DOI: 10.1093/gbe/evz036

Positive Selection Evidence in Xylose-Related Genes Suggests Methylglyoxal Reductase as a Target for the Improvement of Yeasts' Fermentation in Industry

Guilherme Borelli[1], Mateus Bernabe Fiamenghi[1], Leandro Vieira dos Santos[2], Marcelo Falsarella Carazzolle[1,2], Gonçalo Amarante Guimarães Pereira[1,2,*], and Juliana José[1]

[1]Genomics and bioEnergy Laboratory (LGE), Institute of Biology, Unicamp, São Paulo, Campinas, Brazil

[2]Brazilian Bioethanol Science and Technology Laboratory (CTBE), Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, Sao Paulo, Brazil

*Corresponding author: E-mail: goncalo@unicamp.br.

Accepted: February 8, 2019

Abstract

Xylose assimilation and fermentation are important traits for second generation ethanol production. However, some genomic features associated with this pentose sugar's metabolism remain unknown in yeasts. Comparative genomics studies have led to important insights in this field, but we are still far from completely understanding endogenous yeasts' xylose metabolism. In this work, we carried out a deep evolutionary analysis suited for comparative genomics of xylose-consuming yeasts, searching for of positive selection on genes associated with glucose and xylose metabolism in the xylose-fermenters' clade. Our investigation detected positive selection fingerprints at this clade not only among sequences of important genes for xylose metabolism, such as xylose reductase and xylitol dehydrogenase, but also in genes expected to undergo neutral evolution, such as the glycolytic gene phosphoglycerate mutase. In addition, we present expansion, positive selection marks, and convergence as evidence supporting the hypothesis that natural selection is shaping the evolution of the little studied methylglyoxal reductases. We propose a metabolic model suggesting that selected codons among these proteins caused a putative change in cofactor preference from NADPH to NADH that alleviates cellular redox imbalance. These findings provide a wider look into pentose metabolism of yeasts and add this previously overlooked piece into the intricate puzzle of oxidative imbalance. Although being extensively discussed in evolutionary works the awareness of selection patterns is recent in biotechnology researches, rendering insights to surpass the reached status quo in many of its subareas.

Key words: natural selection, comparative genomics, phylogenetics, Saccharomycotina, xylose reductase, xylitol dehydrogenase.

Introduction

Second generation ethanol (2G) production from xylose available in hemicellulose fibers of plants' cell walls is one of the most relevant biotechnological themes of the last decade (Aditiya et al. 2016). The need to restructure the energy matrix aiming the mitigation of global warming and air pollution encouraged the trials for establishing economically viable 2G ethanol plants around the world. Some adjustments are still needed for this industry to reach its payback cost, being one of them the improvement of xylose consumption during fermentation (dos Santos, de Barros Grassi, et al. 2016). The use of molecular biology for engineering yeasts for industrial ethanol production and integration of Biology subareas such as genetics, bioinformatics, and evolution are providing benefits to turning engineered yeasts into biotechnological cell refineries (dos Santos, Carazzolle, et al. 2016). A recently developed approach for finding candidate genes to improve metabolic pathways and increase ethanol production is comparative genomics of different xylose-consuming yeasts (Wohlbach et al. 2011; Riley et al. 2016). The consumption of sugars such as galactose and L-rhamnose is highly predictable from the presence or absence of genes in their metabolic pathways, but xylose is an exception. Many species in which the presence of all xylose metabolism related enzymes is

documented do not consume this pentose, suggesting that the mechanisms underlying this process are more complex and may involve enzyme modifications and other genes not yet studied (Wohlbach et al. 2011; Riley et al. 2016). The genomic comparison among xylose-consumers and nonconsumers has been explored before, but the genomic basis, the regulatory network responsible for xylose fermentation and cofactor imbalance adjustment remain unsolved.

Wohlbach et al. (2011) presented a comparative genomics study of 14 ascomycete genomes searching for genes exclusive to species that consume, or ferment xylose. This comparative approach allowed the identification of some genes that improve xylose assimilation such as the gene encoding an aldo-keto-reductase (AKR) of *Candida tenuis*. Riley et al. (2016) have done a similar analysis including more species and focusing on the effects of genes presence and absence in xylose consumption. Although these two comparative works gave rise to new important insights on the field, they lack a deeper evolutionary perspective on their analysis.

Regarding an evolutionary perspective, comparing traits of species must always account for the underlying evolutionary history that connects them and define the dependency among their traits (Oleksyk et al. 2010). The use of a phylogenetic inference when comparing species' genomes is not only a starting point to the interpretations of exclusivity of genes or metabolic pathways, but it also must be considered when surmising evolutionary processes that gave rise to the pattern of exclusivity, or expansion and retraction of gene groups that we observe in comparative analysis (Han et al. 2013).

When searching for genes involved in metabolic processes, the information on its course of evolution can be a powerful tool. A gene that mainly evolved under negative selection constraints is probably developing a crucial function for the organism as a housekeeping gene (Oleksyk et al. 2010). On the other hand, a gene under positive selection might have changed toward adjusting its function to a relatively novel pressure in its environment (Manel et al. 2016). Nevertheless, the valuable information that can be obtained from studying genomic evolutionary mechanisms is barely considered in biotechnological works. Thus, we conducted a comparative genomics study on the Saccharomycotina subphylum, focusing on evolutionary analysis that can help us to understand the genomic mechanisms for xylose consumption and fermentation in yeasts. We suggest putative gene candidates for rational molecular engineering aiming an increase on xylose fermentation and we also have found evidence of positive selection in primary metabolism enzymes, a report that has not been previously described in literature. Furthermore, we present the methylglyoxal reductase (MGR) pathway as a possible evolutionary response for cofactor imbalance on xylose fermentation by the oxidative pathway in eukaryotes, representing a novel candidate for future studies. We propose that this approach can be implemented in other biotechnological studies, beyond this theme.

Materials and Methods

Data Set

We chose 17 genomes from Saccharomycotina yeasts subphylum accounting for the quality of genome assembly and gene prediction, as well as for a similar number of species in the three major subclades, as indicated in Joint Genome Institute (JGI) Mycocosm database (Grigoriev et al. 2012). We also used *Schizosaccharomyces pombe* (subphylum Taphrinomycotina) as an outgroup for the analysis. The detailed description of the used genomes is shown in table 1.

Yeasts were chosen accounting for two phenotypes: their capacity of naturally metabolizing xylose and capacity of naturally fermenting xylose. For the first phenotype we used the Centraalbureau voor Schimmelcultures (CBS-KNAW) collection database description for xylose consumption as a binary trait. Yeasts' classification of each phenotype as positive/negative for this trait is presented in table 1. In some cases, the phenotype may vary within different strains of one species and was noted as +/−, or −/+ depending on if the used strain in this work was considered +, or − respectively, also using specific strain annotation of CBS-KNAW. For xylose fermentation phenotype, we used previous works of the area describing such trait, being *Spathaspora passalidarum*, *Scheffersomyces (Pichia)* stipitis, and *C. tenuis* the three yeasts a priori positive for this phenotype (Wohlbach et al. 2011).

Gene Families and Homology Assignment

Homology assignment of genes in families used a Markov clustering as implemented in OrthoMCL (Li et al. 2003), defining putative orthologs and paralogs according to the similarity in functional groups, named gene families. Using homologous gene families as a starting point, precise identifications of ortholog genes used phylogenetic inferences as described next.

Phylogenetic Inferences

Phylogenies were inferred for species, in a phylogenomic analysis, and for some gene families individually. Phylogenomic analysis used 1,255 single-copy ortholog proteins aligned with multiple alignment algorithms implemented in MAFFT (Katoh and Standley 2013) using the iterative refinement method with WSP and consistency scores (G-INS-i). All ortholog individual alignments were concatenated in a super matrix (Kück and Longo 2014) for the maximum likelihood phylogenetic inference in RAxML v8 (Stamatakis 2014) with the GTR+GAMMA model of substitutions, and 5,000 bootstrap replicates for branch support.

For homolog gene families, the gene phylogenetic history was inferred using coding sequences, manually checked for mispredictions, aligned in a codon-based alignment in MACSE (Ranwez et al. 2011) to avoid frame shifts. For each analyzed gene family, models of substitutions were evaluated

Table 1
List of Chosen Species for Comparative Analysis

Species	Prefix[a]	Xyl[b]	Genome Size (Mb)	# Genes	Source	Collections[c]		References
						ATCC	CBS	
Blastobotrys (Arxula) adeninivorans	Aa	+	11.8	6,119	JGI	Y-17692 K	8244	Kunze et al. (2014)
Candida arabinofermentans	Ca	+	13.2	5,861	JGI	YB-2248	8468	Riley et al. (2016)
Candida boidinii	Cb	+	19.4	5,978	NCBI	Y-2332 K	2428	Borelli et al. (2016)
Candida sojae	Cs	+	11.9	5,231	NCBI	Y-17909 K	7871	Borelli et al. (2016)
Candida tanzawaensis	Cn	−	13.1	5,895	JGI	Y-17324 K	7422	Riley et al. (2016)
Candida tenuis	Ce	+	10.7	5,533	JGI	Y-1498 K	615	Wohlbach et al. (2011)
Candida troplicalis	Ct	+	14.6	6,254	NCBI	Y-12968 K	94	Wohlbach et al. (2011)
Dekkera bruxellensis	Db	−	13.4	5,636	JGI	Y-12961 K	74	Piškur et al. (2012)
Kazachstania africana	Ka	−	11.1	5,378	JGI	Y-8276	2517	Gordon et al. (2011)
Kluyveromyces lactis	Kl	+/−	10.7	5,076	JGI	Y-8279 K	683	Dujon et al. (2004)
Lipomyces starkeyi	Ls	+	21.3	8,192	JGI	Y-11557 K	1807	Riley et al. (2016)
Komagataella (Pichia) pastoris	Pp	−/+	9.2	5,040	JGI	Y-1603 K	704	De Schutter et al. (2009)
Scheffersomyces (Pichia) stipitis	Os	+	15.4	5,841	JGI	Y-7124 K	5773	Jeffries et al. (2007)
Saccharomyces cerevisiae S288C	Sc	−	12.1	6,575	JGI	Y-12632 N K	1171	Goffeau et al. (1996)
Schizosaccharomyces pombe	So	−	12.6	5,134	JGI	Y-12796 K	356	Wood et al. (2011)
Spathaspora passalidarum	Sp	+	13.2	5,983	JGI	Y-27907	10155	Wohlbach et al. (2011)
Wickerhamomyces anomalus	Wa	+	14.1	6,423	JGI	Y-8168	5759	Riley et al. (2016)
Yarrowia lipolytica	Yl	−	20.5	6,447	JGI	YB-423 K	6124	Dujon et al. (2004)

[a]Species names prefixes.
[b]Information about xylose consumption retrieved from CBS-KNAW website (+ stands for detected xylose consumption, − for no detection, +/− and −/+ for variation among strains, being the first sign the description of the used strain).
[c]Genome attributes and yeasts type strains of two international collections.

in jModelTest2 (Darriba et al. 2012), and the best fits were chosen using Akaike information criteria (Posada and Buckley 2004). The gene phylogenies were inferred using Bayesian methods in MrBayes v3.2 (Ronquist et al. 2012) and BEAST v2 (Bouckaert 2014). Two independent runs of "Metropolis-coupled Monte Carlo Markov Chain" (MCMCMC), each one with two cold and four hot chains, were analyzed for 10 million generations, sampled each thousand generations. The convergence of chains was visually determined, using the software TRACER v1.6 (Rambaut et al. 2018).

For d*N*/d*S* estimates for gene families, we used the species phylogeny when the gene family had no paralogs, and the gene family phylogeny excluding the third base of codons when families had paralogs. This was an attempt accounting for paralogs in analysis but minimizing the effect of substitutions saturation on gene copies that may increase the difference between the gene tree and the species tree. As expected, gene phylogenies that highly differ from species' phylogeny became similar to the species' phylogeny when the third bases from codons were excluded.

Expansion and Retraction of Gene Families

To analyze changes in gene family size in a way that accounts for phylogenetic history and provides a statistical foundation for evolutionary inferences of expansion and retraction, we used the software BadiRate (Librado et al. 2012) that uses birth and death processes to model gene gain and loss across a species' phylogenetic tree, with the following parameters: free-rate branch model (assumes each branch has its own turnover rates); maximum likelihood estimation procedure; birth, death and innovation turnover rates, whereas reporting families that most likely have not evolved under the estimated stochastic process.

Detection of Positive Selection

Evidences for evolution guided by natural selection were searched in gene families of specific proteins on pathways related to xylose consumption. We chose 33 proteins from xylose, glucose, glycerol, methylglyoxal, ethanol, and pentose phosphate pathways. To define the gene families of each protein, we used *Saccharomyces cerevisiae* (Sc) annotated sequences as baits in a BlastP search against yeast proteins already grouped in families. Families including Sc proteins identical to baits were used. Other random 628 gene families of single-copy orthologs were used to calculate an estimated omega distribution for potentially housekeeping genes.

The evolutionary models of positive selection were tested using the ratio of the number of nonsynonymous substitutions per nonsynonymous site (d*N*) to the number of synonymous substitutions per synonymous site (d*S*) (Nei and Gojobori 1986; Nielsen and Yang 1998). The hypothesis that d*N*/d*S* > 1 was tested against the null (d*N*/d*S* = 1), and the negative selection hypothesis (d*N*/d*S* < 1) using model fitting and likelihood ratio test implemented in the codeml

package of PAML4 (Yang 2007). Tests were carried with site models for the whole phylogeny (model 0, NSsites = 0 1 2), and branch-site models (model 2, NSsites = 2) using a clade containing six species (the fermenter's clade—explained in results) as foreground against the remaining clades as background; foreground d*N*/d*S* values were compared against the background assuming foreground sites evolved under positive selection while the background evolved under negative selection (2a) or neutral evolution (2b), if foreground d*N*/d*S* on any of these cases was >1, we checked which sites underwent positive selection, as reported by Bayes Empirical Bayes (BEB) analysis (Yang and Nielsen 2002; Yang 2005; Zhang 2005). In either case, we used the maximum-likelihood estimate of transition/transversion ratio (κ) and an F61 codon substitution model (Yang and Bielawski 2000).

Additionally, a control experiment was done to verify if detected signals of selection could be due to random chance. We selected at random 6 species from the total of 18 to serve as foreground for branch-site tests, reproducing the number of species on fermenter's clade. This species' randomization was done 100 times and used for d*N*/d*S* tests (model 2, NSsites = 2) in each of the 11 families that showed evidences of positive selection for fermenter's clade in results. Species from fermenter's clade were allowed to be selected during randomization as an attempt to check if positive selection signal can still be detected when fewer fermenting species are present.

Results

Gene Families and Homology Assignment

From the 18-species' gene-sets clustered in OrthoMCL, 8,679 gene families were assigned as putative homologs, being 1,255 families composed by ortholog single-copy genes, 2,740 families composed by paralogs shared by all species and 980 gene groups composed by exclusive genes to only 1 species.

Phylogenomics

Using the whole set of 1,255 single-copy orthologs genes (583,536 amino acids positions) among all 18 genomes, the species' phylogeny was reconstructed resulting in species relationships consistent to ones estimated in previous works. We obtained strong branch supports given the large gene data set we used (fig. 1). The phylogenetic positioning of the two *Candida* genomes that we assembled supported once more that this genus is paraphyletic, because *Candida boidinii* is positioned in a different clade than *Candida sojae* and *Candida tropicalis*. *Candida sojae* is placed in the clade that includes three species of well-known xylose-fermenter yeasts and we have previously observed that our isolated strain of this species is a particularly good xylose consumer (Borelli et al. 2016).

Species that are described as xylose-consumers were mapped in the phylogenetic tree and showed a nonclustered distribution among species (check fig. 1 for species and their acronyms), supporting that the ability to consume xylose is not a phylogenetic-dependent trait. Nevertheless, we have searched for gene families composed by genes strictly from xylose-consuming yeasts. We were not able to find any genes fulfilling this condition, but performing a jackknife resample (i.e., resampling leaving out each species at a time) we found family3716 (sequences in Supplementary Material online) in most of consumer species, composed by single-copy genes for eight species (Aa1514, Ca3709, Cb3443, Ce1861, Pp217, Ps5350, Sp3929, Wa4240) and duplicated genes for three species (Cs2534, Cs2538, Ct4213, Ct4214, Ls2898, Ls927). This family was characterized as transmembrane transporter proteins of the Major Facilitator Superfamily (IPR011701), but none of its 13 genes resulted in annotated hits in BLAST searches against NR database. In InterPro searches, all 13 genes were characterized with 12 transmembrane domains, and Gene Ontology terms of transmembrane transport and integral component of the membrane.

Although we could not identify a phylogenetic pattern for xylose-consumers, three-known species that ferment xylose form a monophyletic group (highlighted box in fig. 1) suggesting that their most recent common ancestor might have had some preadaptations to xylose fermentation. This clade was deeply investigated as shown in the ensuing topics and named hereafter fermenters' clade. Within this clade's species, we identified 5,883 gene families from which 3,491 families are shared by all of them and 209 of these were exclusive to the group. The three xylose-fermenting species (Sp, Ps, and Ce) shared exclusively four gene families, being fam6583 functionally annotated as glycosyl-hydrolases from CAZy's family GH115, whereas the other three (fam6584, fam6600, and fam6609) were all functionally annotated as hypothetical proteins. This GH115 protein has a 4-O-methyl α-glucuronidase action that assists other endo-xylanases to reach the main chains of hemicellulose, a very important application to supplement enzymatic cocktails used in lignocellulosic hydrolysis in industrial processes, as previously described (Kolenová et al. 2010). Because these enzymes had already been tested by Bajwa et al. (2016), we went no further in their investigation.

Gene Gains and Losses in Yeasts Evolution

Analysis of expansion and retractions of gene families revealed more gene gain than loss throughout Saccharomycotina's yeasts evolution (fig. 2). Except for *Yarrowia* and *Blastobotrys* clade that shows only gain overcounting losses, the other clades showed both gene gains overcounting losses and gene losses overcounting gains. The most recent common ancestor of the xylose-fermenters' clade presents gene gains overcounting losses, also two gene families that significantly expanded in these strains that could be related to the group's preadaptations for fermenting

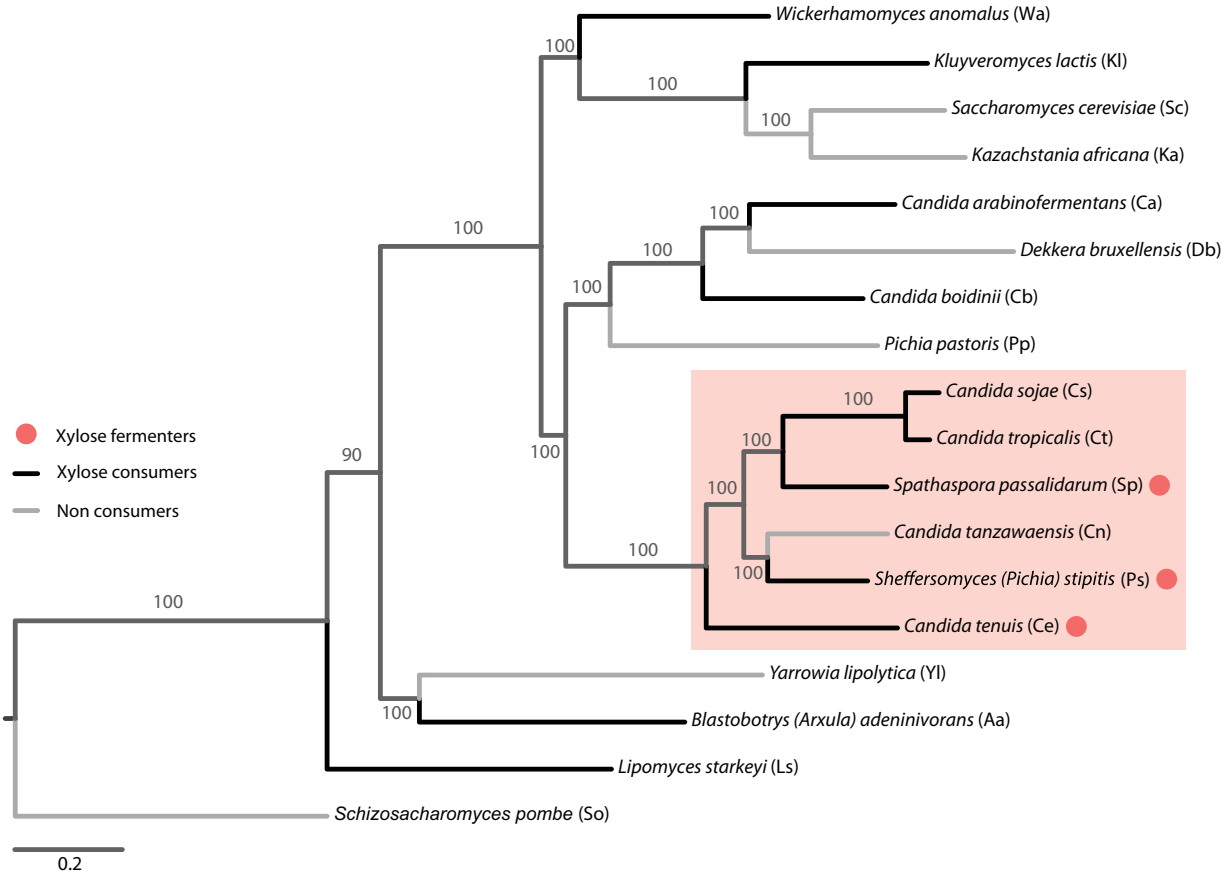

Fig. 1.—Maximum likelihood species phylogenetic inference for 18 yeasts. *S. pombe* was used as outgroup of the 17 species from Saccharomycotina subphylum. Tree was reconstructed in RAxML using 1,255 single-copy ortholog groups and branch supports were obtained using 5,000 bootstrap randomizations. The group referred to as fermenters' clade is highlighted. Phylogeny is scaled in substitution per positions.

xylose. These two gene families are family16 and family167, the first one composed by proteins of predicted flocculin function and the second composed by Nicotinamide adenine dinucleotide phosphate (NADPH)-dependent MGR proteins that are exclusive to the fermenters' clade, except for one copy also appearing on the external group, *S. pombe* (fig. 2). Another highlighting feature of these families is that the flocculin family showed higher gene gains in *C. tenuis* (10 genes), and the MGR showed higher gene gains in *S. passalidarum* (17 genes) and *Scheffersomyces stipitis* (5 genes), the only 2 yeasts described with ability to naturally ferment xylose with alleviated redox imbalance. Family16 of flocculins is also significantly expanded in branches out of fermenters' clade, as in *C. boidinii*, *Dekkera bruxellensis*, *S. cerevisiae*, and *Kluyveromyces lactis*. Because these proteins are not exclusively expanded on the fermenters' clade, and were found inserted in a genomic region rich in transposable elements with high chance of pseudogenization (Tutar 2012), we did not analyze this family further. However, family167 was deeply investigated as presented hereafter.

Methylglyoxal Reductases Expansion in Fermenter's Evolution

Although annotating gene families we found that MGRs were present not only in family167, but also in family7. Their potential homolog relationship was investigated using all proteins from both families in a phylogenetic inference. Based on the resulting phylogeny we suggest that these families are homologs and family167 is positioned as a subgroup that diverged from family7 ancestral proteins (fig. 3). The two phylogenetic recent common ancestors between family167 and family7 proteins, one for *S. passalidarum* and the other for *S. stipitis* genes, support the interpretation that family167 diverged twice from family7 ancestors, resulting in the paraphyletic family167 group (fig. 3). Besides probably being paraphyletic, the overall similarity of protein sequences among family167 allowed the Markov Clustering algorithm (MCL) algorithm to cluster fam7 and fam167 in different families. We suggest that the high similarity among family167 paraphyletic groups is a result of convergent evolution in their proteins, evidenced

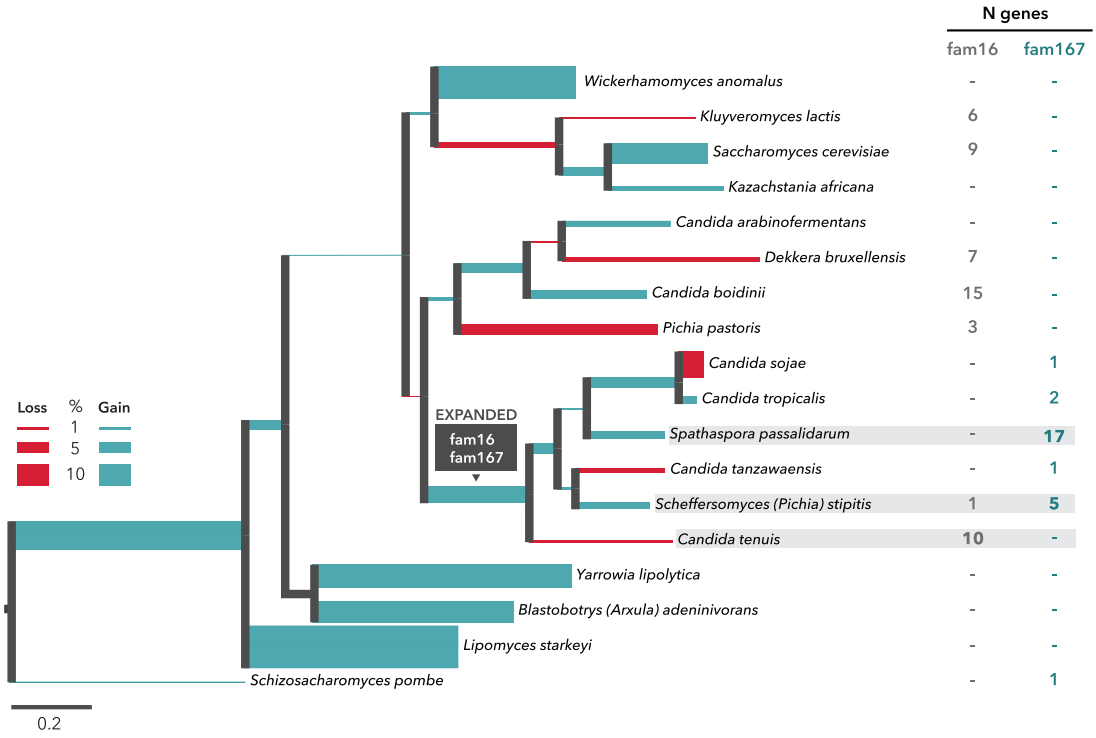

Fig. 2.—Rate of gene gains and losses along species of Saccharomycotina. Inferences were made by maximum likelihood estimates based on gene birth and death models of evolution. Branch rates were calculated as the difference ratio between gene gains and losses to its inferred ancestor gene number. The tree branches' width and colors indicate the percentage of gene gain/loss as shown in caption. Two families were significantly expanded in the fermenters' ancestral branch, shown by the dark gray box. The actual number of genes is indicated in the side table. Light gray box highlights xylose-fermenters. Phylogeny is scaled in substitutions per site.

not only by similarity, but also by shared amino acids substitutions in family167 sequences that converged from different ancestral family7 amino acids, such as at the alignment positions highlighted in figure 3.

Evidences for Positive Selection in Genes of Xylose and Glucose Pathways

We searched for selection clues on 33 proteins involved in the xylose-related metabolic pathways. Many proteins were associated to one gene family, but some returned high similarity to more than one family, such as xylitol dehydrogenase (XDH), aldehyde dehydrogenase (ALD), and alcohol dehydrogenase (ADH), resulting in a total of 48 gene families in our initial OrthoMCL analysis (table 2; see supplementary table 1, Supplementary Material online, for further information about *S. cerevisiae* bait genes, supplementary table 2, Supplementary Material online, for protein IDs, and supplementary table 3, Supplementary Material online, for the protein sites with evidences of positive selection).

The site model on d*N*/d*S* calculations considered the evolution of the protein in the whole group of yeasts and returned low omega values for all families, but some gene families still showed higher values than observed in conserved ortholog genes (0.05 ± 0.02) suggesting an evolution partially guided by positive selection. The branch-site model on d*N*/d*S* calculations indicated 11 families with high probability (>0.95) of having evolved under a positive selection model on the fermenter's clade, relative to a neutral or nearly neutral evolutionary model on the remaining branches of the phylogeny.

Regarding the control analysis for d*N*/d*S*, the randomizations sorted yeasts species including the xylose-fermenters. In this conservative manner, if no positive selection signal was detected even with a subgroup from the sorted species being of fermenter yeasts, the signal might probably be detected just by the exact combination of the fermenter's clade representants. From the 11 families in which we have detected positive selection signals, the randomized tests returned no detection for 8 of them.

For family 175, 16% of the control tests showed some sites with evidence of positive selection, but only a few sequence positions matched the ones found when testing the fermenter's clade, whereas the rest of them were consistent amongst the 16 control tests. The control analysis of family 35 presented 90% of groups containing many sites with evidence of positive selection. Most of the retrieved sites in family 35 randomizations appeared in almost all control tests and were equal to many sites found in codeml site (model 0) analysis, probably reflecting protein sites that highly diversified on yeasts evolution but lacking relationship with xylose fermentation.

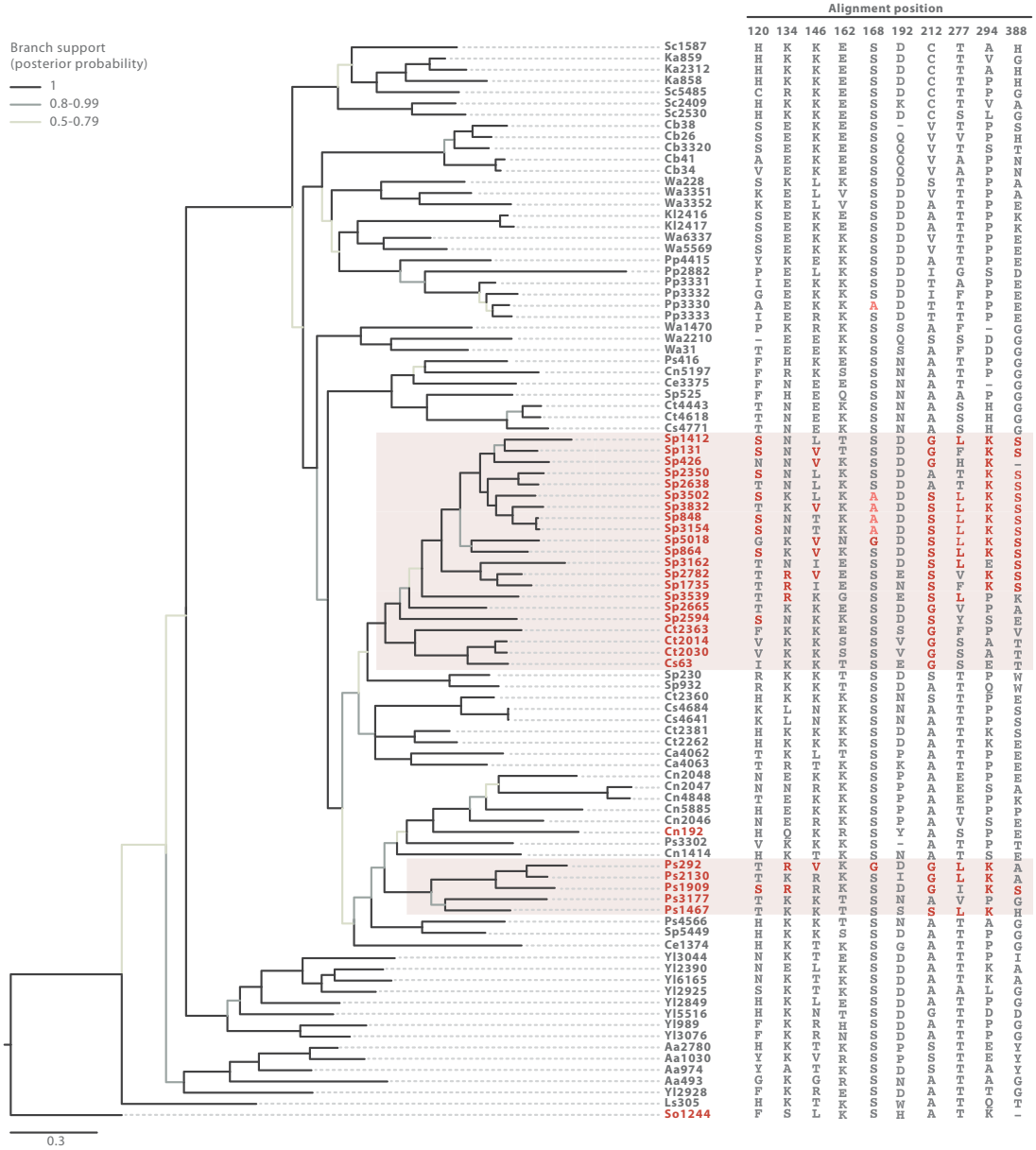

Fig. 3.—Methylglyoxal Bayesian gene tree inference (fam7 and fam167). Columns show amino acid alignment for sites which have evidence for positive selection or parallel evolution. Highlighted boxes represent clades with positive selection evidence on fermenters' group. Red gene names indicate gene copies from fam167, and red amino acids indicate that for this gene the respective amino acid is distinct from others at the alignment site. Amino acids with convergence between families on the alignment are shown in orange. Branches are colored relative to their posterior probabilities. Tree is scaled by substitutions per site.

The last family that showed some positive selection evidence in control tests was the methylglyoxal families 7 and 167. Six percent of control tests presented evidences, but all of the six sorted groups contained at least two representatives of the xylose-fermenters' clade, and all of them retrieved different sites from each other and from the obtained sites with evidence when the fermenters' clade was tested in branch-site model. These findings reinforce our hypothesis that the sites with positive selection evidenced in the branch-site analysis of fermenter's clade are not due to chance and may be related to the clade preadaptations and/or side effects to xylose fermentation.

Selection on Xylose Reductase and Xylitol Dehydrogenase Genes

To test for selection fingerprints on xylose reductase (XR) and XDH genes, we first searched for their gene families on the

Table 2

Characterization of Enzymes from Different Metabolic Pathways Associated with Xylose Fermentation

Metabolic Pathway	Enzyme	Family[a]	Omega (d*N*/d*S*)		Number of Genes[b]																
			m0	m2[b]	Wa	Kl	Sc	Ka	Ca	Db	Cb	Pp	Cs	Ct	Sp	Cn	Ps	Ce	Yl	Aa	Ls	So
Xylose	Xylose reductase (XR)	634	0.13	18.64	1	1	1	1	1	1	1	1	0	2	2	1	1	1	1	1	2	0
	Xylitol dehydrogenase (XDH)	33	0.12	1.00	1	2	2	2	1	1	2	1	3	2	2	3	3	4	5	2	3	2
		34	0.08	1.00	6	1	1	1	2	4	2	1	2	2	2	3	2	3	2	3	2	2
	Xylulose kinase (XK)	578	0.06	1.00	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1
Glycose	Glucokinase (GLK)	436	0.06	1.00	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Glucose-6P-isomerase (GPI)	2728	0.08	11.32	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
	6-Phosphofructokinase (PFK)	86	0.07	1.00	2	2	2	3	2	2	2	2	2	2	2	2	2	2	1	1	1	1
	Fructose-1, 6-bisphosphatase (FBP)	1995	0.04	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Fructose-bisphosphate-aldolase (FBA)	828	0.14	999.00	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
	GA3P-dehydrogenase (TDH)	160	0.12	1.00	1	2	3	3	1	2	1	1	0	1	1	1	3	2	1	1	1	2
	Phosphoglycerate kinase (PGK)	570	0.11	3.01	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Phosphoglycerate mutase (PGM)	807	0.05	3.01	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		1849	0.04	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Phosphopyruvate hydratase (ENO)	175	0.11	5.08	2	1	5	2	1	1	1	1	1	2	1	1	1	1	1	1	1	2
	Pyruvate kinase (PYK)	728	0.09	1.00	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Glycerol	Glycerol kinase (GUT)	2636	0.04	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	Glycerol 3-phosphate (GPP)	3186	0.07	1.00	1	1	2	2	0	0	0	0	1	1	1	2	1	1	0	2	1	0
	Glycerol 3P-dehydrogenase (GPDH)	159	0.08	1.00	2	1	2	2	1	1	1	1	1	2	2	2	2	2	1	1	1	2
Methylglyoxal	Triose-phosphate isomerase (TPI)	2006	0.11	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Lactoyl-glutathione-lyase (GLO1)	2421	0.05	1.00	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
	Hydroxyacylglutathione hydrolase (GLO2)	425	0.03	1.00	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
	D-lactate dehydrogenase (DLD)	45	0.05	1.00	5	3	1	0	3	4	3	1	2	2	1	2	3	3	1	2	1	0
		556	0.06	1.00	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Methylglyoxal reductase (MGR)	7	0.19	1.00	8	2	4	3	2	0	5	6	3	6	4	7	3	2	9	4	1	0
		7_167	0.11	10.95	0	0	0	0	0	0	0	0	1	2	17	1	5	0	0	0	0	1
Ethanol	Pyruvate decarboxylase (THI–PDC)	64	0.09	1.00	1	1	4	3	2	1	2	1	2	3	2	3	2	2	1	2	1	2
		1974	0.07	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Aldehyde dehydrogenase (ALD)	35	0.13	186.16	5	3	3	3	2	3	2	3	1	3	1	2	1	1	2	1	3	1
		128	0.07	1.00	2	2	2	1	3	3	3	1	1	1	1	1	1	1	2	2	1	1
		223	0.10	1.00	2	1	1	1	1	1	1	1	2	2	1	1	1	1	4	2	0	0
	Alcohol dehydrogenase (ADH)	23	0.12	1.00	5	4	4	2	3	3	3	1	2	4	1	2	2	1	4	1	1	1
		149	0.08	10.13	1	1	1	4	3	2	5	1	1	1	1	2	2	2	0	0	1	0
		6178	0.13	–	0	0	1	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
		508	0.06	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2

PPP (oxidative)	Glucose-6-phosphate dehydrogenase (ZWF)	836	0.09	1.00	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	6-phosphogluconolactonase (SOL)	376	0.06	1.00	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		3366	0.09	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
	6-phosphogluconate dehydrogenase (GND)	422	0.08	1.00	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PPP (nonoxidadive)	Ribose-5-phosphate isomerase (RPI)	296	0.04	1.00	1	1	1	1	1	1	3	2	1	1	1	1	1	1	1	1	1	1
	Ribulose-phosphate 3-epimerase (RPE)	3347	0.06	1.00	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1
	Transketolase (TKL)	79	0.12	1.00	2	1	2	2	3	1	3	3	1	1	1	1	2	3	1	3	3	1
	Sedoheptulose-7-phosphate (TAL/NQM)	1014	0.10	15.14	1	1	2	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
PPP (accessory)	Ribose phosphate diphosphokinase (PRS)	540	0.03	1.00	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1
		824	0.03	1.00	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		2251	0.04	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		2882	0.02	32.70	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	0	0
	Ribokinase (RBK)	1802	0.05	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Control[c]	628 single-copy orthologs		0.05 (+/−0, 02)	1.00 (95% = 1)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

[a]Gene groups retrieved from OrthoMCL analysis.
[b]Values for CODEML model 2 and gene copies are shown in a color gradient from low value/copies (gray) and high value/copies (red).
[c]Omega mean and standard deviation for model 0, and mean and fourth quartile for model 2, calculated for 628 single-copy orthologs families.

yeasts' genomes. Many families returned high-identity scores with both proteins. Besides, the search for XDH families was troublesome because *S. cerevisiae* XDH was attributed as an orphan gene in our OrthoMCL analysis.

Because BLAST comparison among known genes of XR and XDH returned too many favorable hits, we decided to use a PFAM Hidden Markov Model for the AKR family of proteins (PF00248) in a search for genes on the 18 genomes using an E-value cutoff threshold of $1e^{-20}$. This first search resulted in the detection of 861 putative AKR proteins organized in 40 different gene families. To reduce our protein data set, we have used sequences of well described XR and XDH genes from *S. passalidarum* and *S. stipitis* as queries in BLAST to retrieve the most similar sequences and their assigned families. These similar sequences were part of the three putative XR families (fam634, fam776, fam2649) and four putative XDH families (fam33, fam34, fam218, fam3425), which were further analyzed. A maximum likelihood phylogenetic inference was made for all proteins in these XR and XDH families to investigate the presence of monophyletic groups. The resulting tree is summarized in figure 4 depicting the relationships of proteins within families and highlighting relationships among families. The nonmonophyletic clustering for each XR and XDH families supports the hypothesis that those genes are in fact homologous. Based on identities with bait sequences, we defined family634 as the main XR family and families 33 and 34 as the main XDH families. In fam634, *Sp* shows a tandem duplication previously identified by Cadete et al. (2016) of the XR gene with a most recent common ancestor with *S. stipitis* gene than to the rest of the clade (fig. 5). In this gene family, we observed a general site model for d*N*/d*S* coupling with strong negative selection along the protein but significant evidence for positive selection were also observed for three codons in fermenters yeast proteins (fig. 6).

For XDH gene families we also observed gene trees that differ from the species' tree. In family33, the gene tree indicates that an ancestor duplication gave rise to two groups of orthologs in yeasts: one composed by a unique copy in *Sp*, *Ps* and *Ce*, with a duplication in *Cs*, and the other group composed by one copy in *Sp*, two in *Ps*, and three in *Ce* (supplementary fig. 1, Supplementary Material online). The general site model for d*N*/d*S* in fam33 indicated positive selection, and the branch-site models indicated high probability of neutral evolution for the fermenters' clade (table 2). An ancestral gene duplication giving rise to two groups of orthologs was also observed in family34, followed by a duplication in *Ce* (Ce5477, Ce2184) that did not diverge as with other fermenter species (supplementary fig. 2, Supplementary Material online). Differently from family33, family34 showed an overall pattern of general site model for d*N*/d*S* coupling with strong negative selection along the protein (table 2 and supplementary fig., Supplementary Material online). For branch-site models, the fermenters' clade showed evidence

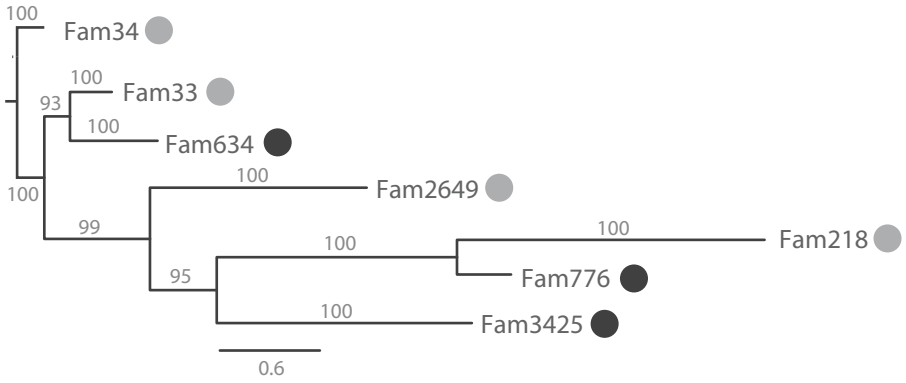

Fig. 4.—Phylogenetic inference for 179 genes (not shown) from 3 gene families of putative xylose reductase (black dots) genes and 4 families of xylitol dehydrogenase genes (gray dots), carried out with maximum likelihood inference on RAxML with branch support obtained with 1,000 bootstraps. Tree is scaled in substitutions per site.

of neutral evolution for family34, as observed in family33 (table 2).

Selection on Methylglyoxal Reductase Genes

Besides the adaptive evidence on family167 of MGR revealed by the significant expansion of this family on fermenter's clade and the convergence in sequence after two independent divergences of these proteins, we also found evidence for positive selection acting on its proteins. When comparing the omega values for the overall phylogeny in a site model for family7 and family167, family7 shows a higher value than family167, but with no specific site significantly fitting positive selection. However, when a branch-site model is tested using fam167 proteins on fermenter's clade as foreground against the remaining fam7 and fam167 branches, a very high omega value was obtained (10.95—table 2), and two specific amino acids significantly fit positive selection (figs. 3 and 6).

Discussion

Xylose Fermentation on Yeasts Evolution

The phylogenomic inference on Saccharomycotina subphylum of yeasts was recently investigated by Shen et al. (2016) and is upheld by our phylogenomic inference for the 18-yeast species. The only incongruence relies on the relationship among *Candida tanzawaensis* and *S. stipitis* that clustered together in our results, but were supported in recent analysis to have diverged before *S. stipitis* and *S. passalidarum* (Riley et al. 2016; Shen et al. 2016). All recent phylogenetic inferences, including the one we present here, indicate *S. stipitis*, *S. passalidarum*, and *C. tenuis* organized in a monophyletic group recurrently being called fermenters' clade (Cadete et al. 2012; Urbina and Blackwell 2012). Although there are other yeasts capable of fermenting xylose, for example, *Pachysolen tannophilus* (Riley et al. 2016), these three species are long-date described and have their genome available, also being *S. passalidarum* and *S. stipitis* the most efficient natural xylose-fermenting yeasts currently known (Veras et al. 2017).

The lack of a phylogenetic pattern for xylose consumption supports that the molecular mechanisms underlying it are diverse. However, xylose fermentation seems to have evolved less frequently than its consumption, and some phylogenetic pattern exists. Previous phylogenetic analysis hypothesized the fermenters' clade ancestor as a xylose-fermenter yeast, even though most of the current lineages are not capable of its fermentation (Urbina and Blackwell 2012). As this clade's diversity is still not deeply evaluated at a genomic level, on our analysis we assume that at least some preadaptations to xylose fermentation may have risen in its ancestors' evolution and investigated selection fingerprints on the whole clade that might be related either directly or as a by-product to the xylose fermentation phenotype.

Although two comparative genomics studies have been recently published (Wohlbach et al. 2011; Riley et al. 2016), none of them deeply investigated the role of natural selection on the evolution of xylose metabolism. Thus, we present evidence for the first time that positive selection might have shaped the evolution of xylose-fermenter yeasts. We have found positive selected codons on XR and other genes related to energy producing metabolic pathways, gene expansion, also convergent evolution and positively selected codons on the MGR genes in the most recent common ancestor of the fermenters' clade. Besides the evolutionary findings, proteins that were identified under positive selection might be new targets for experimental validation and use in rational engineering for biotechnological applications.

Evolution of Xylose-Related Metabolism

We present evidence that 12 enzymes on xylose-related pathways evolved under some positive selection on Saccharomycotina (XR, XDH, FBA, TDH, PGK, ENO, TPI,

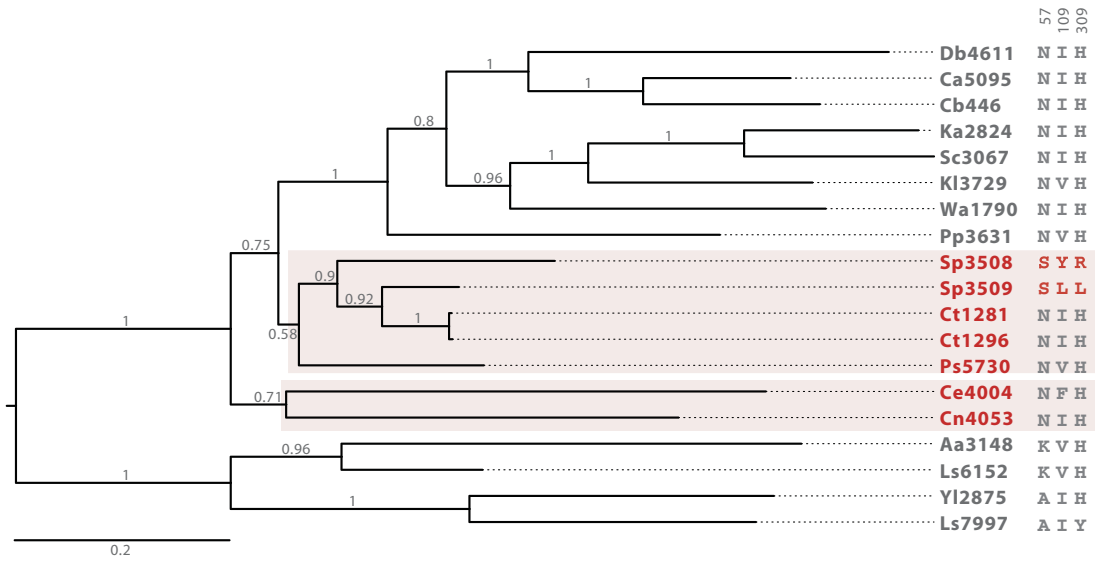

Fig. 5.—Xylose reductase Bayesian gene tree inference of fam634. Posterior probabilities are shown above the branches. The three columns show amino acid alignment for sites which have positive selection evidence. Red amino acids are distinct when compared with other species. Highlighted boxes represent clades with positive selection evidence on fermenter groups. Tree is scaled by substitutions per site.

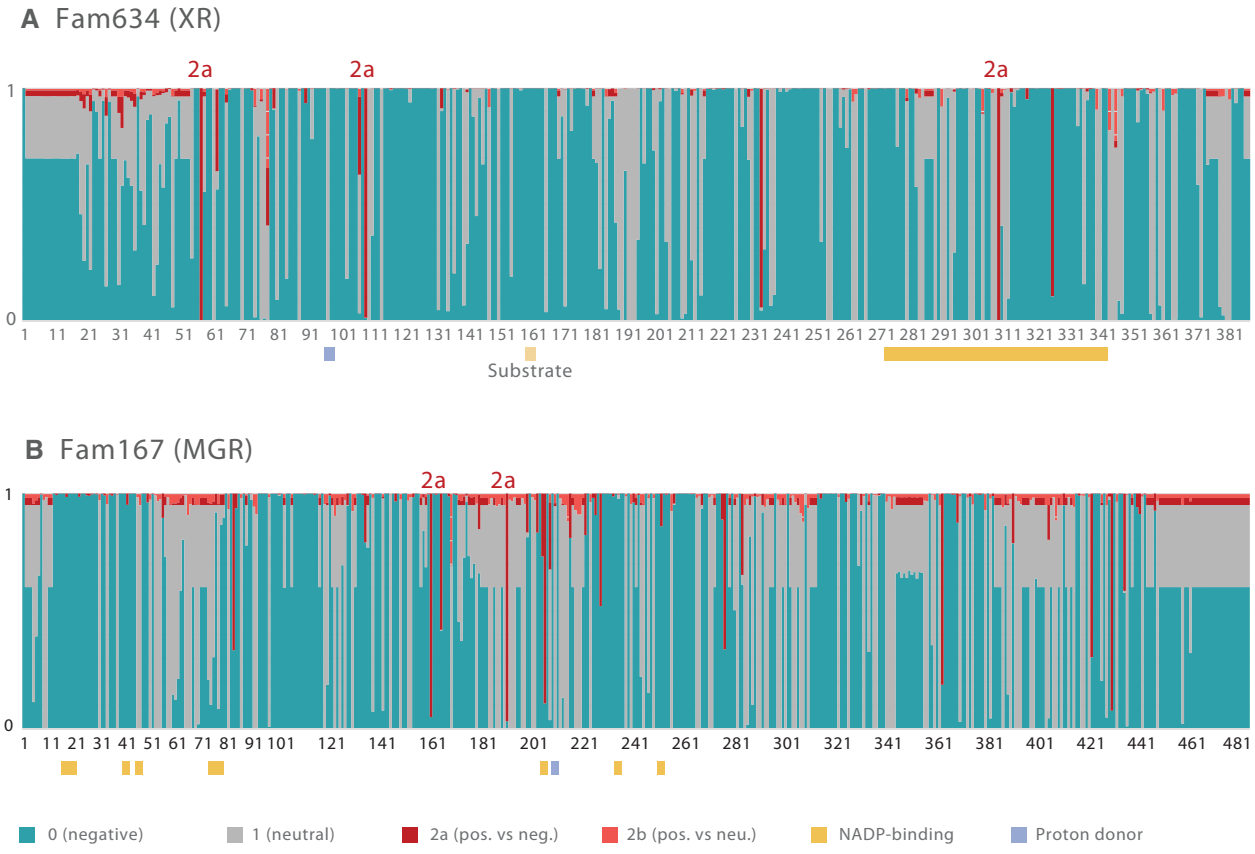

Fig. 6.—Inferred probabilities of evolutionary models along protein sequences. Plots of posterior probabilities (in *y* axis) for each tested evolutionary model (negative selection, neutral evolution, positive selection) in codeml model 2, for each protein site (*x* axis). Sites with higher probability than 0.95 for positive selection are indicated with the model code above graphs. Cofactor binding sites and proton donor sites known for *S. cerevisiae* are also shown.

MGR, ALD, ADH, TKL, and NQM; see table 2 for abbreviation) and 11 have undergone differential selective evolution on the fermenters' clade (XR, GPI, FBA, PGK, PGM, ENO, MGR, ALD, ADH, NQM, and PRS). The finding of positive selection fingerprints on glucose-related enzymes that are central pieces of the energetic cell metabolism was unexpected. Enzymes such as glucose-6P-isomerase (GPI), phosphoglycerate mutase (PGM) and triose-phosphate isomerase (TPI), are mainly coded by low copy gene families and are considered housekeeping. Although the removal of mutations by negative selection has been considered the rule on housekeeping genes evolution, evidence of positive selection has been described for *Drosophila* (Eanes 2011).

Among enzymes with fingerprints of positive selection, the ones comprehending the AKR superfamily are especially variable both in gene copy numbers and functional promiscuity (Kuhn et al. 1995; Jojima et al. 2015). XR and XDH may be attributed to different AKR gene families and, as seen in our analysis, did not group in monophyletic clusters. We suggest that these enzymes may have evolved more than once from different ancestral proteins. Dehydrogenase enzymes on the ethanol pathway (ALD and ADH) can also be assigned to different gene families with high paralog gene numbers. The assignment in different families of homologous genes, and the higher number of paralogs observed, connects with previous works' demonstrations that different AKR enzymes may present variable functions and substrates (Penning 2015).

Three AKR enzymes, the XR, ALD, and ADH, have been considered especially important for redox balance as a result of their cofactor preference that was extensively related to xylose consumption and fermentation efficiency (Ma et al. 2013). We have found sites evolving by positive selection in all three families, not only when generally analyzing all species, but also when focusing on the fermenters' clade. This may hint that these proteins have been gaining potentially adaptive variations along the evolution of yeasts. Specifically in the fermenters' clade, other variants were not only kept during evolution but also strongly selected by specific needs of its yeasts that might be intrinsically related to the evolution of xylose fermentation (Manel et al. 2016). These findings support the well-known idea that the simple presence, or absence of genes in the xylose metabolic pathway does not imply in xylose consumption or fermentation, also discussed by Riley et al. (2016). Besides the existence of genes in the genome, functional issues of the protein and related metabolic pathways must be affecting xylose consumption and fermentation.

Wohlbach et al. (2011) had also found an enzyme described as an AKR protein named CtAKR, which they suggest is an $NADP^+$-dependent glycerol dehydrogenase, because there is about 60% of similarity with one of *S. cerevisiae* enzymes of this group (Gcy1), and have residues known to function using $NADP^+$. Also, when testing this enzyme in a *Sc* mutant lacking three genes possibly related to AKR function, this protein restores closely glycerol and other metabolites to original levels of the parental strain. In our analysis this enzyme was found in family33, which we had annotated as an XDH family. Similar results to those observed in Wohlbach's experiments could be expected if this enzyme was an XDH $NADP^+$-dependent protein, because it is known that a higher cofactor balance between XR and XDH hastens and improves xylose consumption while decreasing glycerol levels in cell (Petschacher and Nidetzky 2008; Lopes et al. 2016). Also, it would be possible that, considering AKRs' promiscuity, this enzyme could act using both xylitol and glycerol as substrates, not being restricted to one of them. Furthermore, other researchers presented some improvement on ethanol production by engineering a $NADP^+$-dependent XDH, in an attempt to correct the cofactor's unbalance (Matsushika et al. 2008). Among enzymes with evidence of positive selection in the fermenters' clade, the XR gene family is of special interest, because one of the three codons identified in our analysis lies within the NAD^+/ Nicotinamide adenine dinucleotide (NADH) binding region of the protein. Observing the variation in positive selected sites, we found that it relies especially on *S. passalidarum* substitutions in one or both gene copies. The second XR copy of *S. passalidarum* (called XYL1.2) presents a cofactor preference change, from NAD^+ mainly, toward the use of NADH (Lopes et al. 2016). Although the variations in XR sequences were already known, we present here the first source of evidence for protein sites that could be related to this trait. The observed changes in sites we identified fitting positive selection models must be considered targets for future experimental tests.

Our findings on gene family expansion in the fermenters' clade also pointed to two gene families that gained more genes than expected by a neutral evolutionary model and could be related to the evolution of xylose fermentation. The first family is composed of genes from a flocculin family (fam16) duplicated with paralogs being retained in many yeasts' species, but with an especially high number of copies in *C. tenuis*, and a potential extinction of this family in the ancestor of *S. stipitis* and *S. passalidarum* clade. The retention of such a high number of paralogs may suggest an adaptive process underlying it (Conant et al. 2014). The other gene family, composed of genes from MGR (fam167), possibly evolved from duplications of fam7 genes that occurred in the same ancestor of *S. stipitis* and *S. passalidarum* clade and highly increased in number in these two fermenter species. The complimentary expansion of these two families on the fermenters' clade, one expanding in *C. tenuis* and the other in *S. stipitis* and *S. passalidarum* could be related to different fermentation strategies evolving in each strain.

The MGR expanded gene family also has a pattern of convergence in sequences between *S. stipitis* and *S. passalidarum* species, as result of parallel evolution from different copies of fam7, suggesting that natural selection is shaping these new

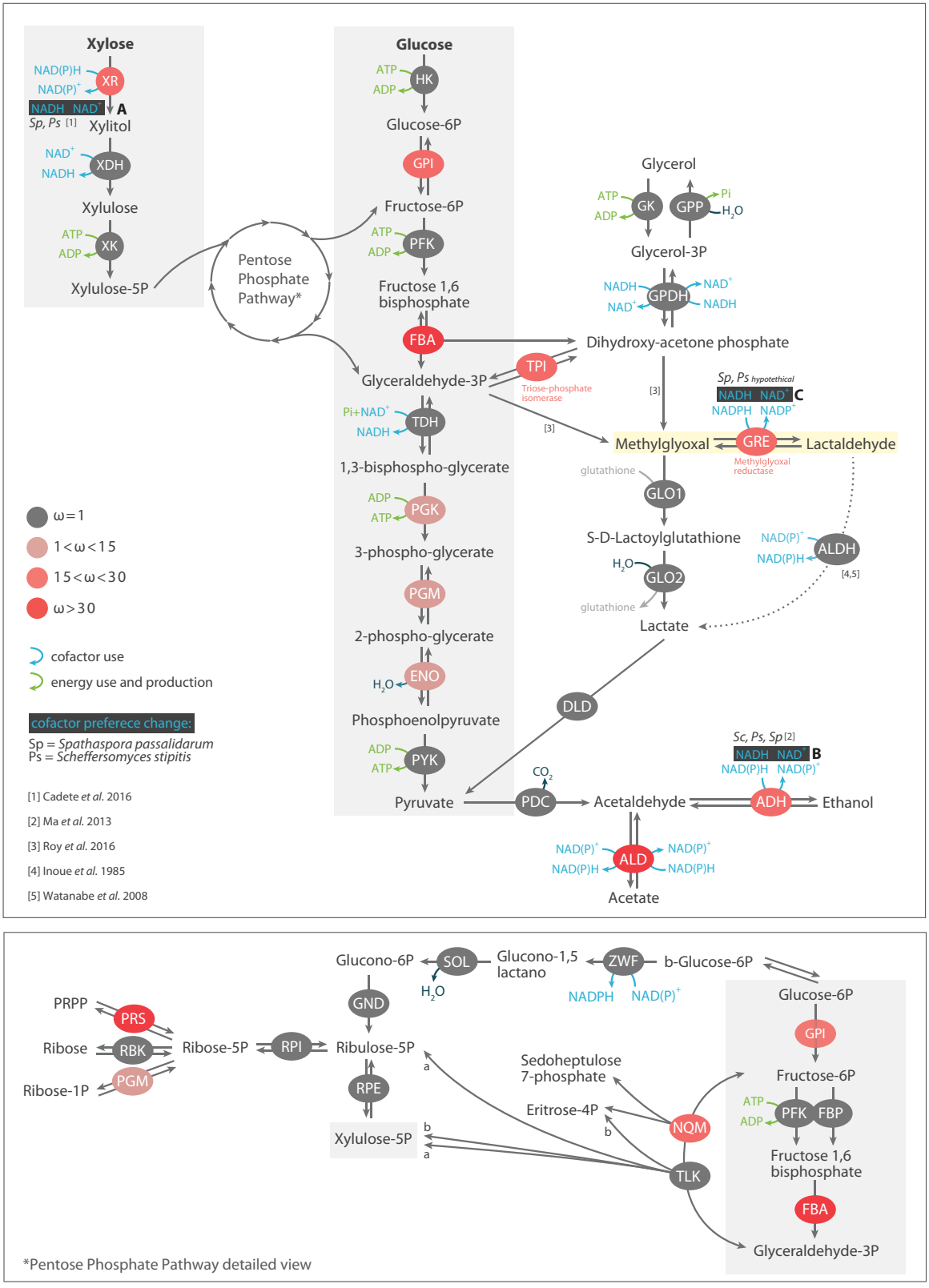

FIG. 7.—*Saccharomyces cerevisiae* metabolic pathways related to xylose and glucose consumption. Enzymes are colored according to omega (d*N*/d*S*, CODEML model 2, NsSites 1–2) values (light red—low values, dark red—high values). (A) and (B) refer to reactions where the cofactor preference is modified in *S. passalidarum* or *S. stipitis*. (C) refers to methylglyoxal reductase hypothetical cofactor preference change in paralog proteins with evidence for positive selection in our analysis.

duplicated genes. Additionally, strong evidence of positive selection was also found on MGR proteins, supporting that these proteins have been driven to accumulate adaptive changes in response to a strong pressure in its recent evolution. These two evidence of selection acting on MGR connects with a protein of high importance in organisms metabolism, with copies that may have slightly diverged in function, or even went under a sub functionalization process (Wisecaver et al. 2014).

Model for MGR Role in Oxidative Imbalance

Paralogs of MGR assigned to fam167 were not only expanded in copy number at the fermenters' clade (fig. 4), but also evolved under positive selection in at least 15 sites in xylose-fermenting lineages (fig. 6). Comparing the site positions within the protein of a homolog gene from fam7 in *S. cerevisiae* (*GRE2*), we found that they are in positions close to predicted cofactor binding sites. We, thus, hypothesize that some of the sites under selection are responsible for cofactor binding in *S. passalidarum* and in *S. stipitis*. Regarding the biochemical properties of changes in sites under selection, it is possible that like other known enzymes that rely mainly on NADPH and have cofactor preference change for NADH in *S. passalidarum* and *S. stipitis*, as pointed in figure 7A and B (Ma et al. 2013; Lopes et al. 2016), MGR displays the same altered preference in these xylose-fermenters, as pointed in figure 7C. This could be another clue to unravel the intricate regulation of NADH/NADPH and NAD^+/$NADPH^+$ regulation in the cell.

The reduction of methylglyoxal to L-lactaldehyde was supposed to end up on lactate in *S. cerevisiae* (as shown in dashed line, fig. 7, [4]). L-Lactaldehyde dehydrogenase (ALDH) was also supposed to be eliminated from *S. stipitis* (Watanabe et al. 2008). However, the so called ALDH enzymes have high similarity with ALD genes, which function similarly through the conversion of acetaldehyde to acetate. *S. cerevisiae* ALDH sequence and function were not found in most protein curated databases, such as Swiss-Prot and UniRef. The experimental conditions where ALDH proteins were first described leave some obscure points, allowing one to speculate that the observed results may be reached in the same way if the flux of reaction was occurring through the reverse reaction of MGR, producing methylglyoxal (fig. 7C). We propose that when using an MGR enzyme, the pathway reaches a reaction dead-end by accumulating L-lactaldehyde (solid lines in fig. 7C).

The glutathione-dependent path of methylglyoxal consumption is shown as the preferred reaction in *S. cerevisiae*, using GLO1 and GLO2 genes (INOUE et al. 1985). Besides this information, concerning the expansion of fam167 and its positive selection in xylose-fermenters, we expect that this enzyme has significant importance for these yeasts. If the cofactor preference is altered as we expect for MGR, in cases of oxidative imbalance as occurs in xylose consumption, these yeasts may consume methylglyoxal using the NADH-preferring enzyme, leading to an accumulation of L-lactaldehyde which is less toxic than methylglyoxal. At opportune moments, such as yeast growth in better cofactor balance conditions, L-lactaldehyde is converted to methylglyoxal in the opposite reaction, which is rapidly directed to the glutathione-dependent pathway, leading to pyruvate, that is used for energy production. This agrees with previous evidence that the GLO pathway is preferred under conditions of cell growth, whereas reductase pathway is preferred under stationary growth condition (INOUE et al. 1985).

Taking an expanded view of this process, NAD^+ regeneration using NADH-preferring enzymes that had natural cofactor preference for NADPH are already confirmed in at least two other enzymes, XR and ADH (fig. 7, black boxes, *A* and *B*). Regarding the influence of cofactor balance in cell for xylose fermentation, this third altered preference may introduce a new source of cofactor regeneration, leading to new insights of how xylose fermentation is much more abundant in the clade of *S. stipitis* and *S. passalidarum*. Even though other fermenting yeasts from outside this clade such as *C. tenuis* display relative high rates of fermentation, they probably do not present the same solution as our hypothesis, probably relying on other yet unknown modifications that converge in phenotype.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by CAPES—National Council for the Improvement of Higher Education, Center for Computational Engineering and Sciences—FAPESP/Cepid (2013/08293-7), grant (#2015/05578-6), (#2017/07008-8), and (#2017/08519-6) from São Paulo Research Foundation (FAPESP).

Literature Cited

Aditiya HB, Mahlia TMI, Chong WT, Nur H, Sebayang AH. 2016. Second generation bioethanol production: a critical review. Renew Sustain Energy Rev. 66:631–653.

Bajwa PK, Harrington S, Dashtban M, Lee H. 2016. Expression and characterization of glycosyl hydrolase family 115 α-glucuronidase from *Scheffersomyces stipitis*. Ind Biotechnol. 12(2):98–104.

Borelli G, José J, Teixeira P, dos Santos LV, Pereira GAG. 2016. De novo assembly of *Candida sojae* and *Candida boidinii* genomes, unexplored xylose-consuming yeasts with potential for renewable biochemical production. Genome Announc. 4:e01551–e01515.

Bouckaert R. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 10:1–6.

Cadete RM, et al. 2012. Diversity and physiological characterization of D-xylose-fermenting yeasts isolated from the Brazilian Amazonian forest. PLoS One 7(8):e43135.
Cadete RM et al. 2016. Exploring xylose metabolism in Spathaspora species: XYL1.2 from Spathaspora passalidarum as the key for efficient anaerobic xylose fermentation in metabolic engineered Saccharomyces cerevisiae. Biotechnol. Biofuels. 9:167. doi: 10.1186/s13068-016-0570-6.
Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. Curr Opin Plant Biol. 19:91–98.
Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 9(8):772.
De Schutter K et al. 2009. Genome sequence of the recombinant protein production host Pichia pastoris. Nat. Biotechnol. 27:561–566. doi: 10.1038/nbt.1544.
dos Santos LV, Carazzolle MF, et al. 2016. Unraveling the genetic basis of xylose consumption in engineered *Saccharomyces cerevisiae* strains. Sci Rep. 6:38676.
dos Santos LV, de Barros Grassi MC, et al. 2016. Second-generation ethanol: the need is becoming a reality. Ind Biotechnol. 12(1):40–57.
Dujon B et al. 2004. Genome evolution in yeasts. Nature. 430:35–44. doi: 10.1038/nature02579.
Eanes WF. 2011. Molecular population genetics and selection in the glycolytic pathway. J Exp Biol. 214(2):165–171.
Goffeau A et al. 1996. Life with 6000 Genes. Science (80-.) . 274:546–567. doi: 10.1126/science.274.5287.546.
Gordon JL et al. 2011. Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. Proc. Natl. Acad. Sci. 108:20024–20029. doi: 10.1073/pnas.1112808108.
Grigoriev IV, et al. 2012. The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res. 40(D1):D26–D32.
Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 30(8):1987–1997.
Inoue Y, et al. 1985. Metabolism of 2-oxoaldehydes in yeasts: purification and characterization of lactaldehyde dehydrogenase from *Saccharomyces cerevisiae*. Eur J Biochem. 153(2):243–247.
Jeffries TW et al. 2007. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis. Nat. Biotechnol. 25:319–326. doi: 10.1038/nbt1290.
Jojima T, et al. 2015. Promiscuous activity of (S, S)-butanediol dehydrogenase is responsible for glycerol production from 1, 3-dihydroxyacetone in *Corynebacterium glutamicum* under oxygen-deprived conditions. Appl Microbiol Biotechnol. 99(3):1427–1433.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.
Kolenová K, Ryabova O, Vršanská M, Biely P. 2010. Inverting character of family GH115 α-glucuronidases. FEBS Lett. 584(18):4063–4068.
Kück P, Longo GC. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool. 11(1):81.
Kuhn A, Van Zyl C, Van Tonder A, Prior BA. 1995. Purification and partial characterization of an aldo-keto reductase from *Saccharomyces cerevisiae*. Appl Environ Microbiol. 61(4):1580–1585.
Kunze G et al. 2014. The complete genome of Blastobotrys (Arxula) adeninivorans LS3 - a yeast of biotechnological interest. Biotechnol. Biofuels. 7:66. doi: 10.1186/1754-6834-7-66.
Li L, Stoeckert CJJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics 28(2):279–281.
Lopes MR, et al. 2016. Genomic analysis and D-xylose fermentation of three novel *Spathaspora* species: *Spathaspora girioi* sp. nov., *Spathaspora hagerdaliae* f. a., sp. nov. and *Spathaspora gorwiae* f. a., sp. nov. FEMS Yeast Res. 16(4):fow044.
Ma M, Wang X, Zhang X, Zhao X. 2013. Alcohol dehydrogenases from *Scheffersomyces stipitis* involved in the detoxification of aldehyde inhibitors derived from lignocellulosic biomass conversion. Appl Microbiol Biotechnol. 97(18):8411–8425.
Manel S, et al. 2016. Genomic resources and their influence on the detection of the signal of positive selection in genome scans. Mol Ecol. 25(1):170–184.
Matsushika A, et al. 2008. Expression of protein engineered NADP+-dependent xylitol dehydrogenase increases ethanol production from xylose in recombinant *Saccharomyces cerevisiae*. Appl Microbiol Biotechnol. 81(2):243–255.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148(3):929–936.
Oleksyk TK, Smith MW, O'Brien SJ. 2010. Genome-wide scans for footprints of natural selection. Philos Trans R Soc B. 365(1537):185–205.
Penning TM. 2015. The aldo-keto reductases (AKRs): overview. Chem Biol Interact. 234:236–246.
Petschacher B, Nidetzky B. 2008. Altering the coenzyme preference of xylose reductase to favor utilization of NADH enhances ethanol yield from xylose in a metabolically engineered strain of *Saccharomyces cerevisiae*. Microb Cell Fact. 7(1):9.
Piškur J et al. 2012. The genome of wine yeast Dekkera bruxellensis provides a tool to explore its food-related properties. Int. J. Food Microbiol. 157:202–209. doi: 10.1016/j.ijfoodmicro.2012.05.008.
Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol. 53(5):793–808.
Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7 Susko, E, editor. Syst. Biol. 67:901–904. doi: 10.1093/sysbio/syy032.
Ranwez V, Harispe S, Delsuc F, Douzery E. 2011. MACSE: multiple alignment of coding SEquences accounting for frameshifts and stop codons. PLoS One 6(9):e22594.
Riley R, et al. 2016. Comparative genomics of biotechnologically important yeasts. Proc Natl Acad Sci U S A. 113(35):9882–9887
Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61(3):539–542.
Shen X-X et al. 2016. Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. G3 (Bethesda). 6:3927–3939. doi: 10.1534/g3.116.034744.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.
Tutar Y. 2012. Pseudogenes. Comp Funct Genomics. 2012:1–4.
Urbina H, Blackwell M. 2012. Multilocus phylogenetic study of the *Scheffersomyces* yeast clade and characterization of the N-terminal region of xylose reductase gene. PLoS One 7(6):e39128.
Veras HCT, Parachin NS, Almeida J. 2017. Comparative assessment of fermentative capacity of different xylose-consuming yeasts. Microb Cell Fact. 16(1):153.
Watanabe S, Piyanart S, Makino K. 2008. Metabolic fate of L-lactaldehyde derived from an alternative L-rhamnose pathway. FEBS J. 275(20):5139–5149.

Wisecaver JH, Slot JC, Rokas A. 2014. The Evolution of Fungal Metabolic Pathways Stajich, JE, editor. PLoS Genet. 10:e1004816. doi: 10.1371/journal.pgen.1004816.
Wohlbach DJ, et al. 2011. Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. Proc Natl Acad Sci U S A. 108(32):13212–13217.
Wood V et al. 2002. The genome sequence of Schizosaccharomyces pombe. Nature. 415:871–880. doi: 10.1038/nature724.
Yang Z. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22(4):1107–1118.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15(12):496–503.
Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19(6):908–917.
Zhang J. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22(12):2472–2479.

Associate editor: George Zhang