DNA Research, 2018, 25(3), 297–306
doi: 10.1093/dnares/dsy002
Advance Access Publication Date: 5 February 2018
Full Paper

OXFORD

Full Paper

Complete genome sequence and analysis of the industrial *Saccharomyces cerevisiae* strain N85 used in Chinese rice wine production

Weiping Zhang[1,†], Yudong Li[1,2,†], Yiwang Chen[2], Sha Xu[1], Guocheng Du[1], Huidong Shi[3], Jingwen Zhou[1,*], and Jian Chen[1,*]

[1]School of Biotechnology, Jiangnan University, Wuxi, Jiangsu 214443, China, [2]Department of Bioengineering, School of Food Sciences and Biotechnology, Zhejiang Gongshang University, Hangzhou 310018, China, and [3]Georgia Cancer Center, Augusta University, Augusta, GA 30912, USA

*To whom correspondence should be addressed. Tel. +86 510 85918312. Fax. +86 510 85918309. Email: zhoujw1982@jiangnan.edu.cn (J.W.Z.); Tel. +86 510 85913660. Fax. +86 510 85918309. jchen@jiangnan.edu.cn (J.C.)
†These authors contributed equally to this work.

Edited by Dr Katsumi Isono

Received 24 May 2017; Editorial decision 5 January 2018; Accepted 10 January 2018

Abstract

Chinese rice wine is a popular traditional alcoholic beverage in China, while its brewing processes have rarely been explored. We herein report the first gapless, near-finished genome sequence of the yeast strain *Saccharomyces cerevisiae* N85 for Chinese rice wine production. Several assembly methods were used to integrate Pacific Bioscience (PacBio) and Illumina sequencing data to achieve high-quality genome sequencing of the strain. The genome encodes more than 6,000 predicted proteins, and 238 long non-coding RNAs, which are validated by RNA-sequencing data. Moreover, our annotation predicts 171 novel genes that are not present in the reference S288c genome. We also identified 65,902 single nucleotide polymorphisms and small indels, many of which are located within genic regions. Dozens of larger copy-number variations and translocations were detected, mainly enriched in the subtelomeres, suggesting these regions may be related to genomic evolution. This study will serve as a milestone in studying of Chinese rice wine and related beverages in China and in other countries. It will help to develop more scientific and modern fermentation processes of Chinese rice wine, and explore metabolism pathways of desired and harmful components in Chinese rice wine to improve its taste and nutritional value.

Key words: rice wine yeast, genome sequence, annotation, transcriptomics

1. Introduction

As early as 5,000 years ago, Chinese rice wine (Huangjiu) was being consumed by people as a fermented alcoholic beverage.[1] Chinese rice wine is made from sticky rice, which is different from other beverages made from malt or fruit. The streamed sticky rice was saccharified by raw wheat *koji* (a variety of *Aspergillus*) and fermented to generate alcohol by yeast strain. Nowadays, Chinese rice wine is still hugely popular in China due to its pleasant taste, and high nutritional and pharmacological value.[2] However, some harmful byproducts are generated during the fermentation and storage of

Chinese rice wine, such as ethyl carbamate (EC),[3] a genotoxic carcinogen of widespread occurrence in fermented food and beverages, with particularly high concentrations in stone fruit spirits.[4,5] In Chinese rice wine, EC is mainly generated from the reaction of ethanol and urea, and its concentration can reach 160 μg/kg.[3] As increasing attention is paid to food safety problems, some researchers have focused on the mechanism of EC formulation in Chinese rice wine, and attempts to reduce its concentration have been made.[3,6–8]

Over the past few decades, researchers have identified proteins responsible for the uptake and catabolism of poorly utilized nitrogen sources such as proline and urea, and found that their genes are repressed in *Saccharomyces cerevisiae* cells cultured in a nitrogen-rich environment. Conversely, in nitrogen-deficient conditions, their transcription is derepressed. This regulatory phenomenon has been termed nitrogen catabolite repression (NCR).[9–11] During the accumulation of EC in wine fermentation, NCR plays an important role in the accumulation urea, the major precursor of EC, by repressing the transcription of urea catabolism-related genes. NCR is reportedly controlled by complicated regulatory systems, such as global GATA family regulators,[12] the TOR pathway,[13] and the Ssy1p-Ptr3p-Ssy5p sensing pathway.[14] However, the exact regulatory mechanism of NCR is still unclear. Above all, a complete genome sequence is fundamental for understanding the genetic and regulatory systems of Chinese rice wine strains. The draft genome sequence of *S. cerevisiae* Chinese rice wine strain YHJ7 has been published,[15] which is closely related to the industrial N85 strain (Supplementary Fig. S1). However, there are still many gaps in the YHJ7 assembly, and detailed annotation of the genome is not available.

In the mid-2000s, the emergence of next-generation sequencing (NGS) platforms dramatically reduced the run time and cost (around US $1000 for the human genome) of genome sequencing, and increased the throughput to hundreds of Gbp per run[16] compared with first-generation sequencing platforms (Sanger sequencers). However, high/low G + C regions, tandem repeat regions, and interspersed repeat regions remain difficult to sequence using NGS platforms.[16] Furthermore, the limited read length from NGS platforms can prohibit sequencing completeness and the accuracy of analysis, resulting in assemblies that are incomplete and fragmented into several thousand contigs.[15,17] In 2011, PacBio RS II was developed as the first commercially available third-generation sequencer and it was marketed to address this issue. The system uses a novel and unique single molecule real-time (SMRT) technology which enables the generation of long reads (half of reads >20 kb, maximum read length >60 kb) and reduces the degree of bias (even coverage across regions of differing G + C content).[18] Because long reads can easily handle complex regions such as repeats, PacBio long reads have the potential to provide accurate and improved genome assemblies.[19]

In this study, a gapless, near-finished genome sequence of the *S. cerevisiae* Chinese rice wine N85 strain was achieved by combining second- and third-generation sequencing technologies. Complex approaches were performed to assemble and annotate the N85 genome, and over 6,000 proteins and 238 long non-coding RNAs (lncRNAs) were identified. Moreover, many genomic variants were identified that could be responsible for genomic evolution. The first complete genome sequence of Chinese rice wine brewing yeast, strain N85, is a milestone in studying Chinese rice wine, as well as other related beverages brewing. It will help to understand the evolutionary history of beverage brewing strains and improve the complicated brewing processes from a traditional and experiential way to a modern and scientific way. In addition, further detail analysis of the genome of strain N85 will provide more genetic information for improvement the general quality of Chinese rice wine.

2. Materials and methods

2.1. Yeast strain and growth conditions

Saccharomyces cerevisiae strain N85 (*MAT***a**/**α**) was the strain generally used for Chinese rice wine production and provided by Guyuelongshan wine company (Shaoxing, China), one of the biggest and oldest Chinese rice wine producer. The yeast strain was precultured on YPD plates (10 g/l yeast extract, 20 g/l peptone, 20 g/l glucose, 20 g/l agar) at 30°C for 24 h. A single colony was activated in YPD media (30°C, 200 rpm) then grown in sole nitrogen medium (1.7 g/l yeast nitrogen base, 20 g/l glucose, 10 mM each of glutamine, arginine, and urea). All cultivations were performed in shake flasks (200 rpm) at 30°C and growth was monitored by determining the optical density at 600 nm (OD_{600}).

2.2. DNA and RNA sequencing

Following a 1 h incubation under the conditions described above, genomic DNA and RNA was isolated as described previously in.[15] DNA/RNA quantity was determined using a Nanodrop 1000 (Thermo Scientific, MA, USA) and integrity was determined with an Agilent 2100 Bioanalyzer (Palo Alto, CA). For NGS of genomic DNA, the libraries were constructed using Nextera DNA sample preparation kit (Illumina, CA, USA). Next, libraries were sequenced using MiSeq reagent kit v3 (Illumina, CA, USA) on MiSeq platform, or TruSeq SBS kit v3-HS (Illumina, CA, USA) on HiSeq 2000 platform. For third generation DNA sequencing, the DNA sample was sequenced using the PacBio RS II technology with a C2 chemistry sequencing kits (Pacific Biosciences, Melon Park, CA). For RNA sequencing (RNA-Seq), all RNA samples were prepared as biological duplicates and subjected to removal of rRNA or Poly-A filtering before cDNA library generation. From these libraries, 100 bp paired-end or strand-specific reads were produced using an Illumina HiSeq 2000. Library preparation and sequencing of DNA and RNA on the Illumina platform was performed at BioMarker Inc. (Beijing, China), and the PacBio sequencing was performed at Shanghai Bohao Inc (Shanghai, China).

2.3. Genome assembly

The genome of strain N85 was assembled using two approaches, one using a classical reference assembly, and the other using a hybrid *de novo* assembly (Supplementary Fig. S2). The hybrid strategy for genome assembly was carried out in three distinct and contiguous steps: (i) *de novo* generation of contigs; (ii) ordering of the *de novo* contigs generated and their concatenation into supercontigs; (iii) closing of the remaining gaps using a *de novo* iterative approach. Specifically, regarding the reference-guided genome assembly, filtered short reads were used for direct alignment to the publically-available *S. cerevisiae* S288c (version Scer3) and YHJ7 genome sequences using the CLC Genomics Workbench version v8.0 (https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/ (19 January 2018, date last accessed)). Regarding the hybrid *de novo* assembly, *de novo* genome assembly was first performed using one or both of the PacBio and Illumina datasets by several methods, including the CLC Genomics Workbench version v8.0, A5-miseq v2,[20] SSPACE-LongRead v1.1,[21] SPAdes v3.5.0,[22] PBcR,[23] and hierarchical genome assembly process (HGAP) v2.0.[24] Second, the order and orientation of contigs from the *de novo* assembly were determined by

aligning contigs to the reference-guided assembly with a minimum identity value of 90% between the *de novo* contigs and the reference genome using CLC Genomics Workbench version v8.0 software. In this manner, most *de novo* contigs could be concatenated into super-contigs by overlapping regions.

2.4. Genome annotation

Gene features were annotated in the high-quality genome sequence using three different approaches: *ab initio*, evidence-, and homology-based predictions (Supplementary Fig. S3). For the *ab initio* prediction, AUGUSTUS v3.0.3[25] was employed with the predefined parameter set for the *S. cerevisiae* genome. For the evidence-based prediction of transcripts, two annotation methods were performed; one used TopHat2[26] to map RNA-Seq reads to the N85 genome with default parameters to identify the possible location of introns and exons[27] that were subsequently integrated in a hint-based run of AUGUSTUS; the other used TRINITY[28] to run mapping-free transcript assembly with the RNA-Seq data. For the homology-based prediction of transcripts, S288c open reading frames (ORFs) downloaded from *Saccharomyces* genome database (SGD), and exonerate[29] was employed to align the ORFs to the genome of strain N85.

To compare coding differences between N85 and S288c, local BLASTp searches were carried out using the amino acid sequences of the predicted ORFs as queries, and amino acid sequences of all ORFs in the S288c as database. The following threshold settings were used: *e*-value $< 1 \times 10^{-5}$, identity >80%, and an alignment length >50%.

2.5. Detection of isoforms and ncRNAs

For the detection of lncRNAs, Tophat2 was used for spliced read mapping with the following non-standard parameters: 'no-mixed', 'no-discordant', 'b2-very-sensitive', 'max-intron-length 10 100', and for strand-specific samples 'libtype fr-first strand'. The number of reads within exons and genes were calculated by using the Cufflinks pipeline,[26] used as gene expression values, and normalized using the number of reads per kilobase on exon regions per million mapped reads. Alternative splicing events (isoforms) were also predicted using the Cufflinks software,[26] and mapped reads were visualized with SpliceGrapher.[30] All identified isoforms and ncRNAs were manually checked using integrative genomics viewer.[31]

2.6. Detection of single nucleotide polymorphisms and analysis of copy-number compare with S288c

To estimate the genetic distance between the N85 and S288c reference strains, all Illumina reads were mapped to reference genomes using BWA-MEM.[32] Biallelic single nucleotide polymorphisms (SNPs) were called by the SAMtools v0.1.19 'mpileup' command and the VCFutils 'varFilter' command with '-D 200 -d 5'.

Per-position read counts were calculated from BAM files of mapped reads using the 'gennomecov' utility of BEDTools v2.15.0.[33] Reads were counted using a sliding window (1 kb) and used to find copy-number variations (CNVs). CNVs were identified using HMMcopy,[34] based on the ReadDepth method. To control false-positives, only CNVs with a length >2 kb were selected. The boundaries of CNVs were confirmed by visual inspection in the integrated genome viewer. We calculated the overlap of each gene with SNPs/CNVs from gene ontology (GO) gene sets using FungiFun2 (https://elbe.hki-jena.de/fungifun/fungifun.php (19 January 2018, date last accessed)). Bonferroni adjusted *P*-values and the false discovery rate were recorded.

2.7. Analysis of the genome variations between N85 and sake strains

The genome sequences of sake strains K7 and K11 were downloaded from SGD (http://www.yeastgenome.org/ (19 January 2018, date last accessed)). The differences between Chinese rice wine strain N85 and sake strains K7 and K11 were analysed at contig and sequencing read levels. First, the genome sequences of strain K7 and K11 were compared with that of strain N85 through local BLASTn with parameters as follows: -outfmt 17 -evalue 1e-5 -num_threads 8 -parse_deflines. The results in SAM format were visualized and manually investigated in integrative genomics viewer.[31] Second, all Illumina sequencing reads of the genome of strain N85, which could be aligned to N85 genome, were mapped into the scaffolds of strain K7 and K11 through CLC Genomics Workbench version v8.0. The resulting unmapped reads were assembled again to detect the unique regions of N85 genome.

2.8. Nucleotide sequence accession number

Data from the whole-genome shotgun project have been deposited at the EMBL database under the accession number LN907784-LN907800.

3. Results

3.1. Genome sequencing and assembly of the N85 genome

Genomic DNA from *S. cerevisiae* strain N85 was sequenced using PacBio RS and Illumina HiSeq/MiSeq platforms. The average read length and coverage value from both sequencing platforms are summarized in Supplementary Table S1. *De novo* assembly of PacBio reads was performed using the RS HGAP[24] assembly protocol version 3.3 in SMRT analysis version 2.2.0 (Pacific Biosciences) and resulted in 324 contigs (each with a length >500 bp, and an N50 value of 75,828 bp for all contigs). Subsequently, hybrid assembler PBcR[35] and SPAdes[22] were used to assemble the PacBio and Illumina reads together, and this generated 601 and 284 contigs, respectively. Additionally, we assembled the genome using only Illumina MiSeq and HiSeq reads with SPAdes, resulting in >1,000 contigs, and the N50 value was <10,000 bp when assembled with only paired-end MiSeq reads. The gap between contigs from each assembly method was further closed using SSPACE-LongReads.[21] The assembly processes and results are summarized in Table 1 and Supplementary Fig. S2.

De novo genome assemblies were further improved by combining the different hybrid assembly contigs and reference-guided assembly (see Section 2.3). Illumina reads were mapped to the S288c reference genome sequence using CLC genomic workbench to generate consensus sequences, which were used to place these contigs into 16 chromosomal and 1 mitochondrial sequences. The final assembly of the strain N85 genome showed high collinearity and structural conservation with that of the S288c reference (Fig. 1), excluding some repetitive or telomeric regions. Genome sequences that are derived from assembling reads potentially suffer from errors, especially around regions with repetitive sequences. Specifically, the ribosome DNA repeats in Chromosome XII appeared to be incorrectly assembled, and were manually adjusted based on the assembly of the closely related YHJ7 strain.[15] Finally, the gapless and near-finished genome sequence of *S. cerevisiae* strain N85 was obtained, which contains 16 chromosomal and one mitochondrial sequences and is 12.09 million bp in total (Supplementary Table S2).

Table 1. Summary of the *de novo* hybrid assembly results

Library type	No. of contigs[a]	Maximum contig size (bp)	N50 contig (bp)	Total length (bp)	Software
HiSeq	1,379	200,122	50,511	12,737,887	CLC genomic workbench
MiSeq	2,382	39,836	7,243	11,888,873	A5-miseq pipeline
PacBio	324	242,389	75,828	11,883,288	HGAP
Hybrid[b]	601	27,619	6,928	3,893,255	PBcR
Hybrid	284	862, 32	201,497	11,857,571	SPAdes
Close Gap	204	1,107,090	477,098	11,917,338	SSPACE-LongRead

[a]For each assembly, only contigs >500 bp in length were considered.
[b]Hybrid represents the combination of HiSeq, MiSeq, and PacBio datasets.

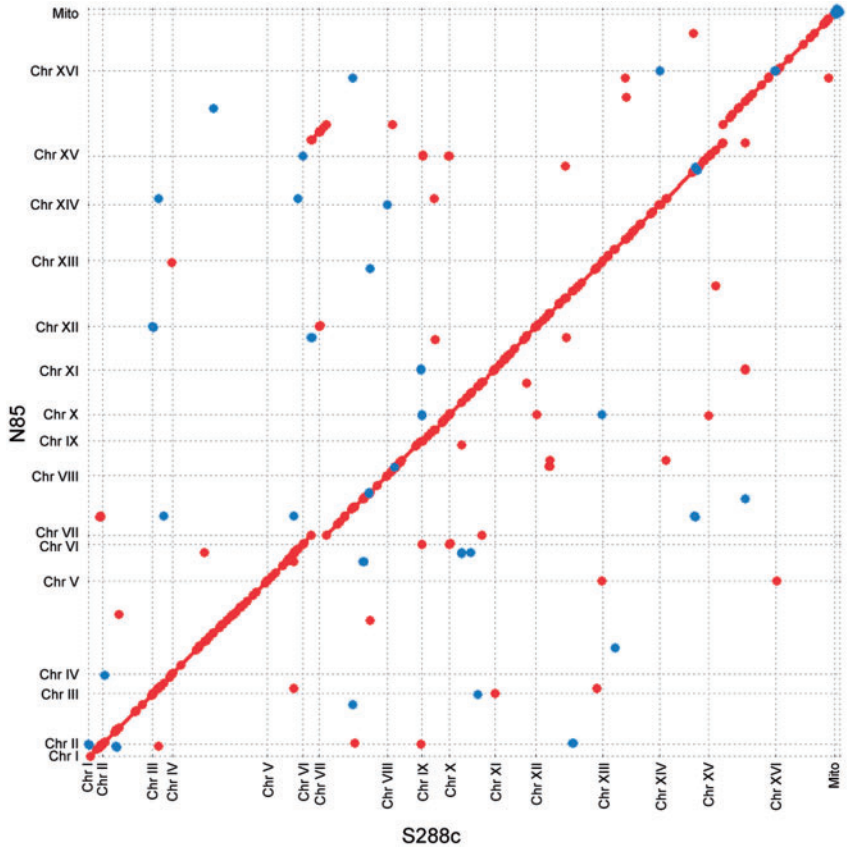

Figure 1. Dot plot of sequence similarity between the assembly scaffolds of the N85 and S288c strains. The majority of N85 assembly sequences are collinear with the chromosome of the reference S288c strain.

3.2. Multiple approaches used to annotate the N85 genome

In order to achieve a high-quality gene annotation of the *S. cerevisiae* strain N85 genome, four different annotation approaches were performed. First, all ORFs in the model *S. cerevisiae* S288c were downloaded from the SGD and aligned to the N85 genome using exonerate.[29] A total of 8,082 predicted ORFs were generated, and 6,236 of these were retained when the percentage coverage of the alignment cut-off was >50%. Second, *ab initio* gene prediction was performed with AUGUSTUS,[25] which generated 6,786 draft predicted ORFs. Thirdly, the total reads from RNA-Seq experiments were mapped onto the genome of *S. cerevisiae* strain N85 using TOPHAT2 to generate exon and intron 'hints' in AUGUSTUS.[25] With the help of these hints, local AUSGUSTUS genome prediction was carried out, resulting in 5,320 hints-based predicted ORFs. Lastly, total RNA-Seq reads were used to create the unbiased mapping-free transcriptome assembly using TRINITY,[36] which generated 15,371 draft transcripts. The results of the final three methods were filtered by amino acid length (>100), and 6,405, 5,339, and 10,487 ORFs were retained, respectively.

To compare coding differences between N85 and S288c, the amino acid sequences of the predicted ORFs were aligned to those of all S288c ORFs. 5,187, 4,968, and 5,267 S288c homologous genes were extracted from AUGUSTUS, AUGUSTUS-Tophat, and TRINITY predictions, respectively (Supplementary Table S3). Combining these S288c homologous genes and those from the exonerate prediction, 6,464 S288c homologous genes were identified in strain N85. Among them, 4,694 S288c homologous genes were

present in all four predictions (Fig. 2A). Moreover, we identified 111, 86, and 82 non-S288c genes that were missing from the S288c genome annotation using AUGUSTUS, AUGUSTUS-Tophat, and TRINITY predictions, respectively (Supplementary Table S3). In total, 171 non-S288c genes were identified by merging all these non-S288c genes, and 17 non-S288c genes were common to all three predictions (Fig. 2B). Consistent with a previous study, we detected a large region (24 kb) in Chromosome XIV that includes three non-S288c genes (g5159, g5160, and g5161) that are only present in Asian yeast strains (Fig. 2C). In addition, we also detected some N85-specific regions that are not present in YHJ7, including the region in Chromosome XII that includes two novel genes g4241 and g4242 (Fig. 2C). The expression of these novel non-S288c genes was validated by RNA-Seq data, but their function requires further investigation.

3.3. Transcription of lncRNAs and isoforms

RNA-Seq was used to determine lncRNAs in *S. cerevisiae* strain N85. Approximately, 20 million poly-adenylated RNA-Seq reads and 2 billion strand-specific ribosome-removed RNA-Seq reads were obtained, which allowed us to confirm the orientation of transcripts and predict anti-sense transcripts. After read mapping and transcript assembly, we classified all expressed transcripts longer than 200 nucleotides into coding genes and lncRNAs. Using these sequencing datasets, we detected 238 lncRNAs, most of which are novel lncRNAs not annotated in the databases (Fig. 3A). Consistent with previous studies, lncRNAs were expressed at significantly lower levels than coding genes (Fig. 3B, Wilcoxon test, $P < 10^{-5}$). However, the function of these novel lncRNAs requires further investigation in the future. Besides lncRNAs, 619 transcript isoforms of N85 genes were also annotated using transcriptome analysis (Fig. 3A), and the expression levels of isoforms were slightly higher than those of genes without alternative splicing (Fig. 3B).

3.4. Genome variations in strain N85 compared with the model strain S288c

To examine genetic variation in N85, the Illumina short reads were mapped to the S288c reference genome, which identified 57,278 SNPs and 8,624 indels in N85 (Table 2). As expected, ~99% of these SNPs have been observed in the closely related YHJ7 strain. Although N85 is almost homozygous, 2,131 heterozygous sites were found in the diploid sequenced N85 strain (Table 2). The SNP distribution was not random, and approximately one-third of all detected SNPs were intergenic (Table 2), even though only about 25% of the *S. cerevisiae* genome is noncoding. More than 30% of SNPs detected in coding regions are nonsynonymous, resulting in changes to the encoded protein sequence. Interestingly, nonsynonymous mutations are more frequent in homozygous SNPs than in heterozygous SNPs (24 vs. 22%, respectively; $P < 0.01$, Chi-squared test). Moreover, several genes have gained or lost stop codons (Table 2).

The nonsynonymous to synonymous substitution rate *(Ka/Ks)* relative to the S288c strain was assessed to identify fast-evolving genes. Genes with *Ka/Ks* values significantly >1 were classified as under positive selection. We identified 232 genes with *Ka/Ks* values > 1 (Supplementary Table S4). GO term analysis revealed that these genes were mostly associated with transcriptional control. Several stress-responsive transcription factors appear to have evolved rapidly in N85, including the GATA zinc finger protein *GZF3*, which regulates nitrogen catabolic gene expression, and two genes (*HKR1*, *Ka/Ks* = 1.94; *MSB2*, *Ka/Ks* = 1.22) involved in the HOG pathway, which have been shown to play a key role in the adaption of Chinese rice wine strains.[37]

Sequencing of the N85 strain in its natural ploidy state allowed analysis of gross chromosomal rearrangements and aneuploidies. Based on the sequencing read depth, there were no gains or losses of whole chromosomes in the N85 genome (Supplementary Fig. S4), discounting polyploidy or aneuploidy. A total of 76 amplification

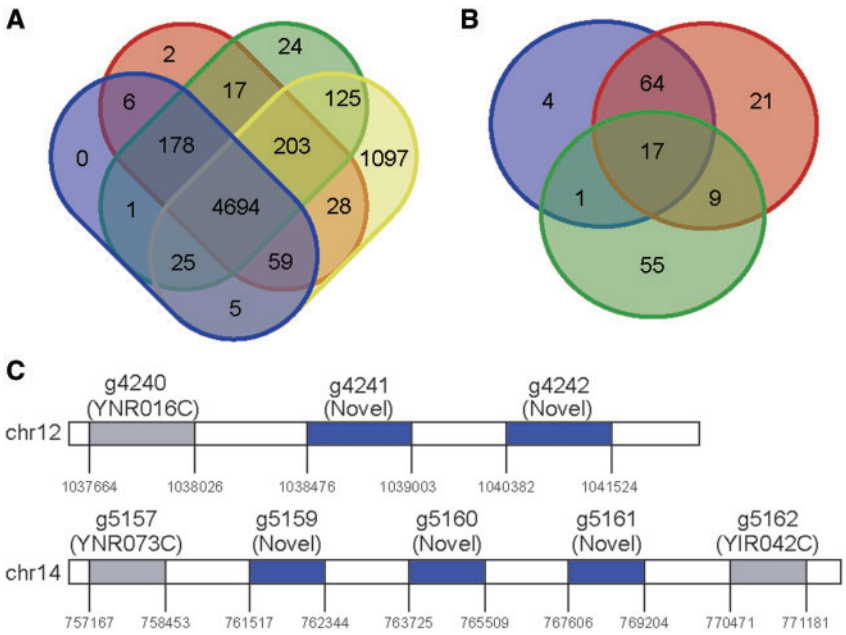

Figure 2. Annotation of the *S. cerevisiae* N85 genome. Analysis of S288c homologous and non-S288c genes in *S. cerevisiae* N85 through different approaches. (A) Number of S288c homologous genes identified using exonerate (yellow), AUGUSTUS (red), AUGUSTUS-Tophat (blue), and TRINITY (green). (B) Number of non-S288c genes identified using AUGUSTUS (red), AUGUSTUS-Tophat (blue), and TRINITY (green). (C) Genomic architecture of non-S288c genes in N85. Gene locations are shown below each gene box. Color figures available in online version.

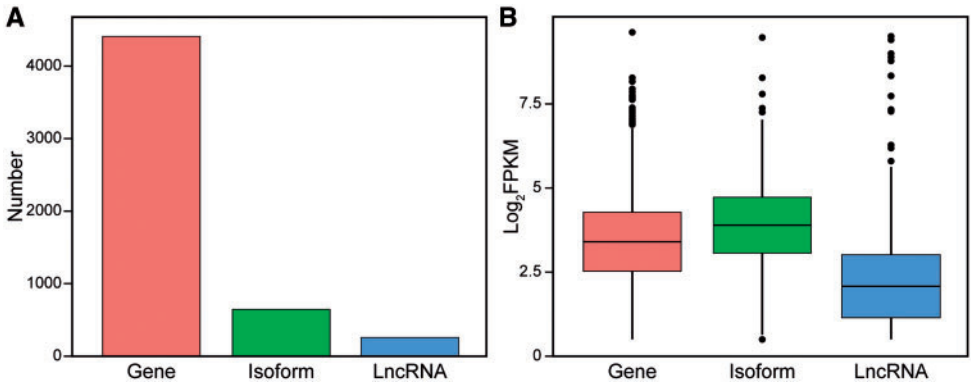

Figure 3. General characteristics of coding and non-coding transcripts. (A) The number of different transcript variants in *S. cerevisiae* N85. (B) Box-plots of transcript expression levels in log2 (FPKM) units. FPKM, fragments per kilobase of exon per million reads mapped.

Table 2. Genetic variations identified in the *S. cerevisiae* strain N85

	N85			YHJ7-homologues		
	Total	Hom[a]	Het	Total	Hom	Het
Exonic	38,478	37,564	914	214	73	141
Synonymous	23,717	23,350	367	75	21	54
Nonsynonymous	12,380	12,121	259	72	26	46
Frameshift	274	243	31	6	2	4
Nonframeshift	1,996	1,748	248	55	20	35
Stop gain or loss	89	82	7	6	4	2
Intronic	428	424	4	5	3	2
Intergenic	26,996	25,783	1,213	1,175	852	323

[a]Hom: homozygous; Het: heterozygous. YHJ7-homologs represents those SNPs that shared with YHJ7 were filtered from the total N85 SNPs.

and deletion events were detected in N85, covering 1.57 Mb of the genome (Fig. 4). The size of these regions ranges from 1 to 18 kb, and most were detected in subtelomeric regions. GO enrichment analysis of CNV regions revealed that genes involved in cellular responses to nitrogen starvation, asparagine metabolic processes, and cellular aldehyde metabolic processes are most heavily influenced by CNVs.

3.5. The genome differences between Chinese rice wine strain N85 and sake strains

In general, the genome of Chinese rice wine strain is quite similar with those of sake strains after manually reviewing the BLASTn results between strain N85 and K7 as well as K11. However, some genomic variations may explain the differences between Chinese rice wine and sake to some content. Four hundred indels were identified in the genome of strain N85 compared with both sake yeasts K7 and K11. Among them, 116 indels were found in gene coding regions. Some affected genes were found to involve in the production of organic acids (*KGD2*, *HSP31*, *BNA3*), amino acids catabolism (*STR2*, *HOM6*, *SPE2*, *CYS3*, *ARG2*). Previous study suggested that strain with deficient α-ketoglutarate dehydrogenase (*KGD2*) produced fewer ethanol during wine brewing.[38] The low ethanol content was compensated by an increase of organic acids, such citrate succinate, fumarate, and malate.[38] *BNA3* encodes a putative carbon-sulphur lyase responsible for volatile-thiol release. Deletion of *BNA3* reduced the release of 4-mercapto-4-methylpentan-2-one, which makes an important contribution to the aroma of wine, in both laboratory and wine yeast background.[39] Besides, the mutation of *CYS3*, which also encodes a putative carbon-sulphur lyase, led to over production of another sulphur compound, methyltetrathiophen-3-one, which was previously shown to contribute to wine aroma.[39] *ARG2* locates in arginine biosynthesis pathway and encodes acetylglutamate synthase, which catalyses the first step in the biosynthesis of the arginine precursor, ornithine.

In addition to indels, 25 regions in N85 genome were found to be absent either in the genome of strain K7 or that of K11, through mapping unmapped reads back to N85 genome. Five of these regions were identified to be absent in genomes of both sake strains. All of them are coding regions in the genome of strain N85, which encode the alpha-glucoside permease (*MPH3*), the sorbitol dehydrogenase (*SOR1*), a component of vacuolar cation channel (*YVC1*), and two putative proteins with unknown function (YGL262W and YGL263W). *SOR1* and *MPH3* locate on the subtelomeric paralogous blocks of Chromosome X. However, the effects of the gain or loss of these genes on Chinese rice wine and sake are still unclear.

3.6. Genome browser for N85

To visualize gene sequences, annotated genes, and genetic variants in the genome of *S. cerevisiae* strain N85, a JBrowse-based genome browser was developed and deposited on the website (http://www.ligene.cn/hygd (19 January 2018, date last accessed)). As shown in Fig. 5, the basic genome browser functionality can provide genome annotation views via an overhead bar that offers a visual indication of the chromosome position. Although only gene annotation and mutation information is currently available, further large-scale datasets will be integrated in the future. We also aim to implement additional analysis tools for BLAST searching, primer design, and sequence alignment, to help the scientific community to use the developed genome resources.

4. Discussion

Chinese rice wine is a traditional fermented alcoholic beverage with many health benefits.[40] Recently, the annual consumption of Chinese rice wine has been steadily increasing, and now is more than 2 million kilolitres.[41,42] However, the complete genome sequence of the Chinese rice wine strain used for fermentation has not been reported, which makes it difficult to further improve the quality of this popular product. The purpose of the current study was to genetically

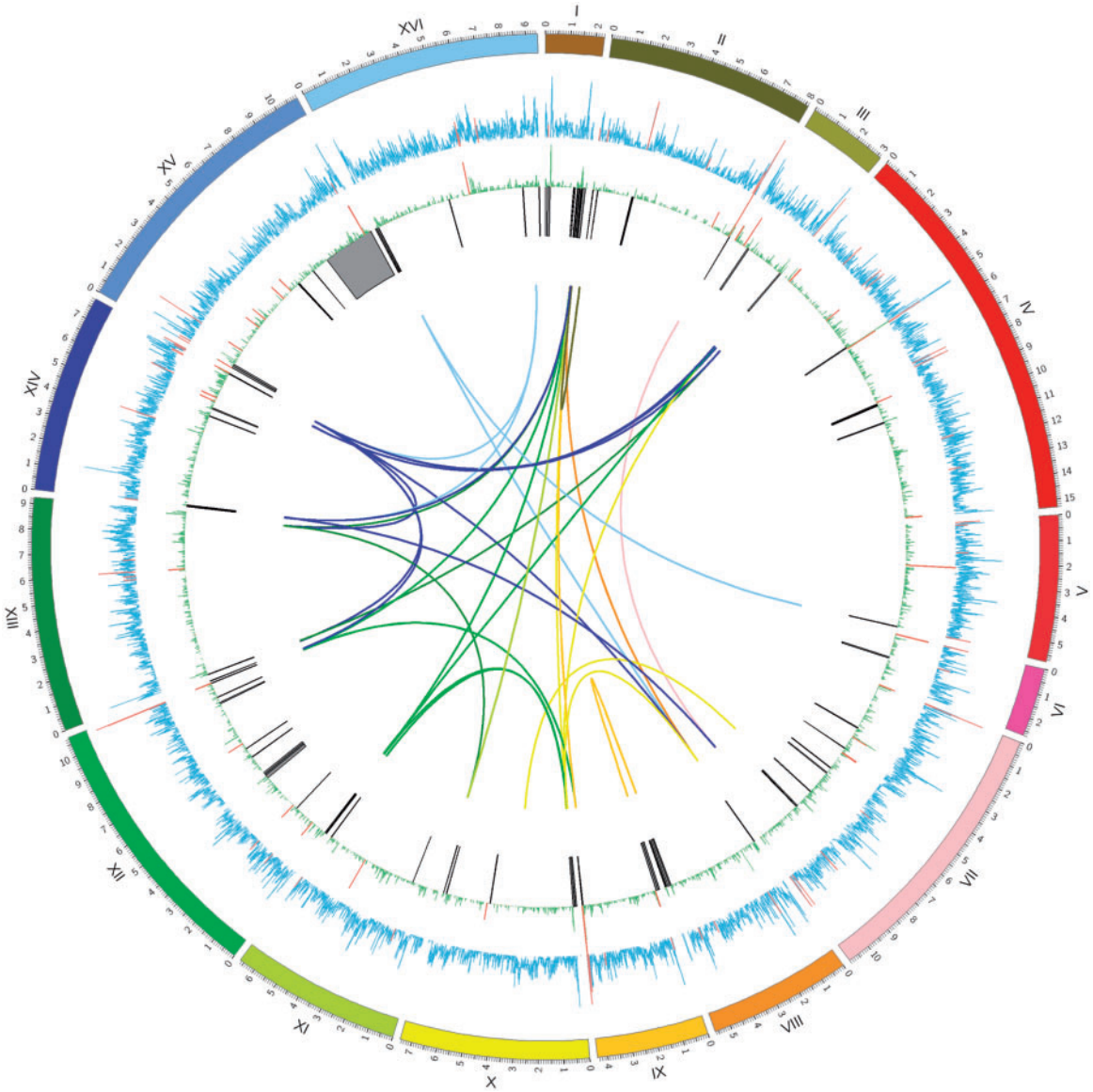

Figure 4. Genetic variation in the *S. cerevisiae* N85 genome. The first and second circles represent SNPs and INDELs relative to the *S. cerevisiae* S288c reference genome, in which the specific variation in N85 relative to YHJ7 is highlighted in red. The third and fourth cycles represent larger duplication/deletion or translocation events relative to YHJ7. Most of the structural rearrangements are localized in subtelomeric regions. Color figures available in online version.

characterize the *S. cerevisiae* strain N85 by sequencing the genome using both PacBio and Illumina platforms. Assembly was achieved by combining the long but relatively low-quality PacBio reads[43] with the short but higher quality Illumina reads using a complex approach, and a high-quality genome of the N85 strain was generated. The assembled gapless and near-complete genome is equivalent in length to that of the model S288c strain. Due to the high integrity of the assembled genome, a larger number of protein coding genes were annotated through multiple annotation approaches compared with the gapped genome of the YHJ7 strain.[15] This comprehensive annotation also identified additional non-S288c genes,[15] which are important for functional and evolutionary analysis. Furthermore, the annotated genes in the N85-specific regions may be to explain the different evolutionary routes of Chinese rice wine strains.

In addition, genome comparison between Chinese strain and Japanese strains revealed a series of variations in coding regions, which encodes genes involved in the production of organic acids and the catabolism of amino acids. These variations might explain the difference of flavour and nutritional value between Chinese rice wine and sake. Notably, some variants were identified in the coding sequence of *ARG2*, who encodes acetylglutamate synthase in arginine biosynthesis pathway. Arginine is the precursor of urea, which is considered to be the major precursor of EC. The mutation in *ARG2* may lead to the higher concentration of EC in Chinese rice wine than that in sake. Besides, previous study revealed that two subtelomeric paralogous blocks in the genome of strain S288c, containing *HXT15-SOR2-MPH2* on Chromosome IV and *HXT16-SOR1-MPH3* on Chromosome X, were lost in the genome of strain K7.[44] However, a part of the second block was identified in the genome of strain N85, containing *SOR1-MPH3*. These differences among the genome of laboratory, Chinese rice wine, and sake strains may suggest strains underwent different evolutionary tracks to cope with each environment.

Figure 5. Screenshot of the homepage of the Huangjiu Yeast Genome Database and genome browser displaying Chromosome 1 of *S. cerevisiae* N85. Gene regions are represented as a horizontal box, and genetic variants are represented as blue dots. Color figures available in online version.

LncRNAs and alternative splicing are two important elements of transcription. LncRNAs are claimed to play an important role in the regulation of gene expression at transcriptional, post-transcriptional, and translational levels.[45,46] Moreover, alternative splicing is a major contributor that determines the environmental fitness of an organism. Recently, various isoforms of the transcription factor *GAT1* were discovered in *S. cerevisiae* that are involved in nitrogen catabolite repression regulation.[47] In this work, numerous isoforms and novel lncRNAs were identified, enriching our existing knowledge in this area. Furthermore, future analysis of their functions could deepen our understanding of Chinese rice wine fermentation. It should be emphasized that current methods for annotating isoforms are not reliable; some annotated isoforms may be false-positives, and the accuracy of the results will be improved when the sequencing technology evolves and generates longer sequences.

The presence of highly conserved SNP sites shared in the N85 and YHJ7 genomes but not in the S288c genome indicated that the two strains originated from a common ancestor that diverged from the ancestor of the model S288c strain, and underwent adaptation during Chinese rice wine fermentation. However, the enrichment of SNPs in intergenic regions in the N85 genome might be due to reduced functional constraints in intergenic sequences during the evolutionary history of this strain. The enrichment of nonsynonymous mutations in homozygous genes suggests the function of heterozygous genes is much more stable, and homozygous SNPs appear to make a greater contribution to functional adaption in the N85 strain. During the fermentation of Chinese rice wine, yeast suffer high concentrations of sugars resulting from the digestion of rice starches in the fermentation mash. The enrichment of nonsynonymous SNPs in the HOG pathway is indicative of adaptive evolution of the strain over its long history in Chinese rice wine fermentation. Notably, this phenomenon was also observed in the genome of another Chinese rice wine strain, namely YJH7.[15] Similar enrichment was also identified in *GZF3*, one of four global transcriptional regulators of nitrogen catabolite repression, which may be responsible for the accumulation of EC in Chinese rice wine. Variations in genome

structure, such as polyploidy, aneuploidy and copy number, have repeatedly been associated with domestication and adaptation to specific niches in experimentally evolved microbes.[48] Moreover, due to their plastic and dynamic nature, loss or gain of genes in subtelomeric regions occurs more frequently than in other regions, which may accelerate adaptive evolution.[49] Analysis of structural variation suggests that CNVs in the N85 genome may underlie niche adaptation.

5. Conclusion

We combined sequencing data from Illumina and PacBio sequencing platforms to generate the first gapless and near-complete genome assembly of the *S. cerevisiae* strain N85 industrial strain used in Chinese rice wine brewing. Our study revealed many genes and genetic variations that may help the strain to cope with the high glucose and ethanol concentrations in the fermentation environment. Industrial rice wine producers, even other beverages producers, will likely benefit from having access to a complete N85 genome to improve their production progresses and products quality. In addition, our findings provide a rich genetic resource for the *S. cerevisiae* fundamental and applied research communities.

Funding

This work was supported by The National Key Research and Development Programme of China (2017YFC1600403), the National Natural Science Foundation of China (31670095, 31770097), the Key Research and Development Programme of Jiangsu Province (BE2016689), the Fundamental Research Funds for the Central Universities (JUSRP51701A), the Six Talent Peaks Project in Jiangsu Province (2015-JY-005), the Distinguished Professor Project of Jiangsu Province, and a grant from Zhejiang Provincial Natural Science Foundation of China (LY14C060001) to Y. Li.

Conflict of interest

None declared.

Accession numbers

EMBL:LN907784, LN907785, LN907786, LN907787, LN907788, LN907789, LN907790, LN907791, LN907792, LN907793, LN907794, LN907795, LN907796, LN907797, LN907798, LN907799, LN907800.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. McGovern, P. E., Zhang, J., Tang, J., et al. 2004, Fermented beverages of pre- and proto-historic China, *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 17593–8.
2. Guangfa, X. 2008, Functional constituents and health function of Chinese rice wine, *Liquor Making*, **35**, 14–6.
3. Zhao, X., Du, G., Zou, H., Fu, J., Zhou, J. and Chen, J. 2013, Progress in preventing the accumulation of ethyl carbamate in alcoholic beverages, *Trends Food Sci. Tech.*, **32**, 97–107.
4. Wu, D., Li, X., Lu, J., Chen, J., Zhang, L., Xie, G. and Poeggeler, S. 2016, Constitutive expression of the *DUR1, 2* gene in an industrial yeast strain to minimize ethyl carbamate production during Chinese rice wine fermentation, *FEMS Microbiol. Lett.*, **363**, fnv214.
5. Dahabieh, M. S., Husnik, J. I. and Van Vuuren, H. J. 2010, Functional enhancement of Sake yeast strains to minimize the production of ethyl carbamate in Sake wine, *J. Appl. Microbiol.* , **109**, 963–73.
6. Zhao, S., Zhao, X., Zou, H., Fu, J., Du, G., Zhou, J. and Chen, J. 2014, Comparative proteomic analysis of Saccharomyces cerevisiae under different nitrogen sources, *J. Proteomics*, **101**, 102–12.
7. Zhao, X., Zou, H., Fu, J., Chen, J., Zhou, J. and Du, G. 2013, Nitrogen regulation involved in the accumulation of urea in *Saccharomyces cerevisiae*, *Yeast*, **30**, 437–47.
8. Zhao, X., Zou, H., Fu, J., Zhou, J., Du, G. and Chen, J. 2014, Metabolic engineering of the regulators in nitrogen catabolite repression to reduce the production of ethyl carbamate in a model rice wine system, *Appl. Environ. Microbiol.*, **80**, 392–8.
9. Wiame, J. M., Grenson, M. and Arst, H. N. Jr. 1985, Nitrogen catabolite repression in yeasts and filamentous fungi, *Adv. Microb. Physiol.*, **26**, 1–88.
10. Beltran, G., Novo, M., Rozes, N., Mas, A. and Guillamon, J. M. 2004, Nitrogen catabolite repression in Saccharomyces cerevisiae during wine fermentations, *FEMS Yeast Res.*, **4**, 625–32.
11. Tate, J. J., Buford, D., Rai, R. and Cooper, T. G. 2017, General amino acid control and 14-3-3 proteins Bmh1/2 are required for nitrogen catabolite repression-sensitive regulation of Gln3 and Gat1 localization, *Genetics*, **205**, 633–55.
12. Zhao, X., Zou, H., Chen, J., Du, G. and Zhou, J. 2016, The modification of Gat1p in nitrogen catabolite repression to enhance non-preferred nitrogen utilization in *Saccharomyces cerevisiae*, *Sci. Rep.*, **6**, 21603.
13. Orlova, M., Kanter, E., Krakovich, D. and Kuchin, S. 2006, Nitrogen availability and TOR regulate the Snf1 protein kinase in *Saccharomyces cerevisiae*, *Eukaryot. Cell*, **5**, 1831–7.
14. Ljungdahl, P. O. 2009, Amino-acid-induced signalling via the SPS-sensing pathway in yeast, *Biochem. Soc. Trans.*, **37**, 242–7.
15. Li, Y., Zhang, W., Zheng, D., et al. 2014, Genomic evolution of Saccharomyces cerevisiae under Chinese rice wine fermentation, *Genome Biol. Evol.*, **6**, 2516–26.
16. Goodwin, S., McPherson, J. D. and McCombie, W. R. 2016, Coming of age: ten years of next-generation sequencing technologies, *Nat. Rev. Genet.*, **17**, 333–51.
17. Gnerre, S., Maccallum, I., Przybylski, D., et al. 2011, High-quality draft assemblies of mammalian genomes from massively parallel sequence data, *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 1513–8.
18. Nakano, K., Shiroma, A., Shimoji, M., et al. 2017, Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area, *Hum. Cell.*, **30**, 149–61.
19. Kamada, M., Hase, S., Sato, K., Toyoda, A., Fujiyama, A., Sakakibara, Y. and Schacherer, J. 2014, Whole genome complete resequencing of *Bacillus subtilis* natto by combining long reads with high-quality short reads, *PLoS One*, **9**, e109999.
20. Coil, D., Jospin, G. and Darling, A. E. 2015, A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data, *Bioinformatics*, **31**, 587–9.
21. Boetzer, M. and Pirovano, W. 2014, SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information, *BMC Bioinformatics*, **15**, 211.
22. Bankevich, A., Nurk, S., Antipov, D., et al. 2012, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J. Comput. Biol.*, **19**, 455–77.
23. Berlin, K., Koren, S., Chin, C. S., Drake, J. P., Landolin, J. M. and Phillippy, A. M. 2015, Assembling large genomes with single-molecule sequencing and locality-sensitive hashing, *Nat. Biotechnol.*, **33**, 623–30.
24. Chin, C. S., Alexander, D. H., Marks, P., et al. 2013, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, *Nat. Methods.* , **10**, 563–9.
25. Stanke, M., Diekhans, M., Baertsch, R. and Haussler, D. 2008, Using native and syntenically mapped cDNA alignments to improve de novo gene finding, *Bioinformatics*, **24**, 637–44.
26. Trapnell, C., Roberts, A., Goff, L., et al. 2012, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nat. Protoc.*, **7**, 562–78.

27. Hoff, K. 2014, Incorporating RNA-Seq into AUGUSTUS with TOPHAT. Avialable at: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=IncorporatingRNAseq.Tophat (19 January 2018, date last accessed).
28. Grabherr, M. G., Haas, B. J., Yassour, M., et al. 2011, Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data, *Nat. Biotechnol*., **29**, 644–52.
29. Slater, G. S. and Birney, E. 2005, Automated generation of heuristics for biological sequence comparison, *BMC Bioinformatics*., **6**, 31.
30. Rogers, M. F., Thomas, J., Reddy, A. S. and Ben-Hur, A. 2012, SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data, *Genome Biol*., **13**, R4.
31. Thorvaldsdottir, H., Robinson, J. T. and Mesirov, J. P. 2013, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, *Brief Bioinform*., **14**, 178–92.
32. Li, H. 2013, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv: 1303.3997*.
33. Quinlan, A. R. and Hall, I. M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841–2.
34. Ha, G., Roth, A., Lai, D., et al. 2012, Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer, *Genome Res*., **22**, 1995–2007.
35. Koren, S., Schatz, M. C., Walenz, B. P., et al. 2012, Hybrid error correction and de novo assembly of single-molecule sequencing reads, *Nat. Biotechnol*., **30**, 693–700.
36. Garber, M., Grabherr, M. G., Guttman, M. and Trapnell, C. 2011, Computational methods for transcriptome annotation and quantification using RNA-seq, *Nat. Methods*, **8**, 469–77.
37. Li, Y., Chen, W., Shi, Y. and Liang, X. 2013, Molecular cloning and evolutionary analysis of the HOG-signaling pathway genes from *Saccharomyces cerevisiae* rice wine isolates, *Biochem. Genet*., **51**, 296–305.
38. Selecký, R., Šmogrovičová, D. and Sulo, P. 2008, Beer with reduced ethanol content produced using *Saccharomyces cerevisiae* yeasts deficient in various tricarboxylic acid cycle enzymes, *J. Inst. Brew*., **114**, 97–101.
39. Howell, K. S., Klein, M., Swiegers, J. H., et al. 2005, Genetic determinants of volatile-thiol release by *Saccharomyces cerevisiae* during wine fermentation, *Appl. Environ. Microbiol*., **71**, 5420–6.
40. Que, F., Mao, L., Zhu, C. and Xie, G. 2006, Antioxidant properties of Chinese yellow wine, its concentrate and volatiles, *LWT Food Sci. Technol*., **39**, 111–7.
41. Wei, X. L., Liu, S. P., Yu, J. S., et al. 2017, Innovation Chinese rice wine brewing technology by bi-acidification to exclude rice soaking process, *J. Biosci. Bioeng*., **123**, 460–5.
42. Fang, R. S., Dong, Y. C., Li, H. J. and Chen, Q. H. 2015, Ethyl carbamate formation regulated by *Saccharomyces cerevisiae* ZJU in the processing of Chinese yellow rice wine, *Int. J. Food Sci. Technol*., **50**, 626–32.
43. Au, K. F., Underwood, J. G., Lee, L. and Wong, W. H. 2012, Improving PacBio long read accuracy by short read alignment, *PLoS One*., **7**, e46679.
44. Akao, T., Yashiro, I., Hosoyama, A., et al. 2011, Whole-genome sequencing of sake yeast Saccharomyces cerevisiae Kyokai no. 7, *DNA Res*., **18**, 423–34.
45. Rinn, J. L. and Chang, H. Y. 2012, Genome regulation by long noncoding RNAs, *Annu. Rev. Biochem*., **81**, 145–66.
46. Thum, T. and Condorelli, G. 2015, Long noncoding RNAs and microRNAs in cardiovascular pathophysiology, *Circ. Res*., **116**, 751–62.
47. Rai, R., Tate, J. J., Georis, I., Dubois, E. and Cooper, T. G. 2014, Constitutive and nitrogen catabolite repression-sensitive production of Gat1 isoforms, *J. Biol. Chem*., **289**, 2918–33.
48. Gallone, B., Steensels, J., Prahl, T., et al. 2016, Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts, *Cell*, **166**, 1397.
49. Dunn, B., Richter, C., Kvitek, D. J., Pugh, T. and Sherlock, G. 2012, Analysis of the Saccharomyces cerevisiae pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments, *Genome Res*., **22**, 908–24.