




Genetic characterization and modification of a bioethanol-producing yeast strain

Ke Zhang¹ · Ya-Nan Di² · Lei Qi² · Yang Sui² · Ting-Yu Wang² · Li Fan² · Zhen-Mei Lv¹ · Xue-Chang Wu¹ · Pin-Mei Wang² · Dao-Qiong Zheng² 

Received: 23 October 2017 / Revised: 16 December 2017 / Accepted: 18 December 2017
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

Yeast *Saccharomyces cerevisiae* strains isolated from different sources generally show extensive genetic and phenotypic diversity. Understanding how genomic variations influence phenotypes is important for developing strategies with improved economic traits. The diploid *S. cerevisiae* strain NY1308 is used for cellulosic bioethanol production. Whole genome sequencing identified an extensive amount of single nucleotide variations and small insertions/deletions in the genome of NY1308 compared with the S288c genome. Gene annotation of the assembled NY1308 genome showed that 43 unique genes are absent in the S288c genome. Phylogenetic analysis suggested most of the unique genes were obtained through horizontal gene transfer from other species. RNA-Seq revealed that some unique genes were not functional in NY1308 due to unidentified intron sequences. During bioethanol fermentation, NY1308 tends to flocculate when certain inhibitors (derived from the pretreatment of cellulosic feedstock) are present in the fermentation medium. qRT-PCR and genetic manipulation confirmed that the novel gene, *NYn43*, contributed to the flocculation ability of NY1308. Deletion of *NYn43* resulted in a faster fermentation rate for NY1308. This work disclosed the genetic characterization of a bioethanol-producing *S. cerevisiae* strain and provided a useful paradigm showing how the genetic diversity of the yeast population would facilitate the personalized development of desirable traits.

Keywords Cellulosic ethanol · *Saccharomyces cerevisiae* · Whole genome sequencing · RNA-Seq · Unique genes

Introduction

Yeast *Saccharomyces cerevisiae* is one of the most useful microorganisms in biological fermentation fields, such as winemaking, baking, brewing, and bioethanol production. The bioconversion of cellulosic biomass to ethanol is a promising technology to address fossil fuels shortages and greenhouse gas over-emission (Koutinas et al. 2016). To release the

sugars that are trapped inside the cross-linking structure of cellulose, pretreatment processes (such steam explosion and dilute acid treatment) are always required prior to the enzymatic hydrolysis (Jönsson et al. 2013; Silveira et al. 2015). Unfortunately, pretreatment processes give rise to certain inhibitors (including acetic acid, furan derivatives, and phenolic compounds) that would greatly inhibit the viability of yeast cells during ethanol fermentation (Jönsson et al. 2013; Sindhu et al. 2016; Zheng et al. 2017).

In the past two decades, a great amount of functional genomic studies on *S. cerevisiae* using the reference genome of strain S288c (the first sequenced eukaryote) have greatly enriched our knowledge of how yeast cells respond to external stimuli (Kitichantaropas et al. 2016). However, *S. cerevisiae* strains isolated from different sources generally show extensive phenotypic diversity (Fay et al. 2004; Strobe et al. 2015). It was shown that certain industrial strains are more tolerant to specific environments stresses compared with S288c-derived laboratory strains (Zheng et al. 2012, 2016). Using high-throughput sequencing technology, the genetic diversity

Ke Zhang and Ya-Nan Di contributed equally to this work.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00253-017-8727-1>) contains supplementary material, which is available to authorized users.

✉ Dao-Qiong Zheng
zhengdaoqiong@zju.edu.cn

¹ Institute of Microbiology, College of Life Sciences, Zhejiang University, Hangzhou, Zhejiang Province 310058, China

² Ocean College, Zhejiang University, Zhoushan, Zhejiang Province 316021, China

between S288c and other strains has been widely determined (Borneman et al. 2011; Coi et al. 2017; Nijkamp et al. 2012; Strope et al. 2015; Zheng et al. 2012; Zhu et al. 2016). Additionally, great efforts have been devoted to exploring how genomic variations (point mutations, chromosomal rearrangements, and novel ORFs) affect stress tolerance and ethanol fermentation performance among different *S. cerevisiae* strains. For example, a non-sense mutation in the gene *AQY2* would result in higher osmotic resistance in many non-S288c strains (Will et al. 2010). The unique genes *AWA1* and *BIO6* in sake strains are responsible for foam formation in sake mash and biotin synthesis, respectively (Akao et al. 2011; Hall and Dietrich 2007). The introgressed *PDR5* gene found in certain clinical strains contributed to tolerance to cycloheximide and ketoconazole (Strope et al. 2015). The chromosomal translocation mediated by the genes *ECM34* and *SSU1* was positively selected for in wine strains, because this event leads to more sulfite resistance (Perez-Ortin et al. 2002). Additionally, our previous studies suggest that large-scale chromosomal rearrangements can cause “global aneuploidy stress” in industrial *S. cerevisiae* strains, as well as specific effects due to the copy number variation of certain functional genes (such as *CUP1*, *SOD2*, and *ERG11*) (Zhang et al. 2015, 2016; Zheng et al. 2014).

In the present study, the genetic characteristics of the bioethanol-producing *S. cerevisiae* strain NY1308 were revealed by whole genome sequencing (WGS). The gene annotation and transcription landscape determination predicted dozens of unique genes in NY1308 that are absent in the S288c genome. One unique gene was responsible for the flocculation phenotype of NY1308 induced by inhibitors from pretreated hydrolysates. Finally, based on the obtained information, effective personalized genomic modification strategies were developed to improve the bioethanol fermentation performances of NY1308.

Materials and methods

Yeast strains and culture conditions

Industrial *S. cerevisiae* strain NY1308 (CICC 1308) has been used for ethanol fermentation for more than 50 years in China. Recently, this strain has also been applied to the cellulosic bioethanol production at a pilot scale by the Henan Tianguan Group Co., Ltd. (China). The S288c-isogenic strain BY4741 (*MATa*; *his3ΔIleu2Δmet15Δoura3Δ0*) was used as control strain in the array comparative genomic hybridization (aCGH) analysis. *S. cerevisiae* strains were grown in yeast extract peptone dextrose (YPD) medium (pH 5.5)

containing 20 g/L glucose, 20 g/L peptone, and 10 g/L yeast extract at 30 °C.

Pulsed-field gel electrophoresis and aCGH analysis

NY1308 and BY4741 cells were cultured in 25 mL YPD with an initial OD₆₀₀ of 0.05 for 30 h and then collected for pulsed-field gel electrophoresis (PFGE) analysis using CHEF Mapper XA equipment (Bio-Rad, Hercules, CA, USA). The detailed PFGE protocols were provided by Argueso et al. (2009). The total genomic DNA from NY1308 and BY4741 was isolated using the yeast DNA kit (Omega, Doraville, GA, USA) and was then sonicated (Bioruptor, Diagenode, Liege, Belgium). The resulting sheared DNA (200–1000 bp) was labeled with Cy5/Cy3 (the NY1308 sample was labeled with Cy5, and the BY4741 sample was labeled with Cy3) and hybridized to *S. cerevisiae* CGH 385K whole-genome tiling arrays (NimbleGen, Madison, WI, USA). Microarray scanning and data analysis were performed as previously described (Zheng et al. 2012).

Whole genome sequencing of *S. cerevisiae* NY1308

NY1308 cells were cultured in 25 mL of YPD medium (initial OD₆₀₀ of 0.05) for 24 h. Then, yeast cells (6×10^8) were collected by centrifugation (8000 rpm for 5 min) for genomic DNA extraction using a yeast DNA kit (Omega, Doraville, GA, USA). DNA was broken into fragments of around 2 kb by sonication for sequencing library construction [SPRIworks Fragment Library System II kit (Beckman Coulter, Fullerton, USA)]. Genome sequencing was performed on a 454 Life Sciences Genome Sequencer FLX platform.

RNA-Seq

NY1308 cells were cultured in 25 mL of YPD medium (initial OD₆₀₀ of 0.05) for 18 h. The total RNA from 3×10^8 cells was extracted using a fungal RNAout kit (Tiandz, Beijing, China). cDNA libraries were prepared and were sequenced as described in our previous study (Zheng et al. 2012).

Ethanol fermentation

Yeast cells were precultured in 25 mL of YPD medium (initial OD₆₀₀ of 0.05) for 24 h. We then collected 3×10^9 cells and transferred them to 200 mL of a regular fermentation medium (100 g/L glucose, 0.5% yeast extract, and 1% peptone; pH 4.5) or an inhibitor-containing fermentation medium (100 g/L glucose, 5 g/L yeast extract, 10 g/L peptone, 4 g/L acetic acid, pH 4.5). Fermentation was performed in a shaking incubator at 100 rpm at 33 °C. The ethanol concentration was measured as was previously described (Zheng et al. 2011).

qRT-PCR

The PrimeScript RT Reagent Kit (TaKaRa, Dalian, China) was used to reverse transcribe RNA samples into cDNA. The qRT-PCR experiments were performed using an ABI Prism 7500 StepOnePlus instrument as described by Zheng et al. (2013). The primers that were used for quantitative PCR are listed in Online Table S1 (Online resource 1).

Genetic manipulation

The cassettes used for gene deletion were amplified by PCR using plasmid pUG6 as a template (Gueldener et al. 2002). The primers used for gene knockout and diagnostic PCR are listed in Online Table S1 (Online resource 1). Yeast transformation was performed using LiAc/SS carrier DNA/PEG method (Gietz and Schiestl 2007). Transformants were selected on YPD plates containing 300 µg/mL G418.

Data deposition

The sequences of NY1308 genome were deposited in DDBJ/EMBL/GenBank under the Whole Genome Shotgun project (GenBank: GCA_000416405.1). The raw data of RNA-Seq was deposited in the Sequence Read Archive database with an accession number of SRP126425.

Results

Karyotype analysis of *S. cerevisiae* strain NY1308

NY1308 is a diploid *S. cerevisiae* strain (confirmed by flow cytometry analysis; data not shown). The sporulation efficiency of this strain was only 0.5%, and no single spore could germinate (79 asci were dissected). PFGE experiment revealed that strains NY1308 and BY4741 differed distinctly in the length of the 16 chromosomes (chrs) (Fig. 1a). We observed that both chromosome (chr)1 and chr6 had two bands (chr1L and chr1S; chr6L and chr6S) in NY1308 (Fig. 1a). Using aCGH, the DNA copy number variations (CNVs) between NY1308 and BY4741 were examined (Fig. 1b). The red and violet regions in Fig. 1b show the amplified and underrepresented DNA segments, respectively, in NY1308 relative to BY4741. Most of these regions are located near the telomeres, on long terminal repeat retrotransposons, or on tandemly repeated arrays. For example, the genes *ADH7*, *RDS1*, *NFT1*, and *YKR104W* on the 3' end of chr3 might be lost in the NY1308 genome compared with the BY4741 genome, while the genes *AQY1*, *HPA2*, *OPT2*, *YPR196W*, *SGE1*, *ARR1*, *ARR2*, and *ARR3* on the 3' end of chr16 were amplified in NY1308 (Fig. 1b). Two amplification signals detected in the central section of chr3 and chr5 corresponded to the *LEU2*

and *URA3* marker locus, respectively (Fig. 1b). These genes were intentionally deleted in BY4741 but were present in NY1308. Unlike some other industrial *S. cerevisiae* strains (such as VL3, FostersB, FostersO, and ZTW1) (Borneman et al. 2011; Zhang et al. 2016), there were no large chromosomal aberrations (whole chromosome amplification or deletion) that occurred in the genome of NY1308 (Fig. 1b).

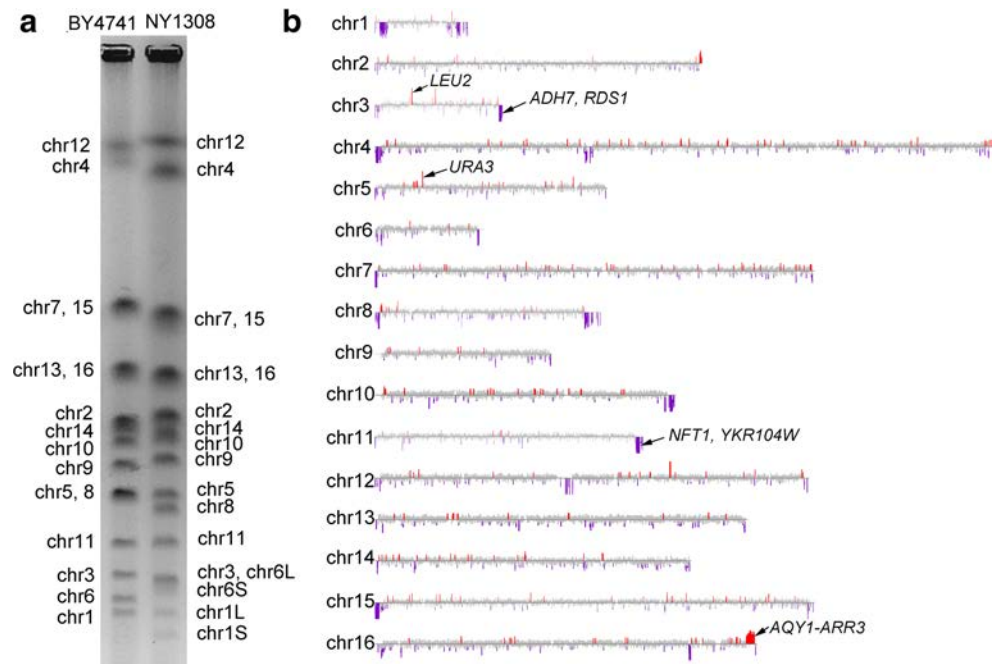
WGS of *S. cerevisiae* strain NY1308

To investigate the genetic traits of NY1308 at a higher resolution, WGS of this strain was performed. An assembly of 564,542 reads (average length of 726 bp) into 378 contigs (> 1 kb) was performed using the software Newbler 2.5.3 (Roche, Basel, Switzerland) with the default parameters. These contigs were sorted using the Mauve 3.0 (Darling et al. 2004) software with the S288c genome as the reference (Darling et al. 2004). We also used the data from Sanger sequencing reactions to close some of the gaps. The NY1308 nuclear chrs were finally covered by 35 large contigs (total length of 11.5 kb).

Using the software gsMapper (Roche, Basel, Switzerland), we identified 71,186 single nucleotide variations (SNVs) (each SNV was supported by at least five reads) within the aligned regions of the NY1308 and S288c genomes (the location of the SNVs and their annotations are listed in Online Dataset S1 in Online resource 2). A total of 43,963 SNVs were found in the coding DNA sequence (CDS) regions, and 33% of them resulted in non-synonymous mutations (the locations and effects of SNVs are listed in Online Dataset S1). We found that there were 51 S288c genes (e.g., *ECM1*, *GIT1*, *TRM2*, and *BLS1*) that had in-frame stop codons and 39 genes had lost their start (e.g., *BIO1*, *MET28*, and *TAD1*) or stop codons (e.g., *COA1*, *BSC1*, and *CRS5*) due to the presence of SNVs (Online Dataset S1 in Online resource 2). Using the number of SNVs separating any two isolates as an estimation of their relatedness, we constructed a neighbor-joining tree that represented the genetic distances among 28 *S. cerevisiae* strains with different backgrounds. The tree shows that NY1308 displayed the closest evolutionary relatedness to the sake strains (K7 and K11) and the bioethanol-producing strain ZTW1 (Fig. 2).

We identified 2218 small insertions and deletions (InDels) after mapping the reads of NY1308 onto the S288c genome. The location of the InDels and their effects on protein sequences are listed in Online Dataset S2 (Online resource 2). In summary, 29.8% of these InDels occurred in the ORF regions, resulting in 59 frame shifts events observed on 60 NY1308 ORFs (Online Dataset S2). InDels that appeared on certain functional genes, such as *PTK1*, *SRN2*, *YHR210C*, and *FIT1*, were confirmed by Sanger sequencing (Online Fig. S1 in Online resource 1).

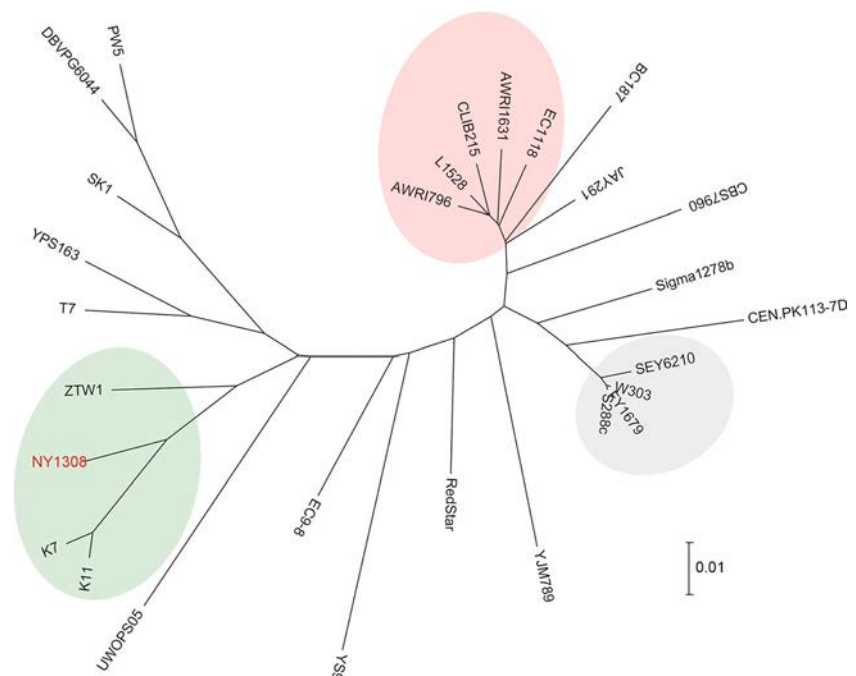
Fig. 1 Genome structure analysis of strain NY1308. **a** Pulsed-field gel electrophoresis of the BY4741 and NY1308 chromosomes. **b** Comparison of the genome structures of NY1308 and BY4741 by array-comparative genomic hybridization (aCGH). Amplified regions and underrepresented regions in NY1308 are shown in red and violet, respectively (Color figure online)



The alignment of the chrs of NY1308 and S288c using the ACT software (Carver et al. 2005) revealed the large deletion and insertion events in the NY1308 genome compared with S288c genome. There were four insertions [region A (19.5 kb), region B (31.0 kb), region C (21.2 kb), and region D (16.2 kb)] located on chr6 (19.3–38.8 and 119.2–150.2 kb; Fig. 3a) and chr14 (3.2–24.3 and 788.4–802.5 kb; Fig. 3b), which contain five, seven, four, and two putative ORFs,

respectively (Fig. 3a, b). The sequencing depth (Seq-depth) calculation showed that the depths of regions A–D were different from the average depth of chr6 (24.3 \times) and chr14 (24.1 \times) (Fig. 3a, b), suggesting that these regions were represented with different copy numbers. For example, region B may exist only on one of a pair of homologous chr6 because the Seq-depth of region B (12.1 \times) was nearly half that of chr6 (Fig. 3a). The sequence analysis of the 33 reads covering the

Fig. 2 A neighbor-joining tree representing the genetic distances between *S. cerevisiae* strains calculated from the SNPs present in the 35-kb aligned DNA regions gathered from all 16 chromosomes. Sake strains, wine strains, and laboratory strains are clustered in the green, pink, and gray regions (Color figure online)



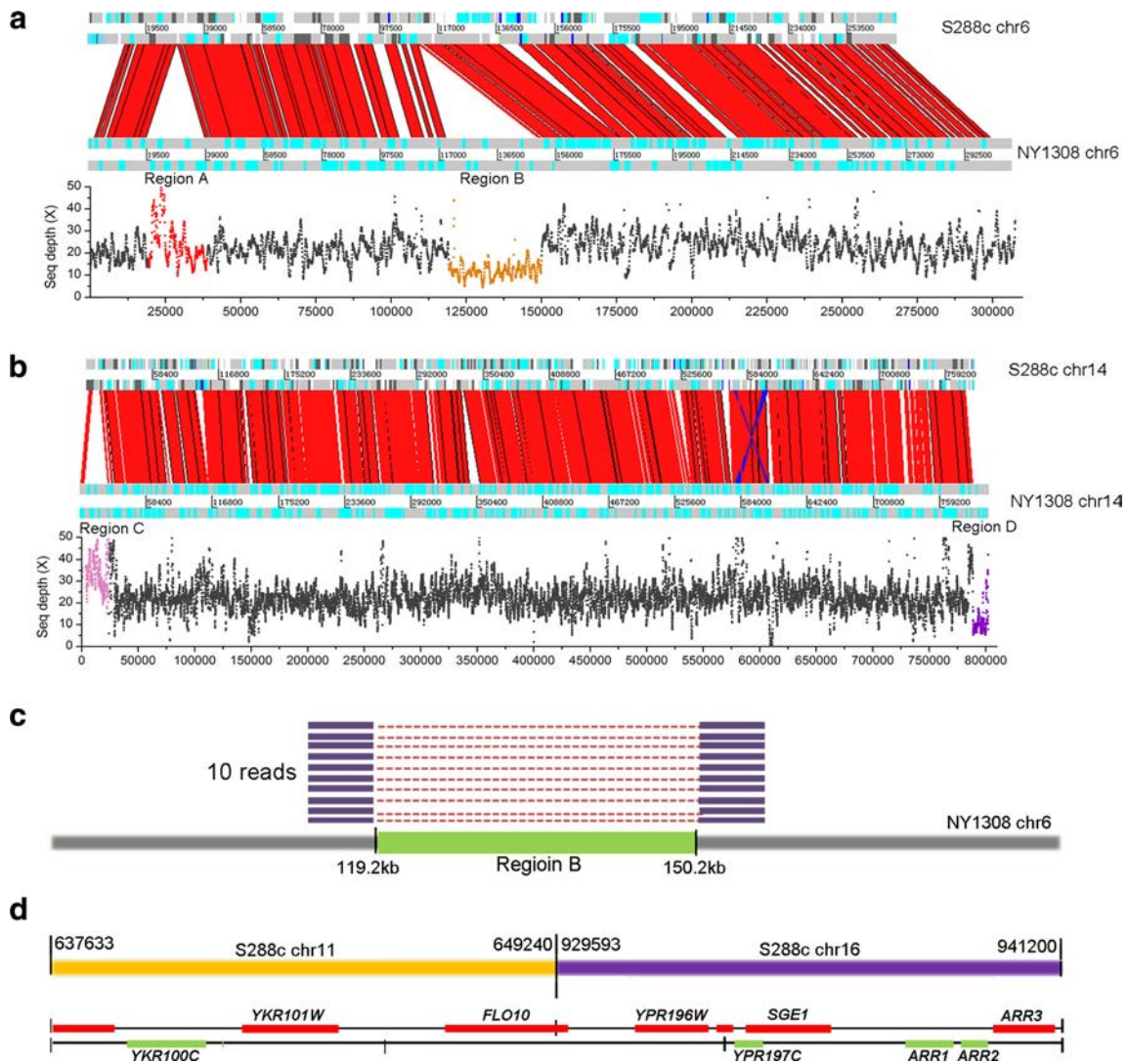


Fig. 3 Insertion and translocation events occurred on the NY1308 genome compared to the S288c genome. Alignments of **a** chr6 and **b** chr14 of NY1308 and S288c. The point diagrams show the read depths of these two chromosomes. The colorful points indicate four inserted

regions that are absent in the S288c genome. **c** Ten reads supported the absence of inserted region B on one large chr6. **d** A translocation event occurred between chr11 and chr16, which was mediated by the gene *FLO10* (Color figure online)

break points of region B (119.2 kb and 150.2 kb on chr6) disclosed that there were 10 reads supporting the direct connection of 119.2–150.2 kb of chr6 and the remaining 22 reads supported the insertion of region B (Fig. 3c). This result explained the different lengths between the pair of homologous chr6s (chr6L and chr6S) on the NY1308 genome (Fig. 1a). In addition to large insertions, certain chromosomal rearrangement events were also discovered in the NY1308 genome compared with the S288c genome. Figure 1b indicates that the DNA sequences that are homologous to region *AQY1-ARR3* were duplicated on the NY1308 genome, but the chromosomal location of the extra copy of *AQY1-ARR3* was not known. The sequence analysis of the reads mapped on these regions suggested that the extra copy of the region (929,593–941,200 bp on chr16) was joined to the *FLO10* gene located on the right end of chr11 through translocation (Fig. 3d).

These observations are consistent with previous findings that the areas near the telomere on *S. cerevisiae* genomes easily undergo rearrangements (Argueso et al. 2009).

A total of 5413 ORFs were predicted for the nuclear genome of NY1308 using Augustus (the location and annotation of these ORFs are listed in Online Dataset S3 in Online resource 2) (Stanke and Morgenstern 2005). Most (99%) of these predicted ORFs have their homolog (identify is above 95%) on the S288c genome (the annotations of these ORFs are listed in Online Dataset S3). In addition, the NY1308 genome has 43 genes that are absent in the S288c genome (Table 1). By aligning the sequences of these unique genes onto the genomes of multiple *Saccharomyces* species (*Saccharomyces bayanus*, *Saccharomyces kudriavzevii*, *Saccharomyces uvarum*, *Saccharomyces paradoxus*, *Saccharomyces castellii*, and *Saccharomyces mikatae*), we

Table 1 Annotation of NY1308 unique genes absent in S288c

Gene	Contig	Start	End	Intron ^a	Function
NYn1	Contig1.01	164,383	165,093	Y	Mst27p
NYn2	Contig1.02	2122	2817	N	Uip3p
NYn3	Contig1.02	5394	6101	N	Prm9p
NYn4	Contig1.02	7112	7834	N	Mst28p
NYn5	Contig2.02	618,257	618,481	N	Unknown
NYn6	Contig4.01	265	954	N	Related to histone acetyltransferase hpa2 and related acetyltransferases
NYn7	Contig4.01	11,785	12,216	N	Aad14p
NYn8	Contig4.03	66,436	67,866	N	Ehd3p
NYn9	Contig4.03	330,184	330,627	N	Hypothetical protein QA23_0892
NYn10	Contig6.01	290	934	N	Related to histone acetyltransferase hpa2 and related acetyltransferases
NYn11	Contig6.01	3567	4586	N	Aad14p
NYn12	Contig6.01	25,204	27,114	Y	Siderophore iron
NYn13	Contig6.01	30,076	31,503	N	Hypothetical protein C1Q_05666
NYn14	Contig6.01	32,325	33,959	N	Aminotriazole resistance protein
NYn15	Contig6.01	119,167	120,405	Y	Thi73p
NYn16	Contig6.01	124,967	125,719	N	Aspartate racemase
NYn17	Contig6.01	126,267	127,394	N	Mdh2p
NYn18	Contig6.01	128,717	129,979	N	Aat2p
NYn19	Contig6.01	131,469	133,210	Y	Dip5p
NYn20	Contig6.01	136,535	136,923	Y	Unknown
NYn21	Contig6.01	141,624	142,178	N	Azole resistance
NYn22	Contig6.01	142,691	143,347	N	Ymr226c-like protein
NYn23	Contig6.01	143,693	145,136	Y	Early growth response protein 1
NYn24	Contig6.01	145,835	146,294	Y	Unknown
NYn25	Contig6.01	146,358	148,478	Y	Hydantoinase B/oxoprolinase
NYn26	Contig6.01	149,042	149,482	N	Thi73p
NYn27	Contig6.01	201,891	202,972	Y	Yfi012w-like protein
NYn28	Contig8.01	1523	2452	N	Pug1p
NYn29	Contig8.01	3150	5402	N	Scy_1426 protein
NYn30	Contig9.01	286,465	287,458	Y	Unknown
NYn31	Contig12.02	375,157	375,558	N	Unknown
NYn32	Contig13.01	452,624	452,887	N	Unknown
NYn33	Contig14.01	6518	8482	N	Transcriptional activator of proline utilization genes
NYn34	Contig14.01	8544	12,407	N	Ykl215c-like protein
NYn35	Contig14.01	12,888	14,399	N	Nicotinic acid plasma membrane transporter
NYn36	Contig14.01	16,210	17,391	N	Unknown
NYn37	Contig14.4	354,753	355,580	N	Related to ammonia transport outward protein 3
NYn38	Contig14.4	356,961	358,745	N	Amidase homolog
NYn39	Contig14.4	360,842	362,440	N	Yps1p
NYn40	Contig14.4	363,707	364,417	N	Yir042c-like protein
NYn41	Contig15.01	153,344	153,550	N	Hypothetical protein CENPK1137D_2436
NYn42	Contig16.01	55,160	56,278	N	Hypothetical protein C1Q_05652
NYn43	Contig6.01	35,943	39,308	N	Flocculin

^a“Y” and “N” indicate whether containing intron or not

identified the existence of genes *NYn2*, *NYn4*, *NYn11*, *NYn37*, and *NYn38* in the genome of *S. paradoxus*. The functions of unique genes were predicted by searching the nr database in

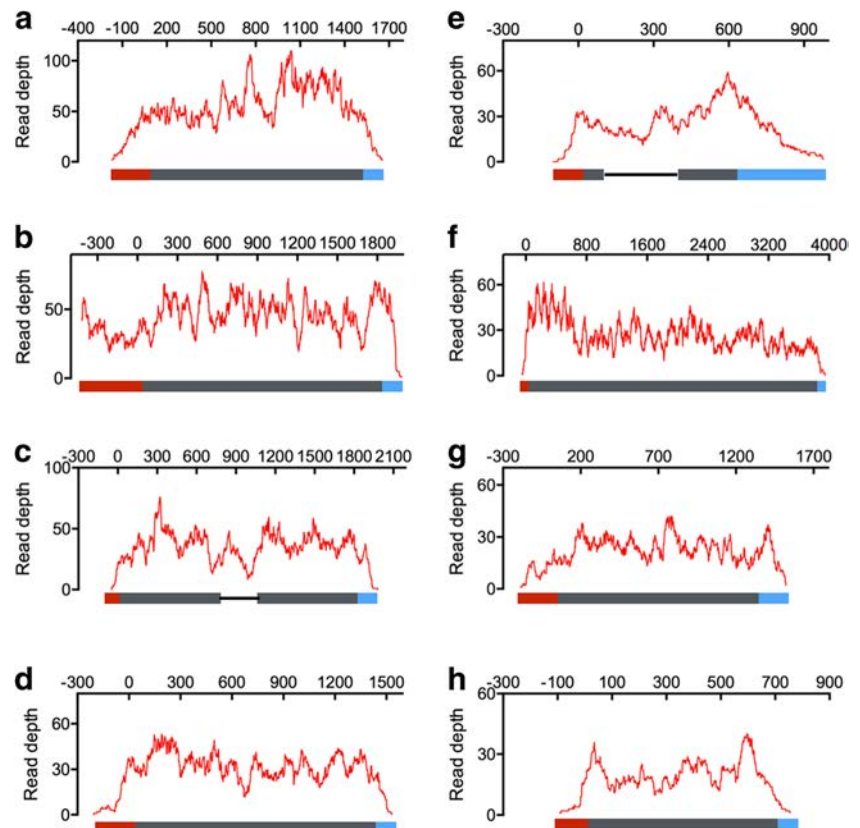
NCBI (<http://www.ncbi.nlm.nih.gov/>) using the Blast2GO software with default settings. Gene function annotation suggested that these genes are mainly involved in

transmembrane transport, transcription regulation, and biological catalytic processes (such as *N*-acetyltransferase activity and hydrolase activity) (Table 1). Interestingly, unlike most of the *S. cerevisiae* S288c genes, some of those 43 unique NY1308 genes contain introns (Table 1). For example, genes *NYn19*, *NYn24*, and *NYn25* were predicted to contain an intron of 170, 76, and 498 bp, respectively. Phylogenetic analysis suggested that some of these genes might have been obtained from other microorganisms through horizontal gene transfer (Online Fig. S2). One such example is the *NYn16* gene that encodes the aspartate racemase (EC 5.1.1.13) that converts L-aspartate to D-aspartate (Table 1). This enzyme is usually found in bacteria, and no homology could be identified in yeast genomes. Sequence alignment and phylogenetic analysis show that the protein sequence of *NYn16* is most closely related to the aspartate racemase from the bacteria *Acinetobacter calcoaceticus* (Online Fig. S2A). Another example is the *NYn19* gene that encodes dicarboxylic amino acid permease, which mediates high-affinity and high-capacity transport of L-glutamate and L-aspartate (Table 1). In *S. cerevisiae* S288c, this enzyme was encoded by the gene *DIP5*, corresponding to the gene contig16.01.g5001 in NY1308 (with a DNA similarity above 99.7%; Online Dataset S3). Online Fig. S2B shows that the protein sequence of *NYn19* was more similar to the dicarboxylic amino acid permease from *Tetrapispora phaffii* than to the S288c *DIP5*.

RNA-Seq to identify the landscapes and activities of unique genes in the NY1308 genome

Although genome sequencing of different *S. cerevisiae* strains has identified dozens of genes that are absent in the S288c reference genome, the expression traits [i.e., untranslated regions (UTR), expression levels under different conditions, and alternative splicing] and physiological roles of those predicted genes have been less frequently investigated. Using RNA-Seq, the UTR regions, relative RNA abundance, and RNA editing of the NY1308 unique genes were determined. Online Dataset S4 (Online resource 2) shows significant differences in the expression activity among the 43 unique NY1308 genes [the expression levels of each gene were normalized by the Reads per Kilobase per Million mapped reads (RPKM) method (Mortazavi et al. 2008)]. The three genes (*NYn35*, *NYn33*, and *NYn12*) with the highest expression levels encode the nicotinic acid plasma membrane transporter, transcriptional activator, and iron siderophore, respectively (Online Dataset S4 and Table 1). Calculation of the RNA-Seq depth allowed us to determine the transcriptional landscapes of these genes. Figure 4a–e shows the 5'UTR, CDS, and 3'UTR region of the eight genes (*NYn35*, *NYn33*, *NYn12*, *NYn8*, *NYn1*, *NYn34*, *NYn34*, and *NYn6*) with an RPKM value above 40. The average length of the 5'UTR and 3'UTR of these genes was 126 and 110 bp, respectively. Interestingly, no difference was observed in the Seq-depth between the exon and the intron

Fig. 4 Transcription analysis of eight unique genes **a** *NYn35*, **b** *NYn33*, **c** *NYn12*, **d** *NYn8*, **e** *NYn1*, **f** *NYn34*, **g** *NYn34*, and **h** *NYn6* identified on the NY1308 genome using RNA-Seq. The 5' UTR, CDS, and 3' UTR of each gene are colored with red, gray, and blue, respectively. The black lines indicate introns of genes *NYn12* and *NYn1* (Color figure online)



regions of the *NYn12* and *NYn1* genes (Fig. 4c, e). Additionally, the sequence analysis of all the reads mapped on the genes with introns (*NYn1*, *NYn12*, *NYn15*, *NYn19*, *NYn23*, and *NYn30*; Table 1) confirmed that the intron sequences were always retained in the transcripts of these genes. These results suggest that NY1308 might not be able to recognize the splice site and remove the intron sequences of these genes.

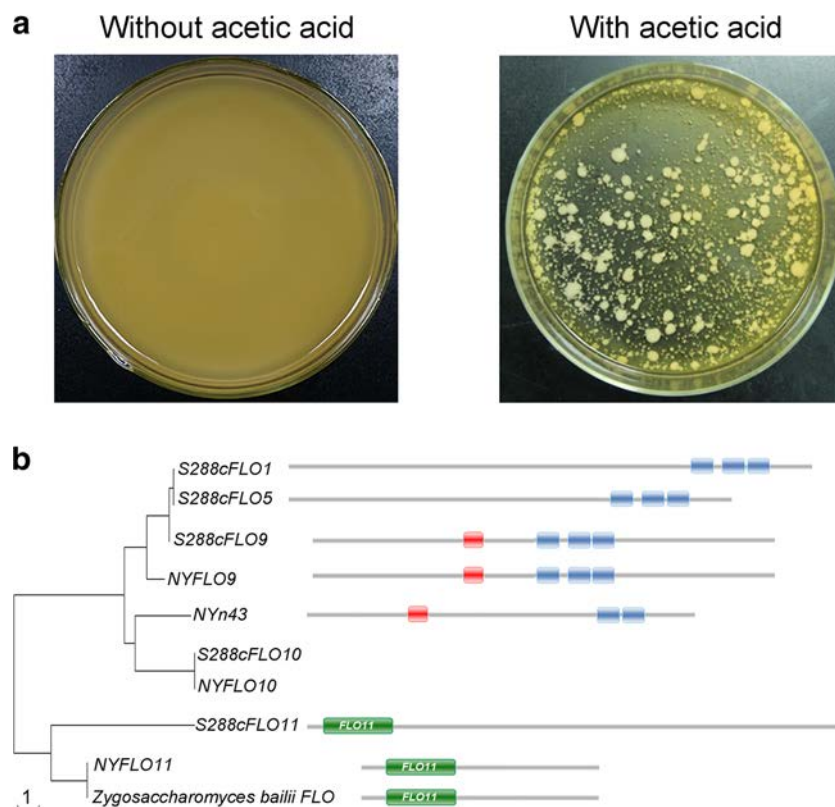
Personalized modifications to improve the ethanol fermentation rate of NY1308

An interesting phenomenon observed in this study is NY1308 cells would form flocculated particles when fermented in medium containing acetic acid (a common inhibitor in cellulosic hydrolysate) (Fig. 5a). Although flocculation may be beneficial to resist stressful conditions for yeast cells (Deed et al. 2017; Westman et al. 2014), the floc sediment would also show a much slower fermentation rate compared to the free cells due to the reduced surface-to-volume ratio of the aggregated cells. It was suggested that the flocculation process was associated with multiple genes, such as *FLO1*, *FLO5*, *FLO9*, *FLO10*, and *FLO11*, which were annotated on the S288c genome (Halme et al. 2004). The genome annotation of the NY1308 genome predicted four flocculin-encoding genes (*NYFLO9*, *NYFLO10*, *NYFLO11*, and *NYn43*). Figure 5b shows the phylogenetic relationships and flocculin domains

(including Flocculin, Flocculin-t3, and Flo11 predicted by the Pfam database) of the flocculin genes of NY1308. The protein sequence of gene *NYn43* is the most similar to the *FLO10* gene of S288c and contains three flocculin domains (Fig. 5b). The sequence of *NYFLO11* had high identity compared to a gene from *Zygosaccharomyces bailii* (Fig. 5b).

To investigate whether and how these four *FLO* genes contributed to the flocculation phenotype, qRT-PCR and genetic manipulation of these flocculin genes were performed. It was found that genes *NYn43* and *NYFLO9* could be greatly induced during fermentation in the presence of acetic acid compared to the 0-h point (Fig. 6a). *NYn43* showed the highest expression at 30 h, which was 110-fold higher than that of the 0-h time point (Fig. 6a). In contrast, the expression levels of the *NYFLO10* and *NYFLO11* genes were not changed dramatically during the fermentation process compared to the 0-h point (Fig. 6a). The flocculation ability of NY1308 cells was greatly diminished after the deletion of *NYFLO9* or *NYn43* during ethanol fermentation (Online Fig. S3), while the deletion of *NYFLO10* and *NYFLO11* had no effect on flocculation for NY1308 (data not shown). Although deletion of *NYn43* and *FLO9* had similar effect on flocculation ability, different ethanol fermentation rates were observed between the *NYn43* deletion mutant (NY Δ NYn43) and *NYFLO9* deletion mutant (NY Δ FLO9). Mutant NY Δ NYn43 showed a significantly shorter fermentation period than NY1308 in acetic acid containing medium (Fig. 6b).

Fig. 5 Flocculation of NY1308 cells and *FLO* genes. **a** NY1308 cells formed flocs during the fermentation process when 4 g/L acetic acid was added into medium. Ethanol fermentation were conducted in 500-mL flasks for 40 h, and the yeast cells were photographed in empty plates. **b** Phylogeny of *FLO* genes on the NY1308 and S288c genomes. The green, red, and blue rectangles represent motifs of Flo11, Flocculin, and Flocculin-t3, respectively (Color figure online)



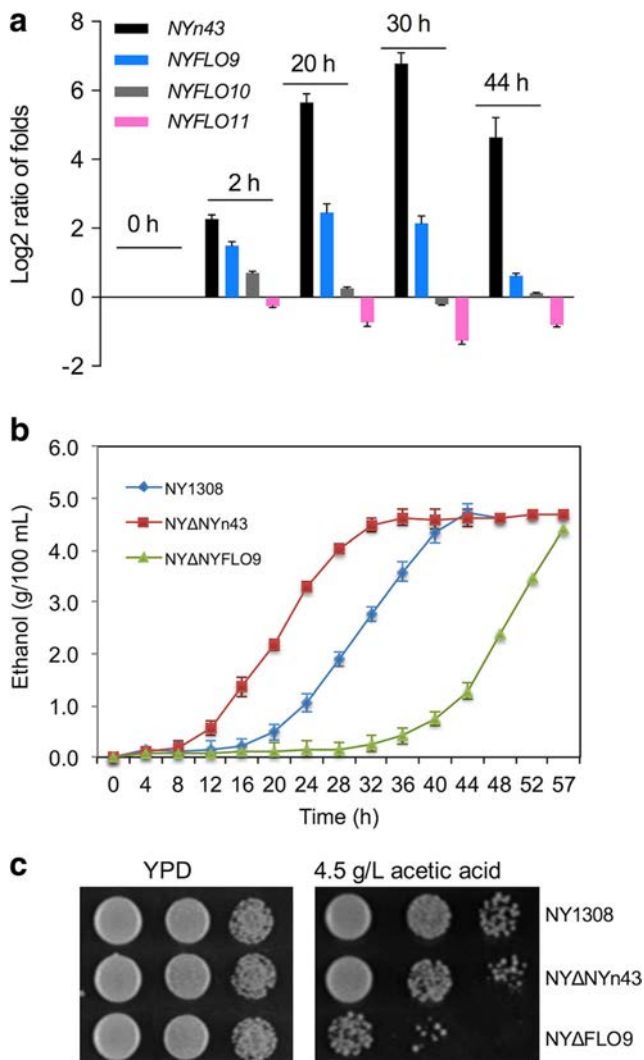


Fig. 6 Effects of *FLO* genes on ethanol fermentation and stress tolerance of NY1308. **a** Expression activities of genes *NYn43*, *NYFLO9*, *NYFLO10*, and *NYFLO11* during ethanol fermentation process in the presence of acetic acid. **b** Effects of deletions of *NYn43* and *NYFLO9* on ethanol fermentation rates. Three independent experiments were conducted, and the data was represented as mean \pm SD. **c** Effects of deletions of *NYn43* and *NYFLO9* on acetic acid tolerance. Three independent experiments were conducted, and one typical result was shown

In contrast, mutant NYΔ*FLO9* had a longer lag phase than NY1308 (Fig. 6b). An explanation to this difference is *NYFLO9* deletion led to great decrease in tolerance to acetic acid for NY1308, whereas *NYn43* deletion had no significant effect on acetic acid of NY1308 (Fig. 6c).

Discussion

This study disclosed the genetic characteristics of a bioethanol-producing *S. cerevisiae* strain, NY1308. Using a PFGE and aCGH array, we confirmed that no chromosomal

aneuploidy (loss or amplification of whole chromosome) events occurred on the diploid genome of NY1308. One cause of the heterogeneity in length between the pair of homologous chr1 and chr6 (Fig. 1a) is the insertion of DNA (Fig. 3). Whole genome sequencing identified that the nucleotide polymorphisms between NY1308 and S288c were 0.71%, which was very similar to the level separating S288c and ZTW1 (another industrial strain used for bioethanol production in China) (Zhang et al. 2015). NY1308 has a closer genetic relationship with sake strains (Fig. 2), which is consistent with the fact that these strains were isolated from East Asia. The frequency of InDels (0.02%) is much less than that of SNVs, but this type of mutation resulted in the functional deactivation of a considerable number of ORFs. A higher ratio of point mutations (up to 40% of the SNVs and 88% of the InDels) was observed in intergenic sequences, which were previously suggested as the main cause of the *cis* regulation of gene expression among yeast populations (Zheng et al. 2010).

Among the 71,186 SNVs listed in Online Dataset S1, 63% were detected in both homologs (the ratio of variation frequency was above 80%; Online Dataset S1). Interestingly, these homologous SNVs were not evenly distributed across the NY1308 genome and were clustered within certain chromosomal segments (Online Fig. S4). For example, it can be clearly seen that the SNVs within the 277 to 348 kb regions of chr5, 202 to 551 kb of chr8, and 1 to 360 kb of chr11 were homologous for the pairs of chromosomes (Online Fig. S4). A possible explanation for this pattern is the loss of heterozygosity (LOH) caused by homologous recombination during mitotic generations. Depending on the different pathways of mitotic recombination (Symington et al. 2014), two main genetic outcomes would be expected: gene conversion (internal LOH; such as 277 to 348 kb of chr5 shown in Online Fig. S4) and crossover (terminal LOH, from a certain break point to the end of a chromosome). Unlike programmed meiotic recombination, mitotic recombination aims to repair DNA double-strand breaks that could be induced by multiple exogenous or endogenous sources (Yin and Petes 2013; Zheng et al. 2016). Although NY1308 had a poor sporulation ratio, mitotic recombination-associated loss of heterozygosity would provide an alternative to rapid genomic purification and evolution.

De novo assembly and annotation of the NY1308 genome identified 43 ORFs that were absent from the S288c genome (Table 1). One possible source of these unique genes is interspecies hybridizations that were followed by the gradual loss of the contributing genomes. On the other hand, some unique genes might be obtained from bacteria through uncertain pathways (Online Fig. S2). Generally, these unique genes are not necessary for the normal growth of yeast cells but may bring certain benefits for individual strains under a specific condition. For example, the horizontal transfer gene *FSY1* (encodes a high-affinity fructose/H⁺ symporter) identified in the

genome of *S. cerevisiae* EC1118 conferred a significant advantage in sugar utilization to *S. cerevisiae* during the wine fermentation process (Galeote et al. 2010), and genes *BIO1* and *BIO6* (obtained from bacteria) contributed to the biotin synthesis capability of sake strains (Hall and Dietrich 2007). However, not all these horizontally transferred genes could function normally in *S. cerevisiae*. Our RNA-Seq results suggest that those genes may be deactivated due to the intron sequences, which could not be identified and spliced from the mRNA transcript.

We found the presence of acetic acid in the fermentation medium would induce the flocculation of NY1308 cells (Fig. 5a). For beer production, flocculation (the calcium-dependent, non-sexual aggregation of cells into clumps) is a necessary process because it causes the yeast to sediment, and the yeast cells can be easily harvested (without the centrifugation step) for the next fermentation (Bauer et al. 2010). However, flocculation is usually an undesirable trait for most *S. cerevisiae* strains used in the bioethanol field (Bauer et al. 2010; Zhao and Bai 2009). In the S288c genome, five *FLO* genes have been annotated. The *FLO5* gene is a paralog of *FLO1*, which arose from a segmented duplication. The *FLO9* and *FLO10* genes are 94 and 58% similar to *FLO1*, respectively. These four genes confer cell-cell adhesion as flocculins could bind to mannose chains on the surface of neighboring cells, leading to the cross binding of cells and ultimately the formation of flocs (Halme et al. 2004). *FLO11* encodes a GPI-anchored cell surface flocculin and contains a Flo11 flocculin domain, which is responsible for adhesion to substrates (Lo and Dranginis 1996). Among the four *FLO* genes identified on the NY1308 genome, *NYn43* and *NYFLO9* were greatly up-regulated during the fermentation process in the presence of acetic acid, particularly at the 20- and 30-h time points (Fig. 6a). The individual deletions of *NYn43* or *NYFLO9* greatly reduced the flocculation capability (Online Fig. S3), suggesting that these two genes synergistically contributed to the flocculation trait of NY1308. Consistent with previous observations that the functions of the *FLO* genes contribute to the resistance of multiple stressors (Deed et al. 2017; Westman et al. 2014), the deletion of *NYFLO9* resulted in a decreased tolerance to acetic acid and a prolonged lag phase to start ethanol fermentation (Fig. 6b, c). Nevertheless, the *NYn43* deletion had no significant influence on acetic acid tolerance and effectively improved the fermentation rate of NY1308 (Fig. 6b, c). It seems that the mechanisms underlying the phenotypic changes are distinctive for each individual Flo protein. In addition, it was noted that the genetic background of strains may also affect the phenotypic effects of the expression of an individual *FLO* gene (Govender et al. 2010). These observations underline the importance of the strain-by-strain approach to improve the commercial traits of yeasts.

Funding This study was funded by the National Natural Science Foundation of China (31401058 and 31370132) and Natural Science Foundation of Zhejiang Province (LY18C060002).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

- Akao T, Yashiro I, Hosoyama A, Kitagaki H, Horikawa H, Watanabe D, Akada R, Ando Y, Harashima S, Inoue T, Inoue Y, Kajiwara S, Kitamoto K, Kitamoto N, Kobayashi O, Kuhara S, Masubuchi T, Mizoguchi H, Nakao Y, Nakazato A, Namise M, Oba T, Ogata T, Ohta A, Sato M, Shibasaki S, Takatsume Y, Tanimoto S, Tsuboi H, Nishimura A, Yoda K, Ishikawa T, Iwashita K, Fujita N, Shimoi H (2011) Whole-genome sequencing of sake yeast *Saccharomyces cerevisiae* Kyokai no. 7. DNA Res 18(6):423–434. <https://doi.org/10.1093/dnares/dsr029>
- Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OV, Missawa SK, Galzerani F, Costa GG, Vidal RO, Noronha MF, Dominska M, Andrietta MG, Andrietta SR, Cunha AF, Gomes LH, Tavares FC, Alcarde AR, Dietrich FS, McCusker JH, Petes TD, Pereira GA (2009) Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. Genome Res 19(12):2258–2270. <https://doi.org/10.1101/gr.091777.109>
- Bauer FF, Govender P, Bester MC (2010) Yeast flocculation and its biotechnological relevance. Appl Microbiol Biotechnol 88(1):31–39. <https://doi.org/10.1007/s00253-010-2783-0>
- Borneman AR, Riches D, Affourtit JP, Forgan AH, Pretorius IS, Egholm M, Chambers PJ (2011) Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*. PLoS Genet 7(2): e1001287. <https://doi.org/10.1371/journal.pgen.1001287>
- Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J (2005) ACT: the Artemis comparison tool. Bioinformatics 21(16):3422–3423. <https://doi.org/10.1093/bioinformatics/bti553>
- Coi AL, Bigey F, Mallet S, Marsit S, Zara G, Gladieux P, Galeote V, Budroni M, Dequin S, Legras JL (2017) Genomic signatures of adaptation to wine biological ageing conditions in biofilm-forming flor yeasts. Mol Ecol 26(7):2150–2166. <https://doi.org/10.1111/mec.14053>
- Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14(7):1394–1403. <https://doi.org/10.1101/gr.2289704>
- Deed RC, Fedrizzi B, Gardner RC (2017) *Saccharomyces cerevisiae* *FLO1* gene demonstrates genetic linkage to increased fermentation rate at low temperatures. G3 (Bethesda) 7(3):1039–1048. <https://doi.org/10.1534/g3.116.037630>
- Fay JC, McCullough HL, Sniegowski PD, Eisen MB (2004) Population genetic variation in gene expression is associated with phenotypic variation in *Saccharomyces cerevisiae*. Genome Biol 5(4):R26. <https://doi.org/10.1186/gb-2004-5-4-r26>
- Galeote V, Novo M, Salema-Oom M, Brion C, Valerio E, Goncalves P, Dequin S (2010) *FSY1*, a horizontally transferred gene in the *Saccharomyces cerevisiae* EC1118 wine yeast strain, encodes a high-affinity fructose/H⁺ symporter. Microbiology 156(12):3754–3761. <https://doi.org/10.1099/mic.0.041673-0>
- Gietz RD, Schiestl RH (2007) High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat Protoc 2(1):31–34. <https://doi.org/10.1038/nprot.2007.13>

- Govender P, Bester M, Bauer FF (2010) *FLO* gene-dependent phenotypes in industrial wine yeast strains. *Appl Microbiol Biotechnol* 86(3): 931–945. <https://doi.org/10.1007/s00253-009-2381-1>
- Gueldener U, Heinisch J, Koehler GJ, Voss D, Hegemann JH (2002) A second set of *loxP* marker cassettes for Cre-mediated multiple gene knockouts in budding yeast. *Nucleic Acids Res* 30(6):e23–223. <https://doi.org/10.1093/nar/30.6.e23>
- Hall C, Dietrich FS (2007) The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. *Genetics* 177(4):2293–2307. <https://doi.org/10.1534/genetics.107.074963>
- Halme A, Bumgarner S, Styles C, Fink GR (2004) Genetic and epigenetic regulation of the *FLO* gene family generates cell-surface variation in yeast. *Cell* 116(3):405–415. [https://doi.org/10.1016/S0092-8674\(04\)00118-7](https://doi.org/10.1016/S0092-8674(04)00118-7)
- Jönsson LJ, Aliksson B, Nilvebrant N-O (2013) Bioconversion of lignocellulose: inhibitors and detoxification. *Biotechnol Biofuels* 6(1):16. <https://doi.org/10.1186/1754-6834-6-16>
- Kitichantaropas Y, Boonchird K, Sugiyama M, Kaneko Y, Harashima S, Auesukaree C (2016) Cellular mechanisms contributing to multiple stress tolerance in *Saccharomyces cerevisiae* strains with potential use in high-temperature ethanol fermentation. *AMB Express* 6(1): 107. <https://doi.org/10.1186/s13568-016-0285-x>
- Koutinas A, Kanellaki M, Bekatorou A, Kandylis P, Pissaridi K, Dima A, Boura K, Lappa K, Tsafrakidou P, Stergiou P-Y (2016) Economic evaluation of technology for a new generation biofuel production using wastes. *Bioresour Technol* 200:178–185. <https://doi.org/10.1016/j.biortech.2015.09.093>
- Lo WS, Dranginis AM (1996) *FLO11*, a yeast gene related to the *STA* genes, encodes a novel cell surface flocculin. *J Bacteriol* 178(24): 7144–7151. <https://doi.org/10.1128/jb.178.24.7144-7151.1996>
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621–628. <https://doi.org/10.1038/nmeth.1226>
- Nijkamp JF, van den Broek M, Datema E, de Kok S, Bosman L, Luttk MA, Daran-Lapujade P, Vongsangnak W, Nielsen J, Heijne WH, Klaassen P, Paddon CJ, Platt D, Kotter P, van Ham RC, Reinders MJ, Pronk JT, de Ridder D, Daran JM (2012) De novo sequencing, assembly and analysis of the genome of the laboratory strain *Saccharomyces cerevisiae* CEN.PK113-7D, a model for modern industrial biotechnology. *Microb Cell Factories* 11(1):36. <https://doi.org/10.1186/1475-2859-11-36>
- Perez-Ortín JE, Querol A, Puig S, Barrio E (2002) Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. *Genome Res* 12(10):1533–1539. <https://doi.org/10.1101/gr.436602>
- Silveira MH, Morais AR, da Costa Lopes AM, Oleksyszyn DN, Bogel-Lukasik R, Andreus J, Pereira Ramos L (2015) Current pretreatment technologies for the development of cellulosic ethanol and biorefineries. *ChemSusChem* 8(20):3366–3390. <https://doi.org/10.1002/cssc.201500282>
- Sindhu R, Binod P, Pandey A (2016) Biological pretreatment of lignocellulosic biomass—an overview. *Bioresour Technol* 199:76–82. <https://doi.org/10.1016/j.biortech.2015.08.030>
- Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 33(Web Server):W465–W467. <https://doi.org/10.1093/nar/gki458>
- Strope PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA, Magwene PM, Dietrich FS, McCusker JH (2015) The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res* 25(5):762–774. <https://doi.org/10.1101/gr.185538.114>
- Symington LS, Rothstein R, Lisby M (2014) Mechanisms and regulation of mitotic recombination in *Saccharomyces cerevisiae*. *Genetics* 198(3):795–835. <https://doi.org/10.1534/genetics.114.166140>
- Westman JO, Mapelli V, Taherzadeh MJ, Franzen CJ (2014) Flocculation causes inhibitor tolerance in *Saccharomyces cerevisiae* for second-generation bioethanol production. *Appl Environ Microbiol* 80(22): 6908–6918. <https://doi.org/10.1128/AEM.01906-14>
- Will JL, Kim HS, Clarke J, Painter JC, Fay JC, Gasch AP (2010) Incipient balancing selection through adaptive loss of aquaporins in natural *Saccharomyces cerevisiae* populations. *PLoS Genet* 6(4):e1000893. <https://doi.org/10.1371/journal.pgen.1000893>
- Yin Y, Petes TD (2013) Genome-wide high-resolution mapping of UV-induced mitotic recombination events in *Saccharomyces cerevisiae*. *PLoS Genet* 9(10):e1003894. <https://doi.org/10.1371/journal.pgen.1003894>
- Zhang K, Tong M, Gao K, Di Y, Wang P, Zhang C, Wu X, Zheng D (2015) Genomic reconstruction to improve bioethanol and ergosterol production of industrial yeast *Saccharomyces cerevisiae*. *J Ind Microbiol Biotechnol* 42(2):207–218. <https://doi.org/10.1007/s10295-014-1556-7>
- Zhang K, Zhang LJ, Fang YH, Jin XN, Qi L, Wu XC, Zheng DQ (2016) Genomic structural variation contributes to phenotypic change of industrial bioethanol yeast *Saccharomyces cerevisiae*. *FEMS Yeast Res* 16(2):fov118. <https://doi.org/10.1093/femsyr/fov118>
- Zhao XQ, Bai FW (2009) Yeast flocculation: new story in fuel ethanol production. *Biotechnol Adv* 27(6):849–856. <https://doi.org/10.1016/j.biotechadv.2009.06.006>
- Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M (2010) Genetic analysis of variation in transcription factor binding in yeast. *Nature* 464(7292):1187–1191. <https://doi.org/10.1038/nature08934>
- Zheng DQ, Wu XC, Tao XL, Wang PM, Li P, Chi XQ, Li YD, Yan QF, Zhao YH (2011) Screening and construction of *Saccharomyces cerevisiae* strains with improved multi-tolerance and bioethanol fermentation performance. *Bioresour Technol* 102(3):3020–3027. <https://doi.org/10.1016/j.biortech.2010.09.122>
- Zheng DQ, Wang PM, Chen J, Zhang K, Liu TZ, Wu XC, Li YD, Zhao YH (2012) Genome sequencing and genetic breeding of a bioethanol *Saccharomyces cerevisiae* strain YJS329. *BMC Genomics* 13(1):479. <https://doi.org/10.1186/1471-2164-13-479>
- Zheng D, Zhang K, Gao K, Liu Z, Zhang X, Li O, Sun J, Zhang X, Du F, Sun P (2013) Construction of novel *Saccharomyces cerevisiae* strains for bioethanol active dry yeast (ADY) production. *PLoS One* 8(12):e85022. <https://doi.org/10.1371/journal.pone.0085022>
- Zheng DQ, Chen J, Zhang K, Gao KH, Li O, Wang PM, Zhang XY, Du FG, Sun PY, Qu AM, Wu S, Wu XC (2014) Genomic structural variations contribute to trait improvement during whole-genome shuffling of yeast. *Appl Microbiol Biotechnol* 98(7):3059–3070. <https://doi.org/10.1007/s00253-013-5423-7>
- Zheng D-Q, Zhang K, Wu X-C, Mieczkowski PA, Petes TD (2016) Global analysis of genomic instability caused by DNA replication stress in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 113(50):E8114–E8121. <https://doi.org/10.1073/pnas.1618129113>
- Zheng DQ, Jin XN, Zhang K, Fang YH, Wu XC (2017) Novel strategy to improve vanillin tolerance and ethanol fermentation performances of *Saccharomyces cerevisiae* strains. *Bioresour Technol* 231:53–58. <https://doi.org/10.1016/j.biortech.2017.01.040>
- Zhu YO, Sherlock G, Petrov DA (2016) Whole genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. *G3 (Bethesda)* 6(8):2421–2434. <https://doi.org/10.1534/g3.116.029397>