Mol Genet Genomics (2012) 287:485–494
DOI 10.1007/s00438-012-0695-7

ORIGINAL PAPER

# Whole-genome sequencing of the efficient industrial fuel-ethanol fermentative *Saccharomyces cerevisiae* strain CAT-1

**Farbod Babrzadeh · Roxana Jalili · Chunlin Wang · Shadi Shokralla · Sarah Pierce · Avi Robinson-Mosher · Pål Nyren · Robert W. Shafer · Luiz C. Basso · Henrique V. de Amorim · Antonio J. de Oliveira · Ronald W. Davis · Mostafa Ronaghi · Baback Gharizadeh · Boris U. Stambuk**

Received: 30 October 2011 / Accepted: 23 April 2012 / Published online: 6 May 2012

**Abstract** The *Saccharomyces cerevisiae* strains widely used for industrial fuel-ethanol production have been developed by selection, but their underlying beneficial genetic polymorphisms remain unknown. Here, we report the draft whole-genome sequence of the *S. cerevisiae* strain CAT-1, which is a dominant fuel-ethanol fermentative strain from the sugarcane industry in Brazil. Our results indicate that strain CAT-1 is a highly heterozygous diploid yeast strain, and the ~12-Mb genome of CAT-1, when compared with the reference S228c genome, contains ~36,000 homozygous and ~30,000 heterozygous single nucleotide polymorphisms, exhibiting an uneven distribution among chromosomes due to large genomic regions of loss of heterozygosity (LOH). In total, 58 % of the 6,652 predicted protein-coding genes of the CAT-1 genome constitute different alleles when compared with the genes present in the reference S288c genome. The CAT-1 genome contains a reduced number of transposable elements, as well as several gene deletions and duplications, especially at telomeric regions, some correlated with several of the physiological characteristics of this industrial fuel-ethanol strain. Phylogenetic analyses revealed that some genes were likely associated with traits important for bioethanol production. Identifying and characterizing the allelic variations controlling traits relevant to industrial fermentation should provide the basis for a forward genetics approach for developing better fermenting yeast strains.

**Keywords** Bioethanol · Genome · *Saccharomyces* · Sugarcane · Industrial strains

## Introduction

Concerns regarding the depletion, environmental impact, and security of fossil fuel sources make renewable fuel alternatives highly attractive. Fuel ethanol is an ecologically friendly, clean, and renewable alternative to gasoline that can be produced from plentiful biomass (Farrell et al.

Communicated by T. Ito.

The authors F. Babrzadeh and R. Jalili contributed equally to this work.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00438-012-0695-7) contains supplementary material, which is available to authorized users.

F. Babrzadeh · R. Jalili · C. Wang · S. Shokralla · S. Pierce · A. Robinson-Mosher · R. W. Davis · M. Ronaghi · B. Gharizadeh
Stanford Genome Technology Center, Stanford University, Stanford, CA, USA

P. Nyren
Department of Biochemistry, School of Biotechnology, KTH Royal Institute of Technology, Stockholm, Sweden

R. W. Shafer
Department of Medicine, School of Medicine, Stanford University, Stanford, USA

L. C. Basso
Biological Science Department, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, São Paulo, Brazil

H. V. de Amorim · A. J. de Oliveira
Fermentec Ltda, Piracicaba, São Paulo, Brazil

B. U. Stambuk (✉)
Departamento de Bioquímica, Centro de Ciências Biológicas, Universidade Federal de Santa Catarina, Florianópolis, SC 88040-970, Brazil
e-mail: bstambuk@mbox1.ufsc.br

2006; Goldemberg 2007). Bioenergy crops are able to offset greenhouse emissions by converting atmospheric carbon dioxide into organic carbon in biomass and soil. In theory, the same amount of carbon that is released when the fuel ethanol is burned is absorbed from the atmosphere by the next year's fuel crop through photosynthesis, thus completing the carbon cycle. The utility of fuel-ethanol production platforms depends on both their economic and environmental viability, and Brazilian sugarcane ethanol has been shown to be fully competitive with gasoline due to several technological achievements (Goldemberg and Guardabassi 2010; Leal and Walter 2010). The fermentation process in Brazil, an adaptation of the Melle-Boinot process, uses very high yeast cell densities in a semi-continuous fed-batch mode to ferment broths (cane juice and/or diluted molasses) containing high concentrations of sugar, producing high ethanol concentration with high yield (90–92 % of the theoretical maximum) and productivity (each fermentation cycle lasts 6–10 h). After centrifuging and washing in dilute sulfuric acid, the yeast cells are recycled back for each subsequent fermentation during the entire 6–9 months crop season (Andrietta et al. 2007). Thus, an ideal yeast strain for ethanol production would thrive in this harsh environment, enabling the outgrowth of contaminating species and also producing ethanol efficiently.

The predominant microorganisms responsible for the efficient production of fuel ethanol are selected strains of the yeast *Saccharomyces cerevisiae*. As with other non-sterile industrial fermentations, contamination by bacteria or other wild yeasts severely affect productivity. Indeed, the microbiological dynamics of industrial fermentors revealed a very rapid succession of yeast strains, and consequently the original "starter" yeast (usually commercial baker's yeast strains) was completely replaced by other strains in a matter of weeks (Silva-Filho et al. 2005; Andrietta et al. 2007; Basso et al. 2008). A few highly productive yeast strains then tend to dominate the fermentor during the entire production season, allowing efficient and stable fermentations. The positive impact of selected yeast strains in increasing ethanol yield and reducing production costs is due to their higher fermentation performance (including high ethanol yield and reduced glycerol formation), maintenance of high viability during fermentation and recycling, including no flocculation and foam formation, and very high implantation capability in the stressful industrial fermenters (Basso et al. 2008). Some of these selected strains became commercially available in the late 1990s, and at present, more than half of the distilleries in Brazil use one, or more commonly a mixture of two (e.g., strains PE-2 and CAT-1), of these selected strains as starters in the fermentation process, which collectively produce billions of gallons of fuel ethanol per year (Basso et al. 2008).

Uncovering the genetic elements that determine the most desirable traits for industrial fermentative microbes is essential for understanding the biology behind an effective fermentation reaction, and will provide the basis for a forward genetics approach of targeted modification to make superior ethanol producers that are more tolerant to fermentation by-products and have greater process hardiness. Genome-wide genetic approaches can certainly help to identify and understand the molecular basis of these complex traits. Since the complete genome sequencing of strain S288c (Goffeau et al. 1996), the genomes of several other *S. cerevisiae* strains have been sequenced (Brem et al. 2002; Wei et al. 2007; Borneman et al. 2008; Argueso et al. 2009; Liti et al. 2009; Novo et al. 2009; Araya et al. 2010; Dowell et al. 2010; Otero et al. 2010; Esberg et al. 2011), providing important molecular information regarding the evolutionary and ecological diversity of this key model organism. Such studies have also revealed a strong influence of human selection in the genome plasticity of the different industrial strains. However, most of the published whole-genome sequencing studies of industrial strains used haploid representatives of the strains, and only very few recent reports have analyze diploid genomes (Borneman et al. 2011; Magwene et al. 2011; Akao et al. 2011). We decided to identify sources of genomic variation in the genome of fuel-ethanol yeast strains that might be linked to (and possibly responsible for) important industrial traits. Strain CAT-1 is one of the most common strains used nowadays by the Brazilian fuel-ethanol plants, shows a very efficient fermentation capacity, especially at high sugar concentrations, and has also shown good fermentation characteristics for production of distillates from cereals, when compared with the other fuel-ethanol yeasts (Amorim-Neto et al. 2009; Pereira et al. 2010). Our analysis of the diploid genome of the *S. cerevisiae* strain CAT-1 revealed significant structural and sequence variation when compared with the genome of the reference strain S288c, a subset of which is likely associated with traits for prevalence and persistence during bioethanol industrial fermentations.

## Materials and methods

### Yeast strains and growth media

The industrial CAT-1 strain was isolated by Fementec Ltda. in 1998/1999 from *Usina VO Catanduva* located in the State of São Paulo (Brazil), and became commercially available, distributed initially by Lallemand Inc. (www.lallemand.com) from Canada and more recently by LNF

Latino Americana Ltda. (www.lnf.com.br) from Brazil (Basso et al. 2008). Yeast cells were routinely grown on rich YPD medium (1 % yeast extract, 2 % peptone, and 2 % glucose) under standard conditions (Rose et al. 1990).

### Genome sequencing

Genomic DNA was isolated from cells of strain CAT-1 with YeaStar columns (Zymo Research) according to the manufacturer's recommendations. Shotgun genome sequencing was performed using both the GS 20 and the GS FLX system from the 454 Life Sciences (Margulies et al. 2005). Contigs were obtained with automated shotgun assembly and BLASTN-based contig end joining. Since it is still a significant technical challenge to develop an assembly process that faithfully maintains the integrity of the allelic contribution from an underlying set of reads originating from a diploid DNA source (Levy et al. 2007), in this study we chose the software Newbler from 454 Life Sciences to assembly the genome sequences first and identify heterozygous alleles later by mapping reads onto contigs with a customized ASW algorithm (Wang et al. 2007). Briefly, using the ASW algorithm, sequences reads were mapped onto the assembled contigs. The contribution of each sequence read to a single position was evaluated to identify positions that contain more than one allele, including small nucleotide polymorphisms (SNP) or InDels (nucleotide insertion or deletion). Typically, more than four reads were required for the initial identification of an alternate allele where minor allele has to be present at more than 30 % reads covering that region. Regions with more than two apparent alleles represented either collapsed repetitive sequence or a group of reads with systematic base calling error, rather than true genetic variation. The assembled contigs were then compared to all sequence reads using the program BLASTN (Altschul et al. 1990) to identify reads that landed at the ends of two contigs, which was then joined together with bridging reads upon manual inspections. Using tools in the Mummer package (Kurtz et al. 2004), we further mapped these contigs to S288c reference genome sequences, including those of the *2-micron* plasmid and the mitochondrial genome (Goffeau et al. 1996). The CAT-1 genome project has been deposited at the NCBI Sequence Read Archive (SRA) (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession number SRA012578.

### Genome sequence analysis

Highly heterozygotic diploid genome sequences complicate *ab initio* gene predictions based on hidden Markov model (HMM) approach such as AUGUSTUS (Stanke et al. 2006). We chose to map S288c coding regions onto CAT-1 contigs using the modified protein-nucleotides mapping program LAP (Huang and Zhang 1996). First, S288c coding regions were translated into amino acid sequences, which were then mapped onto contigs. After that, amino acid sequences were reverse-translated into nucleotide sequences. Each predicted coding region was then manually verified. For SNP detection, tools from the Mummer package (Kurtz et al. 2004) were used to identify SNPs between any two sets of sequences compared. Briefly, we first ran the program NUCMER to determine the position and orientation of contigs in relation to the reference genome sequence, filtered the alignments to keep only one-to-one mapping between contigs and reference sequence using the program $\delta$-filter, and reported difference in these one-to-one unambiguous alignments. We reversed the order of the reference and query sequence in the previous process and only kept those SNPs reported in both processes. To construct a reliable phylogeny of the yeast strains and species analyzed, we identified SNPs between any pair of genomes and kept those shared among all of them with a customized script and generated an alignment of all species by concatenating all SNPs of each species accordingly. We then used tools of the Phylip package (Felsenstein 1989) to construct a phylogenetic tree based on the pseudo-genome of concatenated SNPs.

The unbiased 454 sequencing technique and the depth of sequencing ($\sim 26\times$) of CAT-1 genome allowed us to calibrate gene dosages and thus detect copy number variations (CNV). All sequence reads were mapped onto the S288c reference genome using the Program BLASTN (Altschul et al. 1990), and coverage for each position was then computed. Those positions with coverage greater than 1.5 times or less than 0.5 times of the average coverage of the entire chromosome were identified. The microarray karyotyping analysis of strain CAT-1 was performed essentially as described previously (Stambuk et al. 2009). To identify genes that show a significant consensus CNV relative to the genome of the sequenced reference strain S288c, the CGH-Miner program (Wang et al. 2005) was employed using the default parameters for BAC analysis. Four separate S288c self-self hybridizations were used as "normal controls" in the CGH-Miner program, and these were compared with four independent DNA samples (from independent yeast colonies) of strain CAT-1.

## Results

### Genome sequencing

A preliminary analysis (through sporulation and tetrad dissection) of the industrial strain CAT-1, showing variable growth rates of the haploid cells even on solid rich YPD medium (measured by colony size), indicated that the genome of this diploid strain probably contained several heterozygous alleles (data not shown). Thus, we decided to sequence the whole diploid genome of strain CAT-1 to uncover all possible genome variations that might be involved in the superior industrial performance of this fuel-ethanol yeast. We generated 2.5 million reads totalling 344 million bases for the diploid genome of strain CAT-1 with a shotgun approach using both the GS 20 and the GS FLX system from the 454 Life Sciences. A total of 1,211 contigs were obtained with automated shotgun assembly and BLASTN-based contig end joining. The N50 contig size is 66 kb and the largest contig size is 282 Kb. Using tools in the Mummer package, we mapped these contigs to the S288c reference genome sequences, including those of the 2-µm plasmid and the mitochondrial genome. We could not find sequences corresponding to the 2-µm plasmid in any read. In total, 11,240,885 nucleotides were placed into the S288c genome, interrupted by 426 gaps of total length 649,458 nucleotides. In this study, we mapped all reads onto de novo built contigs and recovered heterozygous alleles by examining mapping results at each position and identified 34,503 heterozygous alleles in total, a number that reflects the extensive heterozygosity of the industrial strain CAT-1 and certainly one of the highest found among diploid yeast strains (Akao et al. 2011; Borneman et al. 2011; Magwene et al. 2011). Nevertheless, these heterozygous sites were unevenly distributed among all chromosomal regions of strain CAT-1 and there was significant large scale loss of heterozygosity (LOH), for example, in the right (large) arm of chromosomes IV (Chr-IV), Chr-XII, and Chr-XV (see Supplementary Table SI), some of which have already been described in other diploid yeast strains (Akao et al. 2011; Esberg et al. 2011; Magwene et al. 2011). In the case of Chr-XII, the large LOH segment was downstream of the tandem rDNA array toward the right telomere and related to frequent DNA lesions and mitotic recombination events at this chromosomal region (Ide et al. 2007, 2010; Magwene et al. 2011).

## Genomic phylogeny of strain CAT-1

The genomic sequence divergence among the industrial fuel-ethanol production strain CAT-1 and other yeast strains is estimated to be 0.5–1 % (approaching that seen between humans and chimps), a sequence variation distributed throughout the genome. Such short evolutionary distances largely prevent us from constructing a reliable and high-resolution tree based on a few genetic markers. In this study, we instead chose to use concatenated SNPs to mimic each genome and then built a phylogenetic tree based on pseudo-genomes. We first identified pair-wise SNPs in the genome sequences of *S. paradoxus* (Kellis et al. 2003) and five *S. cerevisiae* strains with good genome sequence coverage: the reference laboratory strain S288c (Goffeau et al. 1996), the clinical strain YJM789 (Wei et al. 2007), the natural vineyard isolate RM11-1a (Brem et al. 2002), a haploid derivative (strain JAY291) from the fuel-ethanol production strain PE-2 (Argueso et al. 2009), and strain CAT-1. The number of pair-wise SNPs is listed in Table 1. Since the CAT-1 diploid genome was sequenced, we listed only homozygous SNPs present in CAT-1 to simplify the comparison with other species and/or strains. These homozygous SNPs are differences in orthologous sites between CAT-1 and the other strains, where each of those sites in the CAT-1 genome is homozygous. Note that the number of SNPs listed in Table 1 is not necessarily proportional to the distance between the two yeast strains compared, because this table only reports SNPs from reliable alignments. The number of SNPs reported is therefore related to two factors: the number of sequence fragments that can be aligned undisputedly, and the number of different sites found on those alignments. Therefore, a greater number of SNPs might indicate a closer relationship because more regions can be aligned between two distantly related species (such as *S. paradoxus* and S288c). However, we observe that the number of alignments among closely related strains is relatively constant. Therefore, while not a perfect measure, the number of SNPs reported for a pair of closely related strains such as CAT-1 and RM11-1a could be correlated

**Table 1** Number of pair-wise SNPs among *S. paradoxus* (*S. para*) and five *S. cerevisiae* strains

| Strain: | JAY291 | RM11-1a | YJM789 | S288c | *S. para* |
|---|---|---|---|---|---|
| CAT-1 | 20,331 | 21,537 | 37,386 | 36,902 | 117,182 |
| JAY291 | – | 28,922 | 50,990 | 50,079 | 136,264 |
| RM11-1a | – | – | 44,586 | 43,298 | 127,208 |
| YJM789 | – | – | – | 56,124 | 133,586 |
| S288c | – | – | – | – | 138,921 |

Only homozygous SNPs of the CAT-1 genome are counted and listed in this table with the distance between them. Concatenated SNPs, defined as polymorphic nucleotides that are flanked by a fixed number of conserved nucleotides, can be used to provide a less complex representation of the entire genomic sequence (Li et al. 2007). Trees based on concatenated SNPs have better resolution and are more robust than those trees based on alignment of short genome region, because more informative sites are sampled. In total, 92,609 SNPs were identified among the six strains and concatenated to generate an alignment of 92,609 columns. Using the program dnaml from the Phylip package, we constructed the maximum likelihood tree shown in Fig. 1. This tree placed CAT-1 in the same clade with JAY291, another strain widely used in fuel-ethanol production (Basso et al. 2008; Argueso et al. 2009), and close to the vineyard strain RM11-1a. This tree also places the clinical YJM789 isolate close to the laboratory strain S288c, which is consistent with earlier results (Wei et al. 2007).

Genome sequence analysis

As the CAT-1 genome was sequenced in its diploid format, heterozygous alleles could affect the performance of hidden Markov model (HMM)-based gene prediction programs. To take the advantage of the close phylogenetic relationship between this bioethanol strain CAT-1 and the laboratory strain S288c, we chose a similarity searching strategy to annotate the CAT-1 genome by mapping S288c protein-coding region onto the nucleotide sequences of contigs directly using a modified LAP program (Huang and Zhang 1996). Most CAT-1 genes share the same coding pattern as those of S288c, except for the mitochondrially encoded COB, BI2, BI3, and BI4 genes present in the same multigenic loci. The first and second introns of the COB gene in strain S288c are missing in CAT-1's mitochondrial genome. Indeed, the CAT-1 mitochondrial genome could be better aligned with the mitochondrial genome of the clinical isolate YJM789 (Wei et al. 2007). It is unclear whether the missing introns in the COB gene have any impact on the special industrial properties of CAT-1, but it is noteworthy that the bioethanol industrial strain JAY291 does have both introns in its COB gene. A recent survey of the genome content of industrial yeasts has also revealed great variation in the intron content of this mitochondrial gene (as well as other mitochondrial genes) among different strains (Dunn et al. 2012). A total of 20,239 heterozygous sites are found between the start and stop codons of protein-coding regions. Of these, 264 are located in introns and 3,974 (20 %) are at the first codon position, 2,936 (15 %) are at the second, and 13,065 (65 %) are at the third codon position, a distribution proportional to the known position-specific mutation pressure. Nevertheless, 3,830 (58 %) out of the 6,652 predicted protein-coding genes of the *S. cerevisiae* CAT-1 genome constitute different alleles when compared with the genes present in the reference S288c genome.

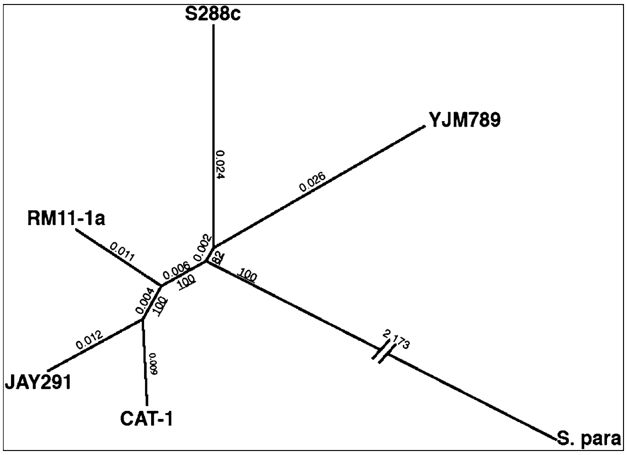

**Fig. 1** Maximum likelihood tree based on the concatenated SNPs of the *S. paradouxus* (S. para) and *S. cerevisiae* CAT-1, RM11-1a, JAY291, S288c, and YJM789 genome sequences. The *numbers* along branches indicate branch length except for those *underlined numbers*, which represent bootstrapping support out of 100 replicates

Because of the unbiased nature of the 454 sequencing technology, where every genome position has roughly an equal chance to be sequenced (Margulies et al. 2005), and the deep coverage with which the CAT-1 genome was determined, we were able to treat our data in a manner akin to comparative genomic hybridization (CGH) data and thus screen for genomic copy number variations (CNVs). We mapped all sequencing reads onto the S288c chromosomal sequences using the program BLASTN (Altschul et al. 1990) and computed the mapping coverage for each S288c position to simulate CGH-array data and look for sequences that were missing or amplified in CAT-1 relative to S288c (Fig. 2). The simulated CGH map closely matches CGH results obtained by classical microarray approach (see Supplementary Figure SI), except for a few differences. While most regions of the genome were present in equal number in comparison to the S288c genome, those regions in Fig. 2 shown in green were underrepresented and the regions shown in red represented sequences that were amplified in CAT-1 relative to S288c. As already pointed out when the genome of another bioethanol strain was analyzed, the telomeres seem to be highly variable regions in the genomes of these industrial strains (Argueso et al. 2009).

Regions near the telomeres of different chromosomes shown in green could result from either being present in only a single CAT-1 chromosome (a common situation when considering telomeric gene families), being missing in the CAT-1 genome, or being too diverged at the sequence level from their corresponding S288c sequences

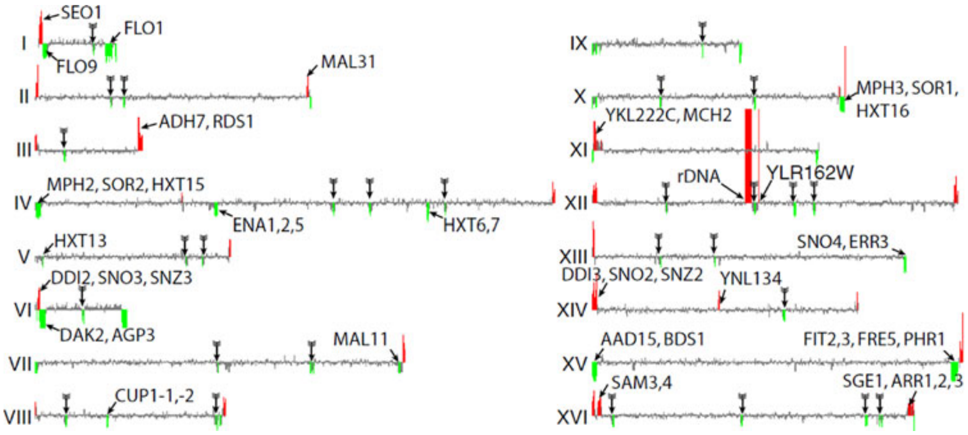

**Fig. 2** Relative gene dosage plots through mapping CAT-1 sequencing reads onto S288c chromosome sequences. Each *horizontal line* corresponds to a specific S288c chromosome (labeled I to XVI). The coverage was smoothed by a sliding window of the size 2 kb and step size 1 kb to reduce noise. *Gray* areas are regions of coverage in the range [0.5–1.5] of average coverage for the particular chromosome; *red areas* are regions of coverage in the range [1.5– +∞], and *green areas* are regions of coverage in the range [0–0.5]. S288c representative genes are marked in regions with under- or overrepresentation peaks. *Arrows* with tails denote Ty transposons. Note that the "amplification" of the ribosomal DNA (rDNA) region into the YLR162W gene (adjacent to the variant ribosomal RDN5-6 gene) shown in Chr-XII is an artifact due to only some rDNA and RDN5 genes annotated in the reference S288c genome sequence, while normally yeast strains harbor several (even hundreds) of these genes in their genome (colour figure online)

to be alignable by BLASTN. Among those S288c genes encoded in these telomeric regions underrepresented in CAT-1 are several genes belonging to gene families involved in flocculation and nutrient uptake: *FLO1* and *FLO9* in Chr-I; *MPH2*, *SOR2,* and *HXT15* in Chr-IV; and *MPH3*, *SOR1,* and *HXT16* in Chr-X; *DAK2* and *AGP3* in Chr-VI; *SNO4* and *ERR3* in Chr-XIII; and *AAD15* and *BDS1* in one arm of Chr-XV; and *FIT2*, *FIT3*, *FRE5,* and *PHR1* in the other arm. Another telomeric gene present in the S288c genome but underrepresented in CAT-1 is *MAL11* (also known as *AGT1*) in Chr-VII. Regions apparently absent in the genome of CAT-1 that were not telomeric corresponded to regions containing tandem repeated genes in S288c (*HXT6* and *HXT7*, and *ENA1*, *ENA2,* and *ENA5* in Chr-IV; and *CUP1-1* and *CUP1-2* in Chr-VIII). Since almost all industrial strains seem to have single copies of these genes, it is likely that they have expanded in the genome of the laboratory strain. Finally, there are also less transposable elements in the genome of the fuel-ethanol strain CAT-1, when compared with the genome of the reference laboratory strain S288c (Fig. 2). Regions labeled in red near the telomeres indicate genes amplified in the CAT-1 genome relative to S228c: *SEO1* in Chr-I; *ADH7* and *RDS1* in Chr-III; *DDI2*, *SNO3,* and *SNZ3* in Chr-VI; and *DDI3*, *SNO2,* and *SNZ2* in Chr-XIV; *MCH2* and YKL222C in Chr-XI; and *SAM3*, *SAM4,* and *SGE1* in Chr-XVI. In the genome of CAT-1, only one non-telomeric overrepresented gene is found: YNL134C in Chr-XIV (Fig. 2), an uncharacterized protein belonging to the quinone oxidoreductase family of medium-chain dehydrogenase/reductases (Riveros-Rosas et al. 2003).

## Identification of gene polymorphisms related to bioethanol production

Yeasts used in industrial fermentation for bioethanol production are exposed to very stressful conditions such as initial high sugar concentrations, high temperature, high ethanol concentrations at the end of the process, as well as several other stresses caused by salts, acidity, sulfite, and bacterial contamination (Andrietta et al. 2007; Basso et al. 2008). *S. cerevisiae* strains that are prevalent and persistent in industrial fermentations must bear traits to tolerate those stressful conditions. CAT-1 and PE-2 are the two *S. cerevisiae* strains most widely used in Brazil industrial fermentation and share similar performances in different industrial fermentation setups (Basso et al. 2008). The availability of JAY291 genome sequence (Argueso et al. 2009), a haploid spore of PE-2 with a close genomic phylogeny to CAT-1 as shown in Fig. 2, allows us to carry out genome-wide comparisons for genetic variants probably involved in bioethanol industrial fermentations.

An example of such approach is illustrated in Fig. 3, were it is shown the phylogenetic trees obtained for two genes (*IRA1* and *IRA2*) that participate as inhibitors of the Ras-cAMP-PKA pathway by increasing the rate at which Ras proteins hydrolyze GTP (Broach 1991). The Ras-cAMP-PKA pathway plays a central role in the regulation of the transcriptional response of yeast cells to the presence of fermentable (glucose, sucrose) sugars, inducing rapid growth and also determining the stress resistance/sensitivity of yeast cells (Thevelein and de Winde 1999; Park et al. 2005; Zaman et al. 2009). The phylogenetic tree of *IRA1* is dramatically different from that of *IRA2*, and only *IRA2* of CAT-1 and JAY291 form a tight clan. The discrepancy of *IRA1* and *IRA2* phylogenetic patterns is consistent with biochemical and genetic evidence that shows that these genes are not functionally redundant and may be involved in responding to different environmental stresses (Tanaka et al. 1990; Park et al. 2005). Indeed, a recent comparison between the transcriptional response of the laboratory strain BY4724 and the vineyard strain RM11-1a to the presence of glucose in the medium revealed that the majority of variations seen in the expression of many growth-related and energy metabolism transcripts were explained by polymorphisms in the *IRA2* gene (Smith and Kruglyak 2008; Litvin et al. 2009), indicative of a selecting pressure acting on this gene.

## Discussion

Presented here is the sequencing and preliminary analysis of the *S. cerevisiae* diploid genome of strain CAT-1, which is one of the most widely used strains in Brazil's fuel-ethanol industries. The diploid genome of CAT-1 reveals a high degree of heterozygosis, and although the fuel-ethanol strains CAT-1 and JAY291 are the nearest neighbors on the SNPs-based phylogenetic tree (Fig. 1), the *IRA1* phylogenetic tree (Fig. 3) suggests that multiplicity of events have taken place throughout the evolutionary history of these strain's genomes. This is also consistent with the difference in the mitochondrial COB gene structure found between JAY291 and CAT-1. Strain CAT-1, as well as many other industrial yeast strains currently in use in Brazil, is known to be defective in flocculation and foam production, traits important for yeast handling and process performance in the fuel-ethanol industry (Basso et al. 2008). Strain S288c is non-flocculant due to a nonsense mutation in the transcription regulator *FLO8* (Liu et al. 1996), but both CAT-1 and JAY291 have a normal coding sequence for this gene. However, in our current CAT-1 genome analysis, the flocculin genes *FLO1* and *FLO9* are both absent (Fig. 2), while other genes related to flocculation have either a large gap in the middle of their coding region (*FLO10*), or only the N-terminal (*FLO5*) or C-terminal (*FLO11*) part of the gene are present, which may explain the lack of flocculation and foam formation phenotype of this industrial strain. Both strains also show several other common CNV, and an example is the amplification of the *SAM3* and *SAM4* genes present at the left telomere of Chr-XVI. These two genes, together with another pair of physically linked telomeric genes (*MMP1* and *MHT1*), are involved in the metabolism of S-adenosylmethionine, a biochemical cofactor that participates in a variety of metabolic pathways (Rouillon et al. 1999; Vinci and Clarke 2010).

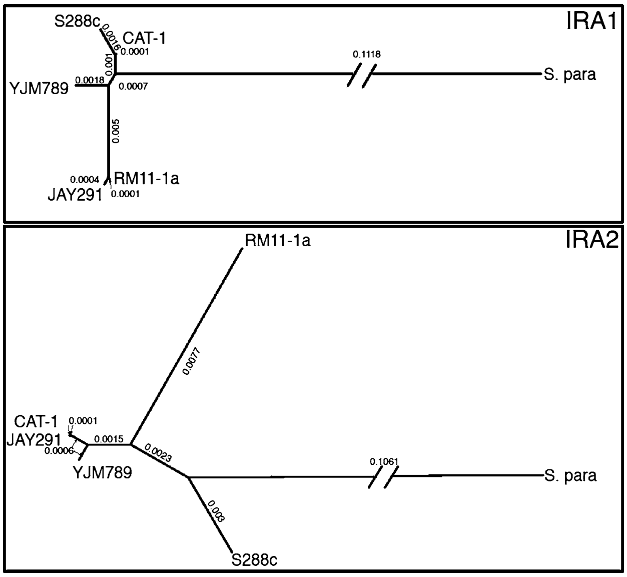

**Fig. 3** The phylogenetic relationship of *IRA1* (*upper*) and *IRA2* (*lower*) genes. The coding regions of *IRA1* and *IRA2* genes from *S. paradouxus* (S. para) and CAT-1, JAY291, YJM789, RM11-1a, and S288c were aligned with Clustalw, and the trees built with the program Phylip. *Numbers* along branch denote phylogenetic distance

Another pair of telomeric genes amplified in the Brazilian bioethanol strains are the *SNO2*/*SNO3* and *SNZ2*/*SNZ3* genes recently shown to be required for efficient sugar utilization through its effects on pyridoxine (vitamin B6) and thiamine (vitamin B1) metabolism (Stambuk et al. 2009). In both CAT-1 and JAY291, these *SNO*/*SNZ* genes are not located in the telomere of Chr-VI as in the genome of S288c (this telomeric region also lacks the *AGP3* and *DAK2* genes; see Fig. 2), but instead these genes are amplified in several other chromosomal telomeres, including Chr-X where these *SNO*/*SNZ* gene pair replace the *MPH3*, *SOR1*, and *HXT16* genes, and probably also the *MPH2*, *SOR2*, and *HXT15* genes of Chr-IV (Stambuk et al. 2009; Argueso et al. 2009). Indeed, our sequence coverage across the entire genome suggests that there is only one copy of this *MPH*/*SOR*/*HXT* cassette present on one chromosome out of the homologous pair of chromosomes in the diploid genome of CAT-1. Another gene present only in a single copy in CAT-1 is the *AGT1* α-glucoside transporter present in the right arm of Chr-VII (Fig. 2). Indeed, recent data indicate that this telomere of Chr-VII is highly heterozygous for the *MAL* locus, having both the *AGT1* and *MAL11* (identical to *MAL31* of Chr-II, which is amplified in CAT-1; see Fig. 2) genes. The CNV of the *MPH2*/*3*, *MALx1* and *AGT1* genes have been recently correlated with the ability of these industrial strains to ferment maltose and maltotriose (Alves-Jr et al. 2008; Duval et al. 2010).

Both in CAT-1 and JAY291 genomes, the *SEO1* gene, encoding a permease of the allantoate transporter subfamily (Isnard et al. 1996), and *ADH7*, encoding for a NADPH-dependent medium-chain alcohol dehydrogenase with broad substrate specificity (Larroy et al. 2002), are amplified. Other amplified genes in CAT-1 are apparently involved in the resistance to toxic materials, including the transcription factor *RDS1* involved in conferring resistance to cycloheximide (Akache and Turcotte 2002), the pleiotropic multidrug transporter *SGE1* (Jacquot et al. 1997), and the three *ARR1–ARR3* genes involved in resistance to arsenic compounds (Bobrowicz et al. 1997). The *RDS1* gene is the most heterozygous among all CAT-1 protein-coding genes, with 44 out of 75 sites non-synonymous. It is important to note that the genes present at the right telomeric region of Chr-XVI (including *SGE1* and *ARR1-ARR3*) have been shown to be also amplified in the genomes of industrial sake yeasts (Ogihara et al. 2008).

Unexpectedly, strain CAT-1 did not show amplification of the *SUC2* gene, encoding for the extracellular $\beta$-fructosidase (invertase) responsible for sucrose hydrolysis and fermentation. Widespread presence of *SUC* genes at several telomeric positions has been shown to be a common feature of both baker's and distillers' yeast strains, and postulated to be an adaptation to sucrose-rich broths (Benitez et al. 1996; Codon et al. 1998). *RTM1*, another gene usually found in association with the telomeric *SUC* genes and implicated in the resistance of yeast cells to molasses (by an unknown mechanism; Ness and Aigle 1995), is absent in the CAT-1 genome, although it has been recently shown to be present in the genome of the pathogenic strain YJM789 and some brewing yeast strains (Wei et al. 2007; Borneman et al. 2011). Since neither *SUC* gene amplifications nor the *RTM1* gene were also observed in strain JAY291 (Argueso et al. 2009), these results indicate that these genetic elements are not necessary by the selected yeast strains for bioethanol production using sucrose-rich sugarcane or molasses broths commonly used by the Brazilian fuel-ethanol industries, and opens new opportunities to further improve the ethanol yield in this important industrial process (Basso et al. 2011).

Although it is not simple to directly correlate genome sequence data to the phenotypic differences among yeast strains (Dowell et al. 2010), some examples of how the genome sequence of a relevant strain can be used for specific biotechnological applications have started to appear (Otero et al. 2010; Madsen et al. 2011). Comparative genomic sequence analysis is a powerful means of inferring functional sequences based on evolutionary constraints (Hardison 2000, 2003; Sidow 2002). With the availability of increasing number of genome sequences of industrial fuel-ethanol yeast strains as well as other yeast strains, a new paradigm could occur in the identification of genetic motifs, or at least prioritized lists of putative functional sequences selected by stressful conditions encountered by fuel-ethanol yeast strains. For example, the *IRA2* allele involved in the Ras-cAMP-PKA pathway identified in this study (Fig. 3) may contribute to increase the stress resistance of the fuel-ethanol industrial yeast cells, but it is a hypothesis that needs further experimental verification. A thorough understanding of the plasticity of industrial yeast genomes is a prerequisite for the systematic understanding of yeast physiology and the development of next-generation strains for dedicated industrial applications. While the phenotypic effects of most of the identified structural and sequence polymorphisms between CAT-1 and the reference S288c genome are largely unknown and remain to be explored, with more bioethanol yeasts to be sequenced in future studies we believe that more of those genetic motifs and alleles could be pinpointed. The genome information now available for the diploid industrial CAT-1 strain will be used not only for their possible industrial applications, including the production of new biofuels from sugarcane (including lignocellulosic bioethanol), but undoubtedly, with more fuel-ethanol yeast genomes available, the phylogenetic analysis could be more powerful to determine the most-desirable traits for biofuel production, ensuring a sustainable and renewable energy future.

**Acknowledgments** The work of B.U.S. at Stanford University was possible through a visiting fellowship (BEX2793-05-9) from CAPES, Brazil.

## References

Akache B, Turcotte B (2002) New regulators of drug sensitivity in the family of yeast zinc cluster proteins. J Biol Chem 277:21254–21260

Akao T, Yashiro I, Hosoyama A, Kitagaki H, Horikawa H, Watanabe D, Akada R, Ando Y et al (2011) Whole-genome sequencing of sake yeast *Saccharomyces cerevisiae* Kyokai no. 7. DNA Res 18:423–434

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Alves-Jr SL, Herberts RA, Hollatz C, Trichez D, Miletti LC, de Araujo PS, Stambuk BU (2008) Molecular analysis of maltotriose active transport and fermentation by *Saccharomyces cerevisiae* reveals a determinant role for the *AGT1* permease. Appl Environ Microbiol 74:1494–1501

Amorim-Neto HB, Yohannan BK, Bringhurst TA, Brosnan JM, Pearson SY, Walker JM, Walker GM (2009) Evaluation of a Brazilian fuel alcohol yeast strain for Scotch Whisky fermentations. J Inst Brew 115:198–207

Andrietta MGS, Andrietta SR, Steckelberg C, Stupiello ENA (2007) Bioethanol—Brazil, 30 years of Proálcool. Int Sugar J 109:195–200

Araya CL, Payen C, Dunham MJ, Fields S (2010) Whole-genome sequencing of a laboratory-evolved yeast strain. BMC Genomics 11:88

Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OV, Missawa SK, Galzerani F et al (2009) Genome structure of a

*Saccharomyces cerevisiae* strain widely used in bioethanol production. Genome Res 19:2258–2270

Basso LC, de Amorim HV, de Oliveira AJ, Lopes ML (2008) Yeast selection for fuel ethanol production in Brazil. FEMS Yeast Res 8:1155–1163

Basso TO, de Kok S, Dario M, do Espirito-Santo JC, Müller G, Schlölg PS, Silva CP et al (2011) Engineering topology and kinetics of sucrose metabolism in *Saccharomyces cerevisiae* for improved ethanol yield. Metab Eng 13:694–703

Benitez T, Gasent-Ramirez JM, Castrejon F, Codon AC (1996) Development of new strains for the food industry. Biotechnol Prog 12:149–163

Bobrowicz P, Wysocki R, Owsianik G, Goffeau A, Ulaszewski S (1997) Isolation of three contiguous genes, *ACR1*, *ACR2* and *ACR3*, involved in resistance to arsenic compounds in the yeast *Saccharomyces cerevisiae*. Yeast 13:819–828

Borneman AR, Forgan AH, Pretorius IS, Chambers PJ (2008) Comparative genome analysis of a *Saccharomyces cerevisiae* wine strain. FEMS Yeast Res 8:1185–1195

Borneman AR, Desany BA, 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:e1001287

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296: 752–755

Broach JR (1991) Ras-regulated signaling processes in *Saccharomyces cerevisiae*. Curr Opin Genet Dev 1:370–377

Codon AC, Benitez T, Korhola M (1998) Chromosomal polymorphism and adaptation to specific industrial environments of *Saccharomyces* strains. Appl Microbiol Biotechnol 49:154–163

Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, Bernstein DA, Rolfe PA et al (2010) Genotype to phenotype: a complex problem. Science 328:469

Dunn B, Richter C, Kvitek DJ, Pugh T, 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. doi:10.1101/gr.130310.111

Duval EH, Alves-Jr SL, Dunn B, Sherlock G, Stambuk BU (2010) Microarray karyotyping of maltose-fermenting *Saccharomyces* yeasts with differing maltotriose utilization profiles reveals copy number variation in genes involved in maltose and maltotriose utilization. J Appl Microbiol 109:248–259

Esberg A, Muller LA, McCusker JH (2011) Genomic structure of and genome-wide recombination in the *Saccharomyces cerevisiae* S288C progenitor isolate EM93. PLoS ONE 6:e25211

Farrell AE, Plevin RJ, Turner BT, Jones AD, O'Hare M, Kammen DM (2006) Ethanol can contribute to energy and environmental goals. Science 311:506–508

Felsenstein J (1989) Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M et al (1996) Life with 6000 genes. Science 274:546–563

Goldemberg J (2007) Ethanol for a sustainable energy future. Science 315:808–810

Goldemberg J, Guardabassi P (2010) The potential for first-generation ethanol production from sugarcane. Biofuels Bioprod Bioref 4:17–24

Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16:369–372

Hardison RC (2003) Comparative genomics. PLoS Biol 1:e58

Huang X, Zhang J (1996) Methods for comparing a DNA sequence with a protein sequence. Comput Appl Biosci 12:497–506

Ide S, Watanabe K, Watanabe H, Shirahige K, Kobayashi T, Maki H (2007) Abnormality in initiation program of DNA replication is monitored by the highly repetitive rRNA gene array on chromosome XII in budding yeast. Mol Cell Biol 27:568–578

Ide S, Miyazaki T, Maki H, Kobayashi T (2010) Abundance of ribosomal RNA gene copies maintains genome integrity. Science 327:693–696

Isnard AD, Thomas D, Surdin-Kerjan Y (1996) The study of methionine uptake in *Saccharomyces cerevisiae* reveals a new family of amino acid permeases. J Mol Biol 262:473–484

Jacquot C, Julien R, Guilloton M (1997) The *Saccharomyces cerevisiae* MFS superfamily *SGE1* gene confers resistance to cationic dyes. Yeast 13:891–902

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12

Larroy C, Pares X, Biosca JA (2002) Characterization of a *Saccharomyces cerevisiae* NADP(H)-dependent alcohol dehydrogenase (ADHVII), a member of the cinnamyl alcohol dehydrogenase family. Eur J Biochem 269:5738–5745

Leal MRLV, Walter AD (2010) Sustainability of the production of ethanol from sugarcane: the Brazilian experience. Int Sugar J 112:390–396

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G et al (2007) The diploid genome sequence of an individual human. PLoS Biol 5:e254

Li Y, Carroll DS, Gardner SN, Walsh MC, Vitalis EA, Damon IK (2007) On the origin of smallpox: correlating variola phylogenics with historical smallpox records. Proc Natl Acad Sci USA 104:15787–15792

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V et al (2009) Population genomics of domestic and wild yeasts. Nature 458:337–341

Litvin O, Causton HC, Chen BJ, Pe'er D (2009) Modularity and interactions in the genetics of gene expression. Proc Natl Acad Sci USA 106:6441–6446

Liu H, Styles CA, Fink GR (1996) *Saccharomyces cerevisiae* S288C has a mutation in *FLO8*, a gene required for filamentous growth. Genetics 144:967–978

Madsen KM, Udatha GD, Semba S, Otero JM, Koetter P, Nielsen J, Ebizuka Y, Kushiro T, Panagiotou G (2011) Linking genotype and phenotype of *Saccharomyces cerevisiae* strains reveals metabolic engineering targets and leads to triterpene hyperproducers. PLoS ONE 6:e14763

Magwene PM, Kayıkçı Ö, Granek JA, Reininga JM, Scholl Z, Murray D (2011) Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 108:1987–1992

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Ness F, Aigle M (1995) *RTM1*: a member of a new family of telomeric repeated genes in yeast. Genetics 140:945–956

Novo M, Bigey F, Beyne E, Galeote V, Gavory F, Mallet S, Cambon B, Legras JL, Wincker P, Casaregola S, Dequin S (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. Proc Natl Acad Sci USA 106:16333–16338

Ogihara F, Kitagaki H, Wang Q, Shimoi H (2008) Common industrial sake yeast strains have three copies of the *AQY1-ARR3* region of chromosome XVI in their genomes. Yeast 25:419–432

Otero JM, Vongsangnak W, Asadollahi MA, Olivares-Hernandes R, Maury J, Farinelli L et al (2010) Whole genome sequencing of *Saccharomyces cerevisiae*: from genotype to phenotype for improved metabolic engineering applications. BMC Genomics 11:723

Park JI, Grant CM, Dawes IW (2005) The high-affinity cAMP phosphodiesterase of *Saccharomyces cerevisiae* is the major determinant of cAMP levels in stationary phase: involvement of different branches of the Ras-cyclic AMP pathway in stress responses. Biochem Biophys Res Commun 327:311–319

Pereira FB, Guimarães PM, Teixeira JA, Domingues L (2010) Selection of *Saccharomyces cerevisiae* strains for efficient very high gravity bio-ethanol fermentation processes. Biotechnol Lett 32:1655–1661

Riveros-Rosas H, Julián-Sánchez A, Villalobos-Molina R, Pardo JP, Piña E (2003) Diversity, taxonomy and evolution of medium-chain dehydrogenase/reductase superfamily. Eur J Biochem 270:3309–3334

Rose MD, Winston F, Hieter P (1990) Methods in yeast genetics. Cold Spring Harbor Laboratory Press, New York

Rouillon A, Surdin-Kerjan Y, Thomas D (1999) Transport of sulfonium compounds. Characterization of the s-adenosylmethionine and s-methylmethionine permeases from the yeast *Saccharomyces cerevisiae*. J Biol Chem 274:28096–28105

Sidow A (2002) Sequence first: ask questions later. Cell 111:13–16

Silva-Filho EA, dos Santos SKB, Resende AM, de Morais JO, De Morais MA Jr, Simoes DA (2005) Yeast population dynamics of industrial fuel-ethanol fermentation process assessed by PCR-fingerprinting. Antonie Van Leeuwenhoek 88:13–23

Smith EN, Kruglyak L (2008) Gene–environment interaction in yeast gene expression. PLoS Biol 6:e83

Stambuk BU, Dunn B, Alves-Jr SL, Duval EH, Sherlock G (2009) Industrial fuel ethanol yeasts contain adaptive copy number changes in genes involved in vitamin B1 and B6 biosynthesis. Genome Res 19:2271–2278

Stanke M, Schoffmann O, Morgenstern B, Waack S (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7:62

Stephens C, Harrison SJ, Kazan K, Smith FW, Goulter KC, Maclean DJ, Manners JM (2005) Altered fungal sensitivity to a plant antimicrobial peptide through over-expression of yeast cDNAs. Curr Genet 47:194–201

Tanaka K, Nakafuku M, Tamanoi F, Kaziro Y, Matsumoto K, Toh-e A (1990) *IRA2*, a second gene of *Saccharomyces cerevisiae* that encodes a protein with a domain homologous to mammalian ras GTPase-activating protein. Mol Cell Biol 10:4303–4313

Thevelein JM, de Winde JH (1999) Novel sensing mechanisms and targets for the cAMP-protein kinase A pathway in the yeast *Saccharomyces cerevisiae*. Mol Microbiol 33:904–918

Vinci CR, Clarke SG (2010) Homocysteine methyltransferases Mht1 and Sam4 prevent the accumulation of age-damaged (R,S)-AdoMet in the yeast *Saccharomyces cerevisiae*. J Biol Chem 285:20526–20531

Wang P, Kim Y, Pollack J, Narasimhan B, Tibshirani R (2005) A method for calling gains and losses in array CGH data. Biostatistics 6:45–58

Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res 17:1195–1201

Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, Gu Z, Bruno D, Miranda M, Nguyen M et al (2007) Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. Proc Natl Acad Sci USA 104:12825–12830

Zaman S, Lippman SI, Schneper L, Slonim N, Broach JR (2009) Glucose regulates transcription in yeast through a network of signaling pathways. Mol Syst Biol 5:245