b/HXT17a/HXT17b: K07_01805/01807/06162/06164) contained frame-shift mutations that may have resulted in the loss of function (Supplementary Table S4), while the sequences corresponding to HXT15/HXT16 were absent in the K7 genome (Supplementary Table S6). Thus, in K7, the functions of gene products in this class appeared to be completely lost, although their molecular functions were unclear.

In the K7 genome, we also identified frame-shift mutations in the GAL3 and GAL4 genes that divided each gene into two ORFs (K07_01156 and K07_01157 for GAL3 and K07_06847 and K07_06846 for GAL4). Thus, both Gal3p and Gal4p in K7 are shorter at their C termini than their orthologs in S288C, suggesting that their molecular functions may be impaired. Since Gal3p and Gal4p function as an inducer and activator, respectively, which constitute the transcriptional induction system of galactose assimilating genes, the loss of Gal3p and Gal4p functions may lead to defective GAL gene induction in response to galactose. Consistent with this speculation, the assimilation and fermentation of galactose are remarkably weakened in several sake yeast strains, including K7.

Our analyses suggest that K7 possesses different sugar uptake and assimilation properties from those of S288C. It is likely that sake yeast cells could tolerate a functional loss from these genes without a negative selective force due to the growth of these strains in the glucose-rich environment specific to sake mash.

3.7. Structure of the HO gene

As K7 is a heterothallic diploid, HO involved in mating-type switching was predicted as non-functional in K7. Indeed, a homozygous mutation of A1424T (H475L), which was reported to cause a loss of function in S288C, was observed in the K7 ho allele. Moreover, a 36-amino acid deletion at 524–559 was also present in K7, as was reported in the heterothallic bioethanol strain JAY291. Collectively, these differences are likely to be responsible for the observed K7 heterothallism.

3.8. Comparison of Ty elements and LTRs

The chromosomal insertion of a Ty element can result not only in the loss or alteration of gene function, but may also modify gene expression levels. The annotated Ty elements and solo LTRs from K7 are summarized in Supplementary Tables S7 and S8. Nearly all Ty and solo LTR insertions followed the target-site-selection rule, displaying preferential insertion within a 1-kb upstream region of RNA polymerase III-target genes (Supplementary Table S8). We compared Ty insertion events, which included intact Ty elements and solo LTRs, between K7 and S288C. If flanking regions of two Ty elements, a Ty element and a solo LTR or two solo LTRs, were identical between the two strains, we estimated that these insertion sequences should be derived from the same Ty element insertion event in the common ancestral strain. A total of 198 Ty insertion events were estimated to be identical between K7 and S288C, while 121 and 137 insertion events were unique to K7 and S288C, respectively (Supplementary Tables S7B and S8). Interestingly, among the same 198 insertion events, only one pair
maintained intact Ty structures in both strains: K7_YCLWTy5-1 and YCLWT Ty5-1. In addition, two pairs kept intact Ty structures in one of the strains: YARCTy1-1 in S288C and K7_YEWTy3-1 in K7. All other insertion events were observed as solo LTRs, representing trace sequences of Ty elements (Supplementary Table S8). This observation supports the idea that same insertion events may occur in the much more distant past than the unique insertion events. In our analyses, we were unable to locate K7 genes that were interrupted by a Ty insertion.

4. Conclusions

We revealed the sequence and structure of the K7 genome that represents the first such study of a sake yeast lineage within S. cerevisiae and provides the basis for future studies on the brewing characteristics, genealogy and evolution of sake yeast. The phenotypic effects of the identified structural polymorphisms between K7 and S288C genomes are largely unknown and remain to be explored. In addition, the uneven heterozygosity distribution found in the K7 genome is suggestive of the microevolution of K7 and related sake yeast strains. Future genetic studies encompassing a wide range of Saccharomyces strains are necessary to resolve the genetic basis for the characteristics and evolution of sake yeast strains in greater detail.

The sequences and annotations reported in this paper have been deposited at DDBJ/EMBL/GenBank under the accession no. BABQ01000001–BABQ01000705, DG000037–DG000052 and AP012028. Information of the sequence and gene annotation are also available on the sake yeast genome database (http://nrfb1.nrib.go.jp/SYGD/) and database of the genomes analyzed at NITE (DOGAN; http://www.bio.nite.go.jp/dogan/top/).

Acknowledgements: The authors thank Tokio Ichimatsu, Hikaru Suenaga (Fukuoka Industrial Technology Center), Takaumi Kubohara, Hiroyuki Senju, Nobuo Yamashita (Hakutsuru Sake Brewing Co., Ltd), Nihonsakari Co., Ltd), Hideki Hirakawa (Kazusa DNA Research Institute), Yasuyuki Masuda, Tasuku Yamada, Yasuhiko Wada (Kiku-Masamune Sake Brewing Co., Ltd), Shigejiro Ikushima (Kirin Holdings Co., Ltd), Kenicho Hara, Jo Kuwashima (Nihonsakari Co., Ltd), Yoshinobu Kaneko, Minetaka Sugiyama (Osaka University), Takayuki Bogaki, Akio Koda (Ozeki Corporation), Naoyuki Kobayashi (Sapporo Breweries Ltd), Takahiro Oura (Tokyo Institute of Technology), Hiroyuki Horiehuci, Ryoichi Fukuda and Jun-ichi Maruyama (University of Tokyo) for their contributions in gene annotation and Taku Kato, Naoki Maeya and Chie Oishi (National Research Institute of Brewing) for their contribution in validating the heterozygosity positions.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

References


15. Wei, W., McCusker, J.H., Hyman, R.W., et al. 2007, Genome sequencing and comparative analysis of
<table>
<thead>
<tr>
<th>No.</th>
<th>Reference</th>
</tr>
</thead>
</table>