

Comparative genomics of xylose-fermenting fungi for enhanced biofuel production

Dana J. Wohlbach^{1,2}, Alan Kuo³, Trey K. Sato², Katlyn M. Potts¹, Asaf Salamov³, Kurt M. LaButti³, Hui Sun³, Alicia Clum³, Jasmyan Pangilinan³, Erika Lindquist³, Susan Lucas³, Alla Lapidus³, Mingjie Jin^{4,5}, Christa Gunawan^{4,5}, Venkatesh Balan^{4,5}, Bruce E. Dale^{4,5}, Thomas W. Jeffries², Robert Zinkel², Kerrie W. Barry³, Igor V. Grigoriev³, Audrey P. Gasch^{1,2}

¹Department of Genetics, University of Wisconsin-Madison, 425G Henry Mall, Madison, WI 53706, USA.

²Great Lakes Bioenergy Research Center, 1550 Linden Drive, Madison, WI 53706, USA.

³United States Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA.

⁴Biomass Conversion Research Laboratory, Department of Chemical Engineering and Materials Science, Michigan State University, University Corporate Research Complex, 3900 Collins Road, Lansing, MI 48910, USA.

⁵Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI 48824, USA.

July 2011

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

Classification:

BIOLOGICAL SCIENCES – Genetics

Comparative genomics of xylose-fermenting fungi for enhanced biofuel production

Dana J. Wohlbach^{1,2}, Alan Kuo³, Trey K. Sato², Katlyn M. Potts¹, Asaf Salamov³, Kurt M. LaButti³, Hui Sun³, Alicia Clum³, Jasmyn Pangilinan³, Erika Lindquist³, Susan Lucas³, Alla Lapidus³, Mingjie Jin^{4,5}, Christa Gunawan^{4,5}, Venkatesh Balan^{4,5}, Bruce E. Dale^{4,5}, Thomas W. Jeffries², Robert Zinkel², Kerrie W. Barry³, Igor V. Grigoriev³, Audrey P. Gasch^{1,2}

¹Department of Genetics, University of Wisconsin-Madison, 425G Henry Mall, Madison, WI 53706, USA. ²Great Lakes Bioenergy Research Center, 1550 Linden Drive, Madison, WI 53706, USA. ³United States Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. ⁴Biomass Conversion Research Laboratory, Department of Chemical Engineering and Materials Science, Michigan State University, University Corporate Research Complex, 3900 Collins Road, Lansing, MI 48910, USA. ⁵Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI 48824, USA.

Corresponding Author:

Audrey P. Gasch
Laboratory of Genetics
University of Wisconsin-Madison
425G Henry Mall
Madison, Wisconsin 53706 USA
p: (608) 265-0859
f: (608) 262-2976
agasch@wisc.edu

Abstract

Cellulosic biomass is an abundant and underutilized substrate for biofuel production. However, many microbes cannot natively metabolize the pentose sugars abundant within hemicellulose, generating specific challenges for microbial biofuel production from cellulosic material. Although engineered strains of *Saccharomyces cerevisiae* can utilize the pentose xylose, the fermentative capacity pales in comparison to glucose, limiting the economic feasibility of industrial fermentations. To better understand xylose utilization for subsequent microbial engineering, we sequenced the genomes of two xylose-fermenting, beetle-associated fungi – *Spathaspora passalidarum* and *Candida tenuis*. To identify genes involved in xylose metabolism, we applied a comparative genomic approach across 14 Ascomycete genomes, mapping phenotypes and genotypes onto the fungal phylogeny, and measured genomic expression across five Hemiascomycete species with different xylose consumption phenotypes. This approach implicated many new genes and processes involved in xylose assimilation. Several of these genes significantly improved xylose utilization when engineered into *S. cerevisiae*, demonstrating the power of comparative methods in rapidly identifying genes for biomass conversion while reflecting on fungal ecology.

Introduction

Biofuel production from cellulosic material utilizes available substrates without competing with food supplies, and therefore presents an economic and environmental opportunity (1). In lignocellulosic plant stocks, which include agricultural residues and wood waste, the second-most abundant sugar after glucose is the pentose xylose. *Saccharomyces cerevisiae* (*Scer*) does not natively consume xylose but can be engineered for consumption with a minimal set of assimilation enzymes, including xylose reductase (Xyl1) and xylitol dehydrogenase (Xyl2) from the xylose-fermenting *Pichia stipitis* (*Psti*; Fig. 1A; 2, 3). However, xylose fermentation remains slow and inefficient in *Scer*, especially under anaerobic conditions when NADH cannot be recycled for NAD⁺-dependent Xyl2 (2, 4, 5). Therefore, improving xylose utilization in industrially relevant yeasts is essential for producing economically viable biofuels from cellulosic material.

A handful of Hemiascomycete yeasts naturally ferment pentose sugars (2, 3, 6). Best known is the xylose fermenting yeast *Psti*, associated with wood-boring beetles that may rely on fungi to release nutrients from wood (7, 8). Other related yeasts cannot ferment pentoses, suggesting that xylose fermentation has evolved in this unique fungal environment (9). Although some details are known (Fig. 1A; 2, 3, 10-12), much of the mechanism of xylose fermentation remains unresolved.

To elucidate genetic features that underlie xylose utilization, we sequenced the genomes of two additional xylose fermenting species and applied a cross-species comparative genomic approach. Comparing phenotypic and genomic differences in diverse Hemiascomycete species indicated many genes important for xylose utilization while reflecting on the unique niche experienced by these beetle-associated species. Here we present the comparative analysis of the genomes and transcriptomes of these yeasts, highlighting aspects of pentose assimilation as well as the ecological significance of these interesting fungi. In the process, we identified several genes that, when expressed in *Scer*, significantly improve xylose-dependent growth and xylose assimilation. By harnessing the power of nature and comparative genomics, this work provides a key improvement to xylose utilization, a significant roadblock to cellulosic biofuel production.

Results and Discussion

We sequenced the genomes of two xylose-fermenting yeasts, *Spathaspora passalidarum* (*Spas*, NRRL Y-27907) and *Candida tenuis* (*Cten*, NRRL Y-1498), for comparison to the existing *Psti* genome (*Materials and Methods* and *SI Appendix, Materials* and Table S1; 13). The *Spas* genome was sequenced to 43.77X coverage over 13.1 Mb arranged in eight scaffolds. The *Cten* genome was sequenced to 26.9X coverage, generating 10.7 Mb in 61 scaffolds representing eight chromosomes. Compared to other sequenced Hemiascomycetes, genome size and genic composition in the xylose-fermenting yeasts span the range from compact (5,533 genes in the 10.7 Mb *Cten* genome) to among the largest (5,841 genes in the 15.4 Mb genome of *Psti* and 5,983 genes in the 13.2 Mb genome of *Spas*; Table 1 and *SI Appendix, Table S2*). Sixty-seven percent of *Spas* and 74% of *Cten* genes are orthologs located in syntenic regions (*SI Appendix, Fig. S1*), and about half of all genes in *Spas*, *Cten*, and *Psti* show three-way synteny.

Xylose Consumers are Members of the ‘CUG Clade’ of Commensal Fungi. We selected 11 other Ascomycetes with available genome sequences (Table 1) for comparison to *Spas*, *Cten*, and *Psti* (Fig. 1*B,C*). Whole-genome phylogenetic analysis placed both *Spas* and *Cten* within the ‘CUG clade’ of yeasts (Fig. 1*B* and *SI Appendix, Materials*), named for the alternative decoding of the CUG codon as serine instead of leucine (14-16). We compared tRNA sequences across the 14 species in our analysis and confirmed that *Spas* and *Cten* harbor the serine tRNA evolved to recognize the CUG codon (14), whereas there were no identifiable sequences similar to standard *Scer* serine tRNAs (*SI Appendix, Fig. S2A,B*). Likewise, a genome-wide scan revealed that the majority of CUG codons from *Candida* and related species (including *Spas* and *Cten*) are decoded as serine in *Scer* orthologs; CUG codons from species outside the CUG clade are decoded as leucine in orthologous *Scer* genes (*SI Appendix, Fig. S2C*). Together, these results support the phylogenetic placement of xylose-fermenting species within the CUG clade. Interestingly, most other species in this CUG group are commensal with humans but can emerge as opportunistic pathogens (17, 18). Thus, commensalism, albeit in association with different hosts, appears to be a feature common to this clade.

Clade-Specific Patterns of Gene Presence. To identify genes associated with xylose utilization, we compared gene content between the 14 Ascomycetes in our phylogeny by assigning orthology and paralogy relationships among the meta-set of 81,907 predicted fungal protein-coding genes (*SI Appendix, Materials*). Over 12,000 orthologous gene groups (OGGs) were resolved, with 5,749 OGGs (91% of all genes) found in at least two species (*SI Appendix, Table S3 and Fig. S3A*). In contrast, the other OGGs (52% of all OGGs representing 9% of all genes) are species-specific paralogs that are distributed non-randomly throughout the phylogeny (*SI Appendix, Fig S3B*). Within the CUG clade, *Debaryomyces hansenii* (*Dhan*) and *P. guilliermondii* (*Pgu*) have the most single-species expansions while the xylose-fermenting fungi (*Spas*, *Cten* and *Psti*) have some of the fewest. Interestingly, amplifications in the xylose-fermenters include sugar transporters and cell-surface proteins, which could be related to their unique sugar environment (*SI Appendix, Table S4 and S5*).

We analyzed conservation patterns of the 5,749 multi-species OGGs through a clustering approach, which identified clade-specific OGGs enriched for different functional properties (Fig. 2A and *SI Appendix, Table S6*). Approximately half of the multi-species OGGs are common to all 14 Ascomycetes. These ubiquitous OGGs are significantly enriched for essential metabolic processes including nucleic acid ($p = 1.32e-42$, hypergeometric distribution), small molecule ($p = 6.28e-35$), and protein ($p = 2.51e-14$) metabolism, as well as transcription ($p = 2.76e-23$) and response to stress ($p = 1.30e-31$).

The remaining OGGs can be clustered into five major clade-specific groups. Remarkably, the majority of clade-specific OGGs (including those unique to well-studied fungi such as *Scer*) are significantly enriched for unclassified and uncharacterized proteins ($p = 4.271e-21$). This finding reveals a general bias in our understanding of gene function and highlights the dearth of information on species-specific processes, even for the best characterized organisms like *Scer*:

OGGs unique to the CUG clade are enriched for genes encoding lipases and cell-surface proteins ($p = 1.306e-6$ and $6.665e-6$, respectively), as previously noted in *Candida* species (19). Although enrichment of these genes in *Candida* species was previously interpreted to be important for pathogenicity (19), their presence in beetle symbionts suggests they may be relevant to commensalism, rather than pathogenicity *per se*.

Additionally, many genes unique to CUG yeasts are involved in *de novo* NAD⁺ biosynthetic processes ($p = 0.00891$), suggesting novel metabolism that may reflect a more complex environment of these commensal organisms.

Surprisingly, orthologs of known xylose-utilization genes are present in all 14 Ascomycetes, even though most Hemiascomycetes cannot utilize xylose (6). This group includes orthologs of *Psti* xylose reductase (*XYL1*; 11), xylitol dehydrogenase (*XYL2*; 12), and xylulokinase (*XYL3*; 10), the minimal set of genes required to engineer *Scer* for xylose assimilation (Fig. 1A; 2, 3, 5). However, these genes show no evolutionary signatures of selection or constraint to suggest functional modification in the xylose-utilizing species (*SI Appendix*, Fig. S4). Thus, other factors must contribute to phenotypic differences in xylose consumption besides the mere presence of this ‘minimal’ gene set.

Conservation of Orthologous Gene Groups Points to Novel Xylose Utilization Genes. To identify genes relevant to xylose fermentation, we devised a phylogenetic approach to correlate genotype to phenotype across the Ascomycetes. First, we examined xylose growth and fermentation (Fig. 2B and *SI Appendix*, Figs. S5 and S6). *Psti*, *Spas*, and *Cten* were the only species able to measurably ferment xylose in our assay (*SI Appendix*, Fig. S6). Intriguingly, these are also the yeasts associated with beetles, many of which are attracted to fermentation byproducts (20). Only three genes are uniquely found in these xylose-fermenting species, one of which contains an α -glucuronidase domain and a signal peptide sequence indicative of secretion (*SI Appendix*, Fig. S7). While its connection to xylose utilization is not clear, this protein may be secreted for degradation of complex carbohydrates in woody biomass.

We expanded our analysis to consider xylose assimilation. Notably, *Lodderomyces elongisporus* (*Lelo*) is the lone member of the CUG clade unable to grow on xylose (Fig. 2B), suggesting that the phenotype was present in the group’s common ancestor but lost in this lineage. Because genes involved in sugar metabolism are not maintained in the absence of selection (21, 22), we reasoned that species unable to grow on xylose may have lost key assimilation genes. We therefore looked for genes whose presence and absence across the fungi correlated with the ability to grow on xylose.

Forty-three genes were absent in xylose non-growers but common to all xylose fermenters, with varying conservation across species that could assimilate xylose (Fig. 2C). Fifteen showed presence and absence patterns that strictly correlated with xylose assimilation. These include orthologs of a putative *Psti* xylose transporter and several endoglucanases that break down higher-order sugars in hemicellulose. Most other genes are unannotated and fungal specific; ten are also found in other fungi capable of plant cell wall degradation. However, two of the proteins have signal peptide sequences: an oxidoreductase and a putative glycoside hydrolase, both of which could be potentially useful for biomass degradation (see *SI Appendix*, Fig. S7 for protein domain and signal peptide analysis). Although the conservation of these genes is suggestive of functional importance, we did not detect any signatures of constraint within the xylose fermenters.

Cross-Species Genomic Expression Identifies Additional Xylose-Responsive Genes. As a second approach to identify xylose metabolism genes, we characterized genomic expression during glucose versus xylose growth in five species including the three xylose-fermenters, xylose-growing *Candida albicans* (*Calb*), and *Lelo*, which is unable to grow on xylose (*Materials and Methods*). We performed a comparative analysis of orthologous gene expression via hierarchical clustering (Fig. 3 and *SI Appendix*, Fig. S8) and significance testing (*SI Appendix*, Tables S7 and S8). The xylose response was strikingly dissimilar across species (Fig. 3A). In particular, *Lelo* altered the expression of thousands of genes, including orthologs of the yeast environmental stress response (ESR) that are induced when *Scer* is stressed (23) or experiences xylose (*SI Appendix*, Fig. S8A and Table S9; 24). This massive expression pattern in *Lelo* likely represents a starvation response to carbon limitation, and demonstrates that the ESR is conserved in this species. In addition, *Lelo*, along with *Cten* and *Calb*, induced ~90 OGGs enriched for fatty acid and lipid catabolism, suggesting reliance on fatty acids as a carbon source (*SI Appendix*, Fig. S8B and Table S10). We also identified two clusters of genes induced by xylose in most or all species, regardless of their xylose growth phenotypes (Fig. 3B,C). These include genes whose expression is required for optimal xylose utilization in engineered *Scer* (e.g. *XYL1*, *XYL2*, *XYL3*, *TKL1*, and *TAL1*; see Fig. 1A). Strikingly, several of these genes were strongly

induced in *Lelo*, even though it cannot utilize xylose. Thus, remnants of the xylose signaling cascade persist in *Lelo*, despite recent loss of xylose assimilation.

In addition to known xylose metabolism genes, others relating to carbohydrate transport and metabolism were highly induced specifically in xylose growers. Genes encoding beta-glucosidases and cellulases were strongly induced, suggesting that xylose participates in a positive feedback loop to catalyze its own release from hemicellulose. Orthologs of genes metabolizing other carbohydrates (including galactose, maltose, and glucose) were also up-regulated. Thus, in their native environment these species may not encounter free xylose in the absence of complex sugars, and are unlikely to rely on it as a sole carbon source. Additionally, the xylose-fermenting species induced several genes linked to redox regeneration, a well-known bottleneck in *Scer* engineered for xylose fermentation (2, 3). Genes encoding NADPH-generating steps of the pentose phosphate pathway (*ZWF1* and *PGIT*) were up-regulated, perhaps to feed NADPH-consuming xylose reductase. Other genes implicated in NAD(P)⁺/H recycling or oxido-reduction were also induced and may function to maintain redox balance during xylose assimilation.

Candidate Genes Improve Xylose Utilization. We tested ten of the genes implicated above for their ability to enhance xylose utilization in two different engineered *Scer* strains (*Materials and Methods* and *SI Appendix*, Note S1). Genetic background influenced the effect of overexpression, and several genes improved growth on both xylose and glucose (*SI Appendix*, Fig. S9), including a putative hexose transporter (*SpHXT*) and a glucose-6-phosphate dehydrogenase (*SpGPD*). Importantly, two genes had a specific positive effect on xylose utilization in one or both strain backgrounds: a *Cten* aldo/keto reductase, *CtAKR*, and a *Spas* unannotated protein, *SpNA*, with homology to uncharacterized fungal-specific proteins (Fig. 4 and *SI Appendix*, Fig. S10).

Expression of plasmid-born *CtAKR* significantly improved xylose consumption during both aerobic and anaerobic growth (Fig. 4*B*). Notably, xylose consumption increased by 32% after 72 h of anaerobic fermentation ($p = 0.0369$, t-test). At the same time, xylitol production relative to xylose consumption was 73% lower (Fig. 4*C*) indicating improved flux through the xylose-assimilation pathway. Glycerol production, which represents a significant drain on ethanol production under anaerobic conditions (25, 26),

was not significantly increased (*SI Appendix*, Fig. S11). However, acetate production was reduced 42% (Fig. 4C). Because acetate is a weak acid stress for yeast, lower acetate levels could facilitate increased cell growth. Indeed, some of the increased xylose utilization went into biomass production (*SI Appendix*, Fig. S11); however, the improved xylose utilization did not increase ethanol titers, revealing that ethanol production was not limited by carbon availability, but by other factors. Nonetheless, the significant effect of p*CtAKR* on anaerobic xylose assimilation and concomitant reduction in xylitol represents a major advance in cellulosic biomass conversion by *Scer*.

CtAKR is a member of the large protein family that includes xylose reductases (*SI Appendix*, Fig. S12A). However, *CtAKR* is most similar to the NADP⁺-dependent glycerol dehydrogenase Gcy1 from *Scer*, which functions in an alternative pathway for glycerol catabolism (Fig. 1A; 27). Notably, *CtAKR* contains residues known to establish NADP⁺ binding (Fig. S12B; reviewed in 28), suggesting *CtAKR* may also function in a NADP⁺-specific manner. We examined the effect of p*CtAKR* expression on glycerol metabolism in a *Scer* mutant lacking three functionally redundant AKRs (*GCY1*, *YPR1*, *GRE3*; *Materials and Methods*). Glycerol levels increased in the mutant strain but were restored to wild-type levels by p*CtAKR* (Fig. 4D). Together, these data suggest that *CtAKR* functions as a NADP⁺-dependent glycerol dehydrogenase in *Scer*. Indeed, like *CtAKR*, overexpression of *Scer* *GCY1* or *YPR1* had a positive effect on xylose utilization (*SI Appendix*, Fig. S13), further supporting our hypothesis for *CtAKR* function.

Summary and Discussion

Previous work aimed at improving *Scer* xylose fermentation focused on metabolic modeling (29), single-species genome and expression analysis (29, 30), or directed evolution (31). In this study, we utilized a comparative genomics approach to understand xylose utilization in several different beetle-associated fungi. Our approach reveals that these species share some features with other commensal fungi, yet display specific traits (*e.g.*, the ability to ferment xylose and expression of genes involved in cellulose degradation) that may be specific to their relationship with wood-boring insects. The ability to assimilate xylose is associated with altered expression of several genes central to glycolysis, xylose catabolism, and the pentose phosphate shuttle, revealing that decades of

directed evolution have largely recapitulated the natural expression response in these species. That some aspects of this response were observed in species that cannot assimilate xylose (namely *Lelo*) indicates that remnants of the genomic expression program can remain long after the ability to consume the sugar has been lost.

Additionally, several induced genes are related to reducing potential. Indeed, one of the biggest challenges for xylose fermentation in *Scer* engineered with *Psti XYL1,2,3* is the cofactor imbalance that emerges under anaerobic conditions. During anaerobic growth, NADH cannot be recycled through respiration, leading to a shortage of NAD⁺ to supply Xyl2 and thus an accumulation of xylitol (2). To reduce this redox imbalance, *Scer* increases NADH-dependent glycerol production. We found that overexpression of a *Cten* glycerol dehydrogenase significantly increased flux through the xylose assimilation pathway, without the typical xylitol accumulation. We hypothesize that *CtAKR* increases cycling through the glycerol metabolic pathway, producing NADPH through alternative glycerol catabolism, which in turn promotes glycerol production and NADH recycling. That glycerol levels do not significantly change in strains engineered with p*CtAKR* is consistent with this cycling hypothesis. The combined effects may promote the first two steps of xylose assimilation, which require NADPH and NAD⁺, by helping to alleviate cofactor imbalance. Decreased acetate levels may also result from increased glycerol cycling, since acetate is otherwise generated as a fermentation byproduct to alleviate cofactor imbalance (4). While the precise mechanism will be the subject of future study, our ability to identify genes that improve xylose assimilation shows the promise of harnessing ecology and evolution through comparative genomics for biofuel research.

Materials and Methods

Genome and EST Sequencing, Assembly and Annotation. We sequenced *Spas* and *Cten* using Sanger (40kb fosmid library) and 454 (standard and paired ended libraries) sequencing platforms. Newbler (Roche, v2.3) was used to produce hybrid 454/Sanger assemblies. Gaps were closed by gapResolution (<http://www.jgi.doe.gov/>), PCR and fosmid clone primer walks, or editing in Consed (32). Illumina reads improved the final consensus quality with Polisher (33). *SI Appendix*, Table S1 lists genome sequencing statistics.

mRNA was purified using Absolutely mRNA™ purification kit (Stratagene) and reverse transcribed with SuperScriptIII using dT₁₅VN₂ primer. cDNA was synthesized with *E. coli* DNA Ligase, polymerase I, and RNaseH (Invitrogen), nebulized, and gel purified for fragment sizes between 500-800 bp. Fragments were end repaired, adaptor ligated, and made into single stranded DNA libraries using the GS FLX Titanium library kit. Single-stranded DNA libraries were amplified in bulk and sequenced using a 454 Genome Sequencer FLX. Reads from each EST library were filtered, screened, and assembled using Newbler. Both genomes were annotated using the JGI annotation pipeline (*SI Appendix, Materials*), and can be accessed through the JGI Genome Portal (<http://www.jgi.doe.gov/spathaspora/> and <http://www.jgi.doe.gov/tenuis/>).

Species Phylogeny and Orthology. We estimated the phylogeny using protein sequences of 136 single-copy orthologs present in all species (*SI Appendix, Materials*). Phylogenies were constructed with MrBayes v3.1.2 (34, 35). We created OGGs using a modified RSD (36) and OrthoMCL (37) method. RSD parameters: significance threshold, 10⁻⁵; alignment threshold, 0.3. OrthoMCL parameters: significance threshold, 10⁻⁵; inflation parameter, 1.5. Pairwise one-to-one orthologs were assigned with RSD between each species and one of four reference species: *Scer*, *Psti*, *Calb*, and *Spom*. Results from the two methods were compared and combined using a custom perl script to maximize high confidence assignments (true positives) and minimize low confidence assignments (false positives; see *SI Appendix, Materials*).

Fungal Strains. All fungal species used in this study are sequenced strains and are listed in Table 1. Heterologous overexpression of selected *Spas* or *Cten* genes was conducted in two different *Scer* strain backgrounds: BY4741 or a wild diploid strain (GLBRCY0A). A codon-optimized DNA cassette (DNA2.0, Inc.) containing the *Scer* *PGK1* promoter (Pr), *TDH3* terminator (t), Pr*TDH3*, t*TEF2*, Pr*TEF2*, and t*CYC1*, followed by the KanMX selection marker (38) was synthesized with or without codon-optimized *Psti* *XYL1*, *XYL2*, and *XYL3* genes between each promoter-terminator pair (in order). The cassettes were integrated at the *HO* locus in single copy. Ten individual *Spas* or *Cten* genes lacking CUG codons that were induced in a majority of the xylose fermenters were cloned between *Scer* Pr *TEF1* and t*TUB1* in a 2-micron pRS426 vector (39) modified with a Hyg selection marker (see *SI*

Appendix, Note S1 for more details on engineered strains). Gene deletions were created by homologous recombination to replace the coding sequence with KanMX or HygMX drug resistance cassettes. All constructs were confirmed by diagnostic PCR and/or DNA sequencing.

Phenotypic Assays. For all assays, cultures were grown in YPD (1% yeast extract, 2% peptone, 2% glucose) or synthetic complete (SC) medium (1.7 g/L yeast nitrogen base, essential amino acids and 1 g/L ammonium sulfate or monosodium glutamate when mixed with Geneticin), with 2% glucose (SCD) at 30°C for at least 16 h to early-mid log phase. Cells were washed once in SC (no sugar), diluted, and transferred either to liquid or solid media containing 2% - 10% glucose or xylose. For solid growth assays, plates were scored after two days at 30°C. For liquid growth assays, OD₆₀₀ was monitored with Spectronic 20D+ (Thermo Scientific), or TECAN F500 or M1000 plate readers. For fermentation, 50 mL of washed cells were transferred to an airlocked 125-mL Erlenmeyer flask and were incubated at 30°C in an orbital shaker at 100 rpm. Supernatant was filtered through a 0.22 µm filter prior to analysis by HPLC with a Biorad Aminex HPX-87H column (40). Concentrations of ethanol were also determined using an Agilent Technologies 7890A gas chromatograph with a 7693 auto sampler and flame ionization detector (*SI Appendix, Materials*).

Microarrays. Cells were collected at OD₆₀₀ 0.5-0.6 after 3 generations growth in 2% glucose or xylose. Cell lysis and total RNA isolation were performed as previously described (41). RNA was further purified with LiCl and Qiagen RNeasy kit. Sample labeling was performed as previously described (41) using cyanine dyes (Amersham), Superscript III (Invitrogen), and amino-allyl-dUTP (Ambion). Whole-genome, species-specific 375K microarrays (Roche-NimbleGen) were designed with chipD (*SI Appendix*, Table S11; 42). Arrays from three biological replicates were hybridized in a NimbleGen hybridization system 12 (BioMicro), washed, and scanned using a scanning laser (GenePix 4000B, Molecular Devices) according to NimbleGen protocols (<http://www.nimblegen.com/>). Data normalization and statistical analyses were performed using Bioconductor (43) and custom perl scripts. The *affy()* package (44) was used to apply probe-level quantile normalization to the log₂ signal of RNA versus species-specific genomic DNA control. Genes with significant expression differences in response to xylose were identified separately for

each species by performing paired t-tests using the Bioconductor package Limma v2.9.8 (45) with a false discovery rate (FDR) correction of 0.05 (46). For cross-species comparisons, genes within OGGs were evaluated for expression differences. When an OGG contained more than one gene from a particular species, genes with the smallest phylogenetic distance (determined with PAML v4.3; (47) were directly compared. Hierarchical clustering of gene expression across species was performed with Cluster 3.0 using the uncentered Pearson correlation as the distance metric (48).

Acknowledgements: We thank Meredith Blackwell, Aviv Regev, Dawn-Anne Thompson, and Cletus Kurtzman for strains; Yann Dufour for assistance in microarray design; Cécile Ané for bioinformatic and phylogenetics support; Alan Higbee and Gwen Bone for GC and HPLC analysis; Thomas Kuster for images of yeasts; and Rebecca Breuer and Ben Bice for technical assistance. This work was performed under the auspices of the US Department of Energy, and was funded in part by the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494), an NLM training grant (5T15LM007359) to D.J.W., and an NSF CAREER Award (0447887) to A.P.G.

Author Contributions: D.J.W. designed and performed experiments, analyzed data, and wrote the manuscript. T.K.S and B.D.B. engineered strains and performed phenotyping. K.M.P. and R.Z. assisted in phenotyping. J.P., E.L., and S.L. performed EST and genome sequencing. K.M.L., H.S., A.C., and A.L. performed genome assembly and finishing. A.K., A.S., and I.V.G. performed genome annotation. K.B. and I.V.G. coordinated genome sequencing. M.J., C.G., V.B., and B.E.D. designed, conducted, and analyzed fermentation experiments. A.P.G. coordinated the project, designed experiments and edited the manuscript. All authors discussed results and commented on the manuscript.

Data deposition: The assemblies and annotations reported here have been deposited to Genbank under accession numbers AEIK000000000 (*Spas*) and AEIM000000000 (*Cten*). Microarray data have been deposited to the Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE24858.

This article contains supporting information.

Figure Legends

Fig. 1. Overview of xylose assimilation and phylogeny of xylose-fermenting fungi. (A) The simplified pathway includes genes that have been engineered in *Scer* via overexpression (red text) for improved xylose fermentation. *XYL1*, xylose reductase; *XYL2*, xylitol dehydrogenase; *xylA*, xylose isomerase; *XKS1/XYL3*, xylulokinase; *ZWF1*, glucose-6-phosphate dehydrogenase; *GND1*, 6-phosphogluconate dehydrogenase; *RK11*, ribose-5-phosphate ketol-isomerase; *RPE1*, ribulose-5-phosphate 3-epimerase; *TKL1*, transketolase; *TAL1*, transaldolase. (B) Maximum likelihood phylogeny from concatenated alignment of 136 universal orthologs, with bootstrap values. (C) Electron microscopy images of *Cten* (top panel), *Psti* (middle panel), and *Spas* (bottom panel). Scale bar, 2 μ m.

Fig. 2. Mapping of phenotype and genotype onto phylogeny. (A) Hierarchical clustering based on ortholog presence (orange) or absence (grey) for 3,073 non-ubiquitous multi-species OGGs. Blue indicates BLAST homology despite no ortholog call. Functional enrichment in indicated clusters is described in *SI Appendix*, Table S6. (B) Average $\pm$ SD ($n = 3$) xylose (blue) and glucose (red) growth curves for fungi growing on 2% (closed circles), 8% (open squares), or 0% (black) sugar. (C) OGG patterns for 43 genes present (orange) in xylose-fermenting species and absent (grey) in non-xylose-assimilating species, as described in text. Species abbreviations as in Table 1. Green text, xylose-growing species; purple box, xylose-fermenting species.

Fig. 3. Transcriptome analysis of xylose growing cultures. (A) Overlap between significantly differentially expressed genes within the xylose-fermenters ($FDR < 0.05$). (B-C) Two clusters of genes, identified by hierarchical clustering, that are highly induced in most species responding to xylose. Data represents average expression change ($n = 3$) for indicated genes (rows) in each species (columns). Red indicates higher, green represents lower, and black shows no change in expression in response to xylose. Grey indicates no ortholog detected. Purple blocks represent statistically significant fold-changes ($FDR < 0.05$) in *Psti*, *Spas*, *Cten*, *Calb*, and *Lelo*. Blue text, genes related to carbohydrate metabolism; purple text, genes related to redox balance; underlined text, known engineering targets for improved *Scer* xylose utilization.

Fig. 4. *CtAKR* improves *Scer* xylose utilization. (A) Average $\pm$ SD ($n = 4$) growth on 8% xylose of *Scer* strain GLBRCY0A carrying *PsXYL123*+p*CtAKR* (blue), *PsXYL123*+VOC (vector only control; green), p*CtAKR* only (grey), or VOC only (black). (B) Average $\pm$ SD ($n = 3$) xylose consumed after 72 hours growth for GLBRCY0A carrying *PsXYL123*+p*CtAKR* (purple) or *PsXYL123*+VOC (grey). Asterisks indicate statistically significant measurements ($p < 0.05$, t-test). (C) Average $\pm$ SD ($n = 3$) xylitol or acetate produced after 72 h anaerobic fermentation for GLBRCY0A carrying *PsXYL123*+p*CtAKR* (blue) or

PsXYL123+VOC (grey). Inset: time course of average $\pm$ SD (n = 3) anaerobic xylitol production relative to xylose consumed. (**D**) Average $\pm$ SD (n = 3) glycerol produced in wild-type (WT, BY4741) or mutant strains carrying p*CtAKR* (aqua) or VOC (grey).

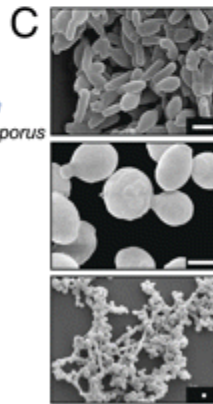
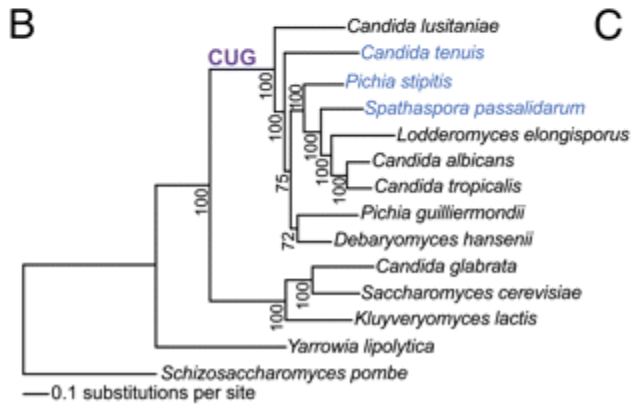
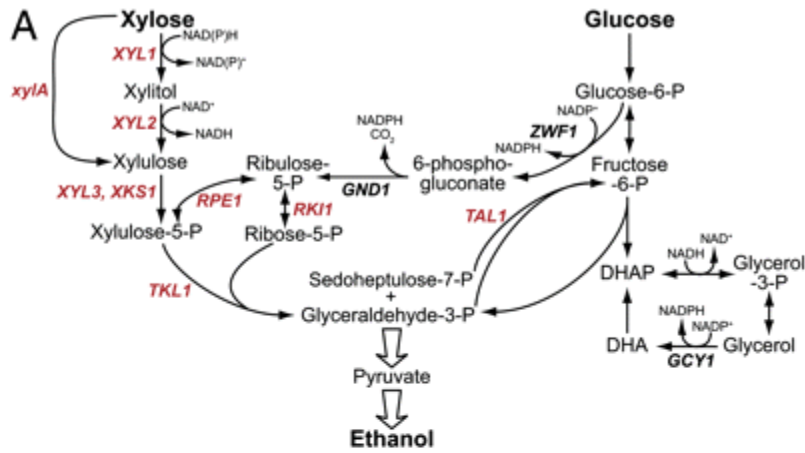
References

1. Solomon BD (2010) Biofuels and sustainability. *Ann N Y Acad Sci* 1185:119-134.
2. Jeffries TW (2006) Engineering yeasts for xylose metabolism. *Curr Opin Biotechnol* 17:320-326.
3. Van Vleet JH & Jeffries TW (2009) Yeast metabolic engineering for hemicellulosic ethanol production. *Curr Opin Biotechnol* 20:300-306.
4. Jeppsson M, Johansson B, Hahn-Hägerdal B, & Gorwa-Grauslund MF (2002) Reduced oxidative pentose phosphate pathway flux in recombinant xylose-utilizing *Saccharomyces cerevisiae* strains improves the ethanol yield from xylose. *Appl Environ Microbiol* 68:1604-1609.
5. Kötter P & Ciriacy M (1993) Xylose fermentation by *Saccharomyces cerevisiae*. *Appl Microbiol Biot* 38:776-783.
6. Jeffries TW & Kurtzman CP (1994) Strain selection, taxonomy, and genetics of xylose-fermenting yeasts. *Enzyme Microb Tech* 16:922-932.
7. Suh SO, Marshall CJ, McHugh JV, & Blackwell M (2003) Wood ingestion by passalid beetles in the presence of xylose-fermenting gut yeasts. *Mol Ecol* 12:3137-3145.
8. Suh SO, McHugh JV, Pollock DD, & Blackwell M (2005) The beetle gut: a hyperdiverse source of novel yeasts. *Mycol Res* 109:261-265.
9. Nguyen NH, Suh SO, Marshall CJ, & Blackwell M (2006) Morphological and ecological similarities: wood-boring beetles associated with novel xylose-fermenting yeasts, *Spathaspora passalidarum* gen. sp. nov. and *Candida jeffriesii* sp. nov. *Mycol Res* 110:1232-1241.
10. Deng XX & Ho NW (1990) Xylulokinase activity in various yeasts including *Saccharomyces cerevisiae* containing the cloned xylulokinase gene. *Appl Biochem Biotechnol* 24-25:193-199.
11. Rizzi M, Erlemann P, Bui-Thanh N-A, & Dellweg H (1988) Xylose fermentation by yeasts. *Appl Microbiol Biot* 29:148-154.
12. Rizzi M, Harwart K, Bui-Thanh N-A, & Dellweg H (1989) Purification and properties of the NAD⁺-xylitol-dehydrogenase from the yeast *Pichia stipitis*. *J Ferment Bioeng* 67:20-24.
13. Jeffries TW, *et al.* (2007) Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat Biotechnol* 25:319-326.
14. Ohama T, *et al.* (1993) Non-universal decoding of the leucine codon CUG in several *Candida* species. *Nucleic Acids Res* 21:4039-4045.
15. Santos MA & Tuite MF (1995) The CUG codon is decoded in vivo as serine and not leucine in *Candida albicans*. *Nucleic Acids Res* 23:1481-1486.
16. Sugita T & Nakase T (1999) Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus *Candida*. *Syst Appl Microbiol* 22:79-86.
17. Lockhart SR, Messer SA, Pfaller MA, & Diekema DJ (2008) *Lodderomyces elongisporus* masquerading as *Candida parapsilosis* as a cause of bloodstream infections. *J Clin Microbiol* 46:374-376.

18. Pfaller MA & Diekema DJ (2007) Epidemiology of invasive candidiasis: a persistent public health problem. *Clin Microbiol Rev* 20:133-163.
19. Butler G, *et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* 459:657-662.
20. Hammons DL, Kurtural SK, Newman MC, & Potter DA (2009) Invasive Japanese beetles facilitate aggregation and injury by a native scarab pest of ripening fruits. *Proc Natl Acad Sci U S A* 106:3686-3691.
21. Hittinger CT, *et al.* (2010) Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* 464:54-58.
22. Hittinger CT, Rokas A, & Carroll SB (2004) Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc Natl Acad Sci U S A* 101:14144-14149.
23. Gasch AP, *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11:4241-4257.
24. Wenger JW, Schwartz K, & Sherlock G (2010) Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from *Saccharomyces cerevisiae*. *PLoS Genet* 6:e1000942.
25. Guadalupe Medina V, Almering MJ, van Maris AJ, & Pronk JT (2010) Elimination of glycerol production in anaerobic cultures of a *Saccharomyces cerevisiae* strain engineered to use acetic acid as an electron acceptor. *Appl Environ Microbiol* 76:190-195.
26. Wang ZX, Zhuge J, Fang H, & Prior BA (2001) Glycerol production by microbial fermentation: a review. *Biotechnol Adv* 19:201-223.
27. Norbeck J & Blomberg A (1997) Metabolic and regulatory changes associated with growth of *Saccharomyces cerevisiae* in 1.4 M NaCl. Evidence for osmotic induction of glycerol dissimilation via the dihydroxyacetone pathway. *J Biol Chem* 272:5544-5554.
28. Sanli G, Dudley JI, & Blaber M (2003) Structural biology of the aldo-keto reductase family of enzymes: catalysis and cofactor binding. *Cell Biochem Biophys* 38:79-101.
29. Sonderegger M, Jeppsson M, Hahn-Hägerdal B, & Sauer U (2004) Molecular basis for anaerobic growth of *Saccharomyces cerevisiae* on xylose, investigated by global gene expression and metabolic flux analysis. *Appl Environ Microbiol* 70:2307-2317.
30. Otero JM, *et al.* (2010) Whole genome sequencing of *Saccharomyces cerevisiae*: from genotype to phenotype for improved metabolic engineering applications. *BMC Genomics* 11:723.
31. Wisselink HW, Toirkens MJ, Wu Q, Pronk JT, & van Maris AJ (2009) Novel evolutionary engineering approach for accelerated utilization of glucose, xylose, and arabinose mixtures by engineered *Saccharomyces cerevisiae* strains. *Appl Environ Microbiol* 75:907-914.
32. Gordon D, Abajian C, & Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res* 8:195-202.
33. Lapidus A, *et al.* (2008) POLISHER: An effective tool for using ultra short reads in microbial genome assembly and finishing. in *AGBT* (Marco Island, FL).
34. Huelsenbeck JP & Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754-755.
35. Ronquist F & Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572-1574.

36. Wall DP, Fraser HB, & Hirsh AE (2003) Detecting putative orthologs. *Bioinformatics* 19:1710-1711.
37. Li L, Stoeckert CJ, Jr., & Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178-2189.
38. Wach A, Brachat A, Pohlmann R, & Philippsen P (1994) New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast* 10:1793-1808.
39. Christianson TW, Sikorski RS, Dante M, Shero JH, & Hieter P (1992) Multifunctional yeast high-copy-number shuttle vectors. *Gene* 110:119-122.
40. Krishnan C, *et al.* (2010) Alkali-based AFEX pretreatment for the conversion of sugarcane bagasse and cane leaf residues to ethanol. *Biotechnol Bioeng* 107:441-450.
41. Gasch AP (2002) Yeast genomic expression studies using DNA microarrays. *Methods Enzymol* 350:393-414.
42. Dufour YS, *et al.* (2010) chipD: a web tool to design oligonucleotide probes for high-density tiling arrays. *Nucleic Acids Res* 38 Suppl:W321-325.
43. Gentleman RC, *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5:R80.
44. Gautier L, Cope L, Bolstad BM, & Irizarry RA (2004) affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20:307-315.
45. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article3.
46. Storey JD & Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100:9440-9445.
47. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586-1591.
48. Eisen MB, Spellman PT, Brown PO, & Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95:14863-14868.
49. Jones T, *et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proc Natl Acad Sci U S A* 101:7329-7334.
50. Dujon B, *et al.* (2004) Genome evolution in yeasts. *Nature* 430:35-44.
51. Goffeau A, *et al.* (1996) Life with 6000 genes. *Science* 274:546, 563-547.
52. Wood V, *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature* 415:871-880.

Figure 1



[illegible]

Figure 3

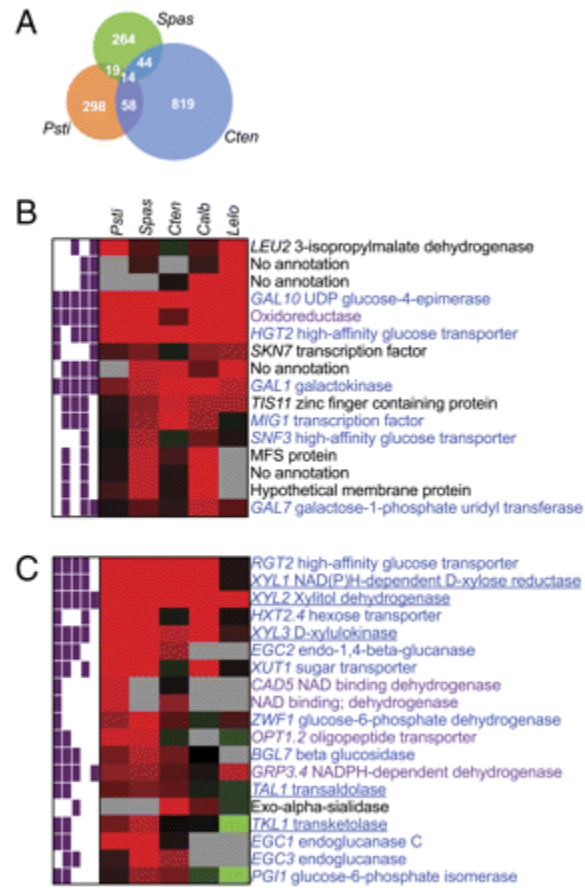


Figure 4

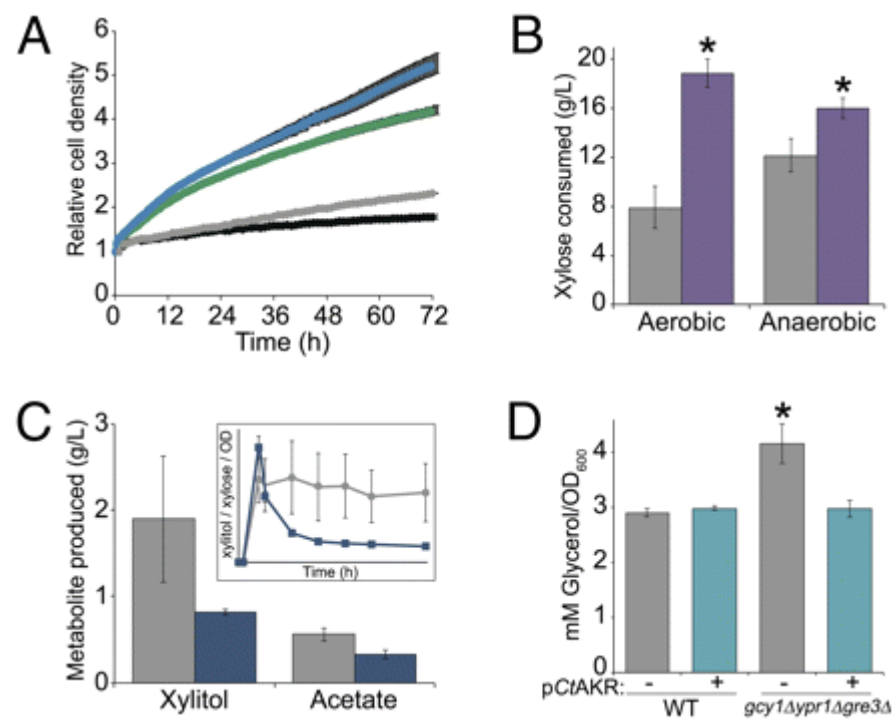


Table 1.

Strain sources and genome statistics

Organism	Strain	Genome size (Mb)	% GC	Total ORFs	Sequencing coverage	Data source	Reference
<i>Sp. passalidarum</i> (<i>Spas</i>)	NRRL Y-27907	13.2	42.0	5983	44X	DOE JGI	This work
<i>C. tenuis</i> (<i>Cten</i>)	NRRL Y-1498	10.7	42.9	5533	27X	DOE JGI	This work
<i>P. stipitis</i> (<i>Psti</i>)	CBS 6054	15.4	42.3	5841	Complete	DOE JGI	¹³
<i>C. albicans</i> (<i>Calb</i>)	W0-1	14.4	33.5	6157	10X	Broad Institute	⁴⁹
<i>C. tropicalis</i> (<i>Ctro</i>)	MYA-3404	14.6	33.1	6258	10X	Broad Institute	¹⁹
<i>C. lusitaniae</i> (<i>Clus</i>)	ATCC 42720	12.1	46.8	5936	9X	Broad Institute	¹⁹
<i>Debaryomyces hansenii</i> (<i>Dhan</i>)	CBS767	12.2	37.5	6887	10X	Genolevures	⁵⁰
<i>L. elongisporus</i> (<i>Lelo</i>)	NRRL YB-4239	15.5	40.4	5796	9X	Broad Institute	¹⁹
<i>P. guilliermondii</i> (<i>Pgui</i>)	ATCC 6260	10.6	44.5	5920	12X	Broad Institute	¹⁹
<i>C. glabrata</i> (<i>Cgla</i>)	CBS 138	12.3	40.5	5215	8X	Genolevures	⁵⁰
<i>Kluyveromyces lactis</i> (<i>Klac</i>)	NRRL Y-1140	10.7	40.1	5327	11X	Genolevures	⁵⁰
<i>S. cerevisiae</i> (<i>Scer</i>)	S288c	12.1	34.4	5695	Complete	SGD	⁵¹
<i>Yarrowia lipolytica</i> (<i>Ylip</i>)	CLIB122	20.5	53.7	6436	10X	Genolevures	⁵⁰
<i>Schizosaccharomyces pombe</i> (<i>Spom</i>)	972h-	12.5	39.6	5004	8X	Wellcome Trust	⁵²

DOE JGI, Department of Energy Joint Genome Institute; SGD, Saccharomyces Genome Database.

Supplementary Information Appendix for:

Comparative genomics of xylose-fermenting fungi for enhanced biofuel production

Dana J. Wohlbach^{1,2}, Alan Kuo³, Trey K. Sato², Katlyn M. Potts¹, Asaf Salamov³, Kurt M. LaButti³, Hui Sun³, Alicia Clum³, Jasmyan Pangilinan³, Erika Lindquist³, Susan Lucas³, Alla Lapidus³, Mingjie Jin^{4,5}, Christa Gunawan^{4,5}, Venkatesh Balan^{4,5}, Bruce E. Dale^{4,5}, Thomas W. Jeffries², Robert Zinkel², Kerrie W. Barry³, Igor V. Grigoriev³, Audrey P. Gasch^{1,2}

¹*Department of Genetics, University of Wisconsin-Madison, 425G Henry Mall, Madison, WI 53706, USA.* ²*Great Lakes Bioenergy Research Center, 1550 Linden Drive, Madison, WI 53706, USA.*

³*United States Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA.* ⁴*Biomass Conversion Research Laboratory, Department of Chemical Engineering and Materials Science, Michigan State University, University Corporate Research Complex, 3900 Collins Road, Lansing, MI 48910, USA.* ⁵*Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI 48824, USA.*

Corresponding Author:

Audrey P. Gasch
Department of Genetics
University of Wisconsin-Madison
425G Henry Mall
Madison, Wisconsin 53706 USA
p: (608) 265-0859
f: (608) 262-2976
agasch@wisc.edu

SI Notes

Note S1: Characterization of engineered *Scer* strains. We selected ten *Spathaspora passalidarum* (*Spas*) and *Candida tenuis* (*Cten*) genes for characterization in *Saccharomyces cerevisiae* (*Scer*) engineered with *PsXYL123*, using the following selection criteria. First, we selected 109 OGGs that were induced in response to xylose in at least two of the three xylose-fermenting species (*Spas*, *Cten*, and *Pichia stipitis*, *Psti*), but were not induced in the non-xylose-utilizing species (*Candida albicans*, *Calb*; and *Lodderomyces elongisporus*, *Lelo*). Next, we examined the coding sequence of the *Spas* and *Cten* genes in these OGGs. We required the genes to contain no CUG codons, enabling heterologous expression in *Scer* without the need for codon optimization. Second, we prioritized our list for genes that were most likely to be involved in some aspect of carbon metabolism based on predicted annotations and protein domain analysis. This list included a *Cten* aldo-keto reductase (*CtAKR*), a *Cten* galactokinase (*CtGalK*), a *Spas* glucose-6-phosphate dehydrogenase (*SpGPD*), a *Spas* UDP-glucose-epimerase (*SpUGE*), a *Spas* glucose phosphate isomerase (*SpGPI*), *RGT2* from *Spas* and *Cten* (*SpRGT2* and *CtRGT2*), and *YBR2* from *Spas* and *Cten* (*SpYBR2* and *CtYBR2*). An additional three genes were included because they were also from the list of 43 OGGs present in xylose-fermenters but absent in non-xylose-utilizers: a *Spas* unannotated protein (*SpNA*), a *Spas* oxidoreductase (*SpOR*), and a *Spas* hexose transporter (*SpHXT*).

SI Materials and Methods

Data Sources. The complete genome sequences of twelve Ascomycete yeasts were obtained and downloaded from their respective online databases (Table 1).

Genome and EST Sequencing and Assembly. Details of library construction and sequencing can be found in the main text and at the JGI website (<http://www.jgi.doe.gov/>). The 13.1 Mb assembly of *Spas* consists of 26 contigs arranged in eight scaffolds. The genome was sequenced to 43.77X coverage (1.78X of Sanger and 41.99X of 454). A total of 53 Sanger finishing reads were produced to close gaps, to resolve repetitive regions, and to raise the quality of the finished sequence. Assembly completeness was confirmed by mapping 8,089 out of 8,349 EST contigs (97%) with 90% identity and 85% coverage.

The *Cten* genome was sequenced to 26.9X coverage (1.13X of Sanger and 24.97X of 454). 15,126 Sanger, 439,285 standard, and 634,050 paired-end pyrosequencing reads were combined into a 10.6 Mb assembly consisting of 1065 contigs organized in 61 scaffolds representing eight chromosomes. Assembly completeness was confirmed by mapping 7,493 out of 8,230 EST contigs (91%) with 90% identity and 85% coverage.

For *Spas*, one EST library consisting of 1,050,790 initial sequence reads led to a set of 1,020,921 "good" reads assembled into 8,349 contigs. For *Cten*, one EST library consisting of 987,487 reads resulted in 964,346 "good" reads assembled into 8,230 contigs. These ESTs and contigs were used in annotation of the corresponding genomes.

Genome Annotation. Genomic assembly scaffolds were masked using RepeatMasker (<http://www.repeatmasker.org/>) and a standard RepeatMasker library of 234 fungal transposable elements (1). tRNAs were predicted using tRNAscan-SE (2). Using repeat-masked assembly, several gene prediction programs were used: *ab initio* FGENESH (3); homology-based FGENESH+ (3) and Genewise (4) seeded by BLASTX alignments against GenBank's database of non-redundant proteins (NR; <http://www.ncbi.nlm.nih.gov/BLAST/>); and cDNA-based EST_map (<http://www.softberry.com/>) seeded by the EST contigs (*SI Appendix*, Table S12).

Genewise models were completed using scaffold data to find start and stop codons. EST BLAT alignments (5) were used to extend, verify, and complete the predicted gene models. The resulting set of models was then filtered for the "best" models, based on EST and homology support, to produce a non-redundant representative set. This representative set was subject to further analysis and manual curation. High (> 90%) proportions of the models are complete with start and stop codons, consistent with ESTs, and supported by similarity with proteins from the NCBI non-redundant protein set (*SI Appendix*, Table S13).

Analysis indicated that both species display the alternate codon decoding of CUG for serine rather than leucine (see main text). Therefore, all predicted gene models were translated using alternative translation table 12 (CUG → Ser) and functionally annotated using SignalP (6), TMHMM (7), InterProScan (8), BLASTP (9) against NR, and hardware-accelerated double-affine Smith-Waterman alignments (deCypherSW; http://www.timelogic.com/decypher_sw.html) against Swiss-Prot (<http://www.expasy.org/sprot/>), KEGG (10), and KOG (11). KEGG hits were used to map EC numbers (<http://www.expasy.org/enzyme/>), and Interpro and Swiss-Prot hits were used to map GO

terms (<http://www.geneontology.org/>; *SI Appendix*, Table S14). Multi-gene families were predicted with the Markov clustering algorithm to cluster the proteins, using BLASTP alignment scores between proteins as a similarity metric (12). Manual curation of the automated annotations was performed using the web-based interactive editing tools of the JGI Genome Portal to assess predicted gene structures, assign gene functions, and report supporting evidence.

Syntenic regions were identified as those containing at least three genes and with 50% of all genes in the region conserved and syntenic in each species. Single species expansions are defined as 3× gene counts in one species compared to two others.

Codon Usage Determination. tRNA gene sequences were identified with the program tRNAscan-SE v1.21 (2). We produced a multiple alignment of the tRNA genes using ClustalW v1.81 (13) with the default settings (*SI Appendix*, Fig. S2A). The alignment shows unambiguously that the tRNA_{CAG} from *Spas* and *Cten* are orthologous to the serine encoding tRNA from other CUG-utilizing species and display the known polymorphisms that converted the codon recognition of this tRNA (14-16).

We also examined CUG codon usage by comparing *Scer* (as the reference), a well-characterized species that uses CUG to encode leucine, to the other 13 species (queries) in our analysis using custom perl scripts. First, we identified all CUG-containing genes within each of the thirteen query species. If the CUG-containing gene had a one-to-one ortholog in *Scer* (as assigned by reciprocal smallest distance, RSD; 17), pairwise protein alignments of the two genes were generated with ClustalW v1.81 (13). We then converted the protein sequence of the query species back to the corresponding DNA sequence. For each CUG codon in the query sequence, we identified the corresponding orthologous amino acid from *Scer* and counted the total number of CUG codons aligned to either leucine or serine and report this value as a percentage of the total aligned CUG codons (*SI Appendix*, Fig. S2C). A clear delineation was observed for the species known to decode CUG with leucine (*C. glabrata*, *Cgla*; *Kluyveromyces lactis*, *Klac*; *Yarrowia lipolytica*, *Ylip*; and *Schizosaccharomyces pombe*, *Spom*), and those known to decode CUG with serine.

Species Phylogeny. We estimated the phylogeny of the 14 Ascomycete species in our analysis using the protein sequences of 136 orthologs present in single copy in all species, identified using our ortholog assignment method described below. For each set of orthologous proteins, we produced multiple alignments using ClustalW v1.81 (13) with the default settings and identified conserved alignment blocks using Gblocks v0.19b (18). The final concatenated alignment used for phylogenetic reconstruction analysis consisted of 28,166 amino acid positions. ModelGenerator v0.85 (19) was used to identify the optimum model of amino acid substitution (RtRev+G+F) for maximum likelihood phylogeny reconstruction. Phylogenies were constructed using the maximum likelihood method with the program RAxML v7.0.4 (20) and using the Bayesian method with the program MrBayes v3.1.2 (21, 22). For both methods, we constrained the topology to require the outgrouping of *Spom*. RAxML was executed with 100 rapid bootstrap inferences followed by a slow ML search using the RtRev+G+F model of amino acid substitution. MrBayes was executed for 500,000 generations with a sample frequency of 10 and a burn-in of 1250 samples using the mixed model of amino acid substitution with a mixture of invariant and gamma distributed rates across sites. Both methods produced identical topologies; consequently, only the ML tree is shown in Fig. 1B.

Ortholog Assignment and Resolution. We created orthologous gene groups (OGGs) using a modified RSD (17) and OrthoMCL (23) method. RSD parameters: significance threshold, 10^{-5} ; alignment threshold, 0.3. OrthoMCL parameters: significance threshold, 10^{-5} ; inflation parameter, 1.5. Pairwise one-to-one orthologs were assigned with the RSD method using four reference species: *Scer*, *Psti*, *Calb*, and *Spom*. These species were chosen for their complete and/or well-annotated genomes, and because they are representative of the Ascomycetes in our study. Pairwise OGGs (including orthologs and paralogs) were also assigned with the OrthoMCL method using the same four reference species. Results from the two methods were compared and combined using a custom perl script to maximize high confidence assignments (true positives) and minimize low confidence assignments (false positives).

In approximately 85% of comparisons, the ortholog assignments between RSD and OrthoMCL agreed perfectly. In cases when the two did not agree, the four reference genomes were used to resolve OGGs by comparing the different results from each reference, and determining a majority consensus when possible. Approximately 150 of the OGGs remain unresolved. Within the amino acid sequences of the genes in these OGGs, there is not sufficient phylogenetic information to determine if the OGG consists of genes derived from a single ancestral gene, or if there are multiple ancestral gene signatures in the OGG. These OGGs generally contain large families of genes with highly similar sequence (*e.g.* sugar transporters). The result of this analysis is a list of 12,038 OGGs containing the entire set of 81,907 genes. Over 90% (74,633) of the genes are contained within 5,749 multi-species OGGs (*SI Appendix*, Fig. S3).

To avoid false negative calls of ortholog absence, we devised a method implemented with custom perl scripts. For each species not assigned a gene in a particular OGG, we examined the complete genome sequence of that species through multiple tBLASTn (24) runs using the protein sequence of all other genes in the OGG as queries. We filtered the results to identify putative missed ortholog assignments (false negatives) attributed to incomplete or incorrect genome sequence or genome annotation. These putative new orthologs were assigned a ‘flag’ for possible orthology and are indicated in blue in all OGG figures.

We classified the OGGs as multi-species or single-species (*SI Appendix*, Table S3). Single-species OGGs are comprised of expansions (a group of paralogous genes from a single species, which are likely to represent real genes) and orphans (genes with no recognizable homolog in our data set that may be annotation artifacts or novel genes.)

Evolutionary analyses. We generated Bayesian gene trees for each OGG using MrBayes v3.1.2 (21, 22) executed for 100,000 generations with a sample frequency of 10 and a burn-in of 250 samples using the mixed model of amino acid substitution with a mixture of invariant and gamma distributed rates across sites. We estimated non-synonymous nucleotide substitutions (dN) and synonymous substitutions (dS) using PAML (25) implemented with custom perl scripts and calculated average dN/dS over all lineages within the xylose-utilizers or the non-utilizers.

Xylose fermentation. For xylose fermentation measurements in untransformed yeast species (*SI Appendix*, Fig. S6), cells were initially grown to saturation for 36 h in YPD (10 g/L yeast extract, 20 g/L peptone, 20 g/L glucose) at 30°C, washed once in SC (synthetic complete; 1.7 g/L yeast nitrogen

base, essential amino acids, 1 g/L ammonium sulfate), and split into two cultures: SC + 8% glucose and SC + 8% xylose. Then, 50 mL of cells were resuspended in a 125-mL Erlenmeyer flask to an OD₆₀₀ of 10 and were incubated at 30°C in an orbital shaker at 100 rpm. Samples were taken every 8 h for 56 h. For xylose fermentation measurements in engineered *Scer* (Fig. 4 and *SI Appendix*, Fig. S10 and S11), 50 mL of cells were resuspended in YPXD (5 g/L yeast extract and 10 g/L tryptone, 58 g/L glucose, 28 g/L xylose) in an airlocked 125-mL Erlenmeyer flask to an OD₆₀₀ of 2 and were incubated at 30°C in an orbital shaker at 150 rpm. Samples were taken every 12-24 h for 168 h.

Concentrations of ethanol in untransformed yeast species were determined using an Agilent Technologies 7890A gas chromatograph with a 7693 autosampler and flame ionization detector (FID). The instrument was operated and data acquired using Agilent GC Chemstation version B.04.02. The GC Inlet was equipped with a 4mm ID deactivated split liner with deactivated glass wool (Restek, Inc) and held at 250°C throughout the run. The helium carrier gas flow through the column was maintained at 1 mL/min with electronic pressure control. A 1 µL sample was injected with a split ratio of 20:1. The GC column was a Stabilwax-DA 30 M x 0.32 mm ID x 0.5 µm stationary phase (Restek, Inc). The GC oven program was as follows: Initial temperature of 110°C was held for 3.5 minutes after injection, increased at 60°C/min to 250°C and held for 5 minutes. The oven was equilibrated at the starting temperature for 3 minutes between runs. The flame ionization detector parameters were: detector temperature 300°C, hydrogen (fuel gas) flow 30 mL/min, air flow 400 mL/min, nitrogen makeup gas flow 25 mL/min.

SI References

1. Jurka J, *et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462-467.
2. Lowe TM & Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 2:955-964.
3. Salamov AA & Solovyev VV (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res* 10:516-522.
4. Birney E & Durbin R (2000) Using GeneWise in the *Drosophila* annotation experiment. *Genome Res* 10:547-548.
5. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Res* 12:656-664.
6. Nielsen H, Engelbrecht J, Brunak S, & von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* 10:1-6.
7. Melén K, Krogh A, & von Heijne G (2003) Reliability measures for membrane protein topology prediction algorithms. *J Mol Biol* 327:735-744.
8. Zdobnov EM & Apweiler R (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847-848.
9. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403-410.
10. Kanehisa M, *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 36:D480-484.
11. Koonin EV, *et al.* (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* 5:R7.
12. Enright AJ, Van Dongen S, & Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575-1584.
13. Thompson JD, Higgins DG, & Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673-4680.
14. Ohama T, *et al.* (1993) Non-universal decoding of the leucine codon CUG in several *Candida* species. *Nucleic Acids Res* 21:4039-4045.
15. Santos MA & Tuite MF (1995) The CUG codon is decoded in vivo as serine and not leucine in *Candida albicans*. *Nucleic Acids Res* 23:1481-1486.
16. Sugita T & Nakase T (1999) Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus *Candida*. *Syst Appl Microbiol* 22:79-86.
17. Wall DP, Fraser HB, & Hirsh AE (2003) Detecting putative orthologs. *Bioinformatics* 19:1710-1711.
18. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540-552.
19. Keane TM, Creevey CJ, Pentony MM, Naughton TJ, & McLnerney JO (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that *ad hoc* assumptions for choice of matrix are not justified. *BMC Evol Biol* 6:29.
20. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688-2690.
21. Huelsenbeck JP & Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754-755.
22. Ronquist F & Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572-1574.

23. Li L, Stoeckert CJ, Jr., & Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178-2189.
24. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.
25. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586-1591.
26. Bendtsen JD, Nielsen H, von Heijne G, & Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340:783-795.
27. Sanli G, Dudley JI, & Blaber M (2003) Structural biology of the aldo-keto reductase family of enzymes: catalysis and cofactor binding. *Cell Biochem Biophys* 38:79-101.
28. Dietrich FS, *et al.* (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* 304:304-307.
29. Jones T, *et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proc Natl Acad Sci U S A* 101:7329-7334.
30. Jackson AP, *et al.* (2009) Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. *Genome research* 19:2231-2244.
31. Dujon B, *et al.* (2004) Genome evolution in yeasts. *Nature* 430:35-44.
32. Butler G, *et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* 459:657-662.
33. Kellis M, Birren BW, & Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428:617-624.
34. De Schutter K, *et al.* (2009) Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nature biotechnology* 27:561-566.
35. Jeffries TW, *et al.* (2007) Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat Biotechnol* 25:319-326.
36. 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.
37. Goffeau A, *et al.* (1996) Life with 6000 genes. *Science* 274:546, 563-547.
38. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article3.
39. Gasch AP, *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11:4241-4257.
40. Boyle EI, *et al.* (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 20:3710-3715.

SI Figures

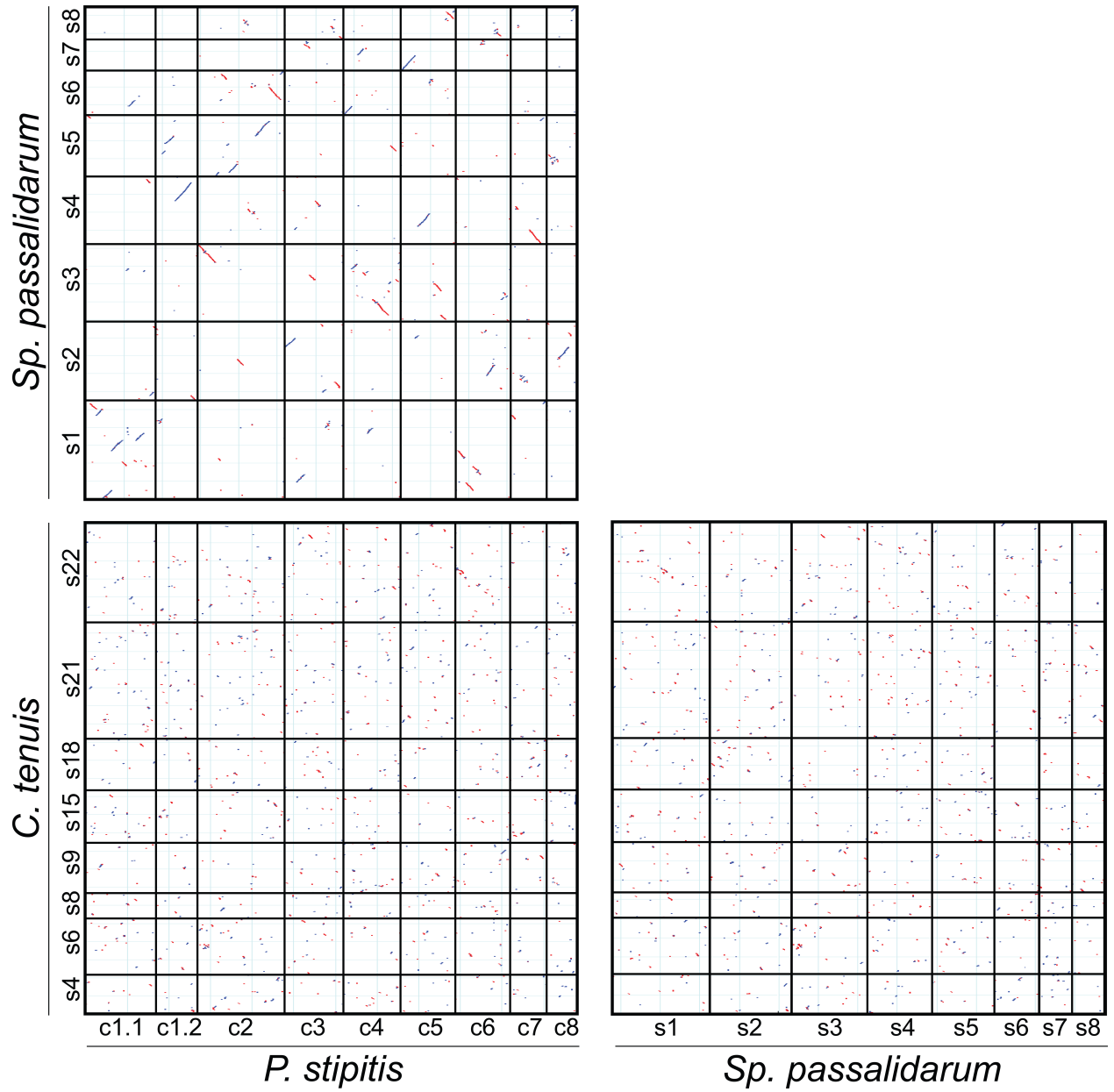


Fig. S1. Pairwise genome-wide synteny dot plots for xylose-fermenting fungi. Diagonal lines display the homologous regions between the two genomes, either on the same strand (blue), or on opposite strands (red). Black grid lines indicate scaffold (s) or chromosome (c) boundaries. Longer regions of co-linearity exist between *Spas* and *Psti*, supporting the constructed species phylogeny.

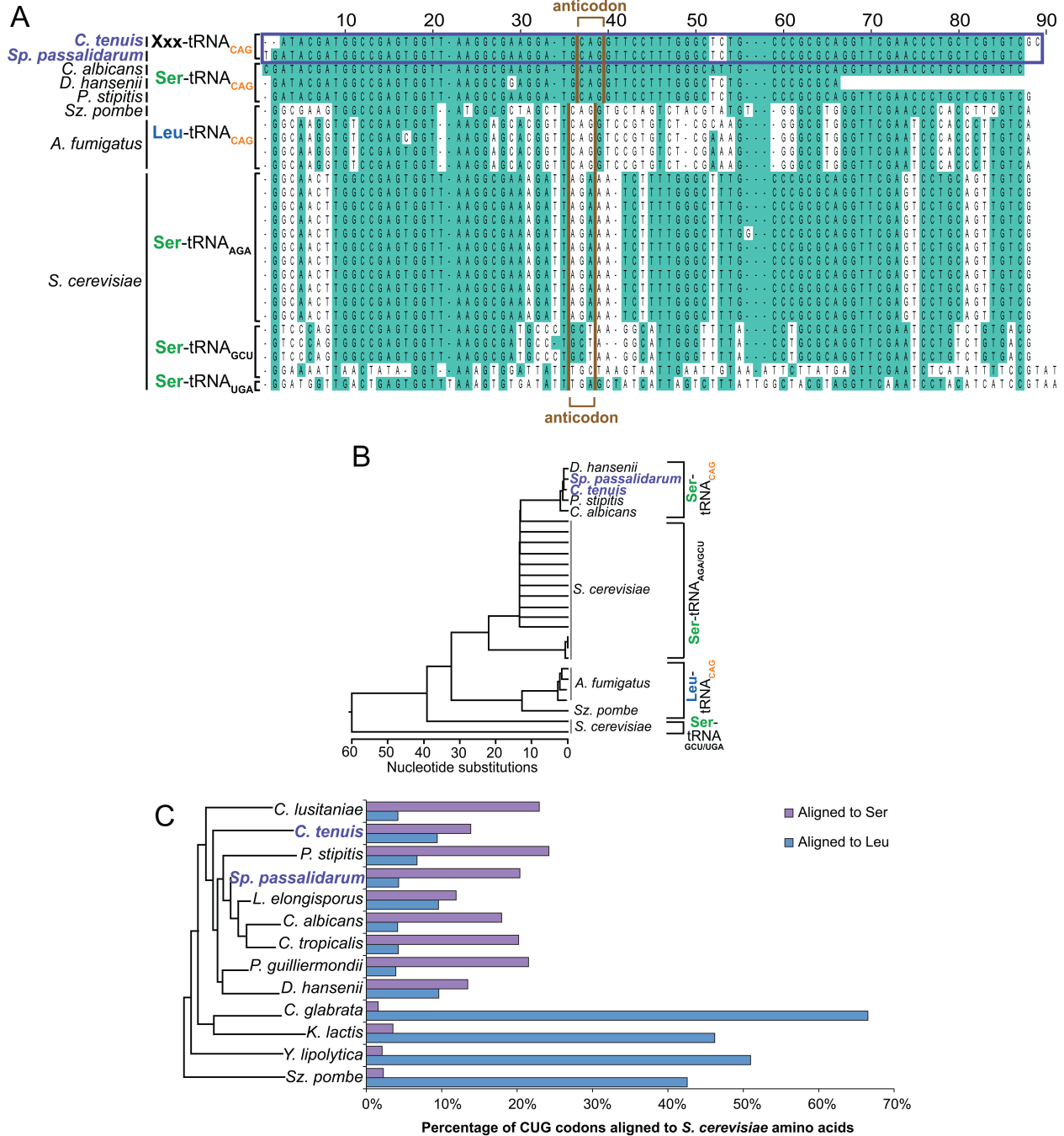


Fig. S2. Analysis of CUG codon usage in *Cten* and *Spas*. (A) ClustalW alignment of 26 tRNAs from several Ascomycete species. The tRNA_{CAG} from *Cten* and *Spas* is outlined in purple. The anticodon loop is outlined in brown. Residues identical to those in the Ser-tRNA_{CAG} from *Calb* are shaded aqua. (B) Neighbor joining tree of tRNAs created from ClustalW alignment in (A). (C) Alignment of CUG codons to orthologous *Scer* amino acids (AAs). For each species, the AA sequence of protein coding genes containing one or more CUG codons was aligned to the orthologous *Scer* protein. The location of each CUG codon was mapped to the orthologous AA position in *Scer*, and the fraction of CUG codons aligned to serine (purple), or leucine (blue) is shown.

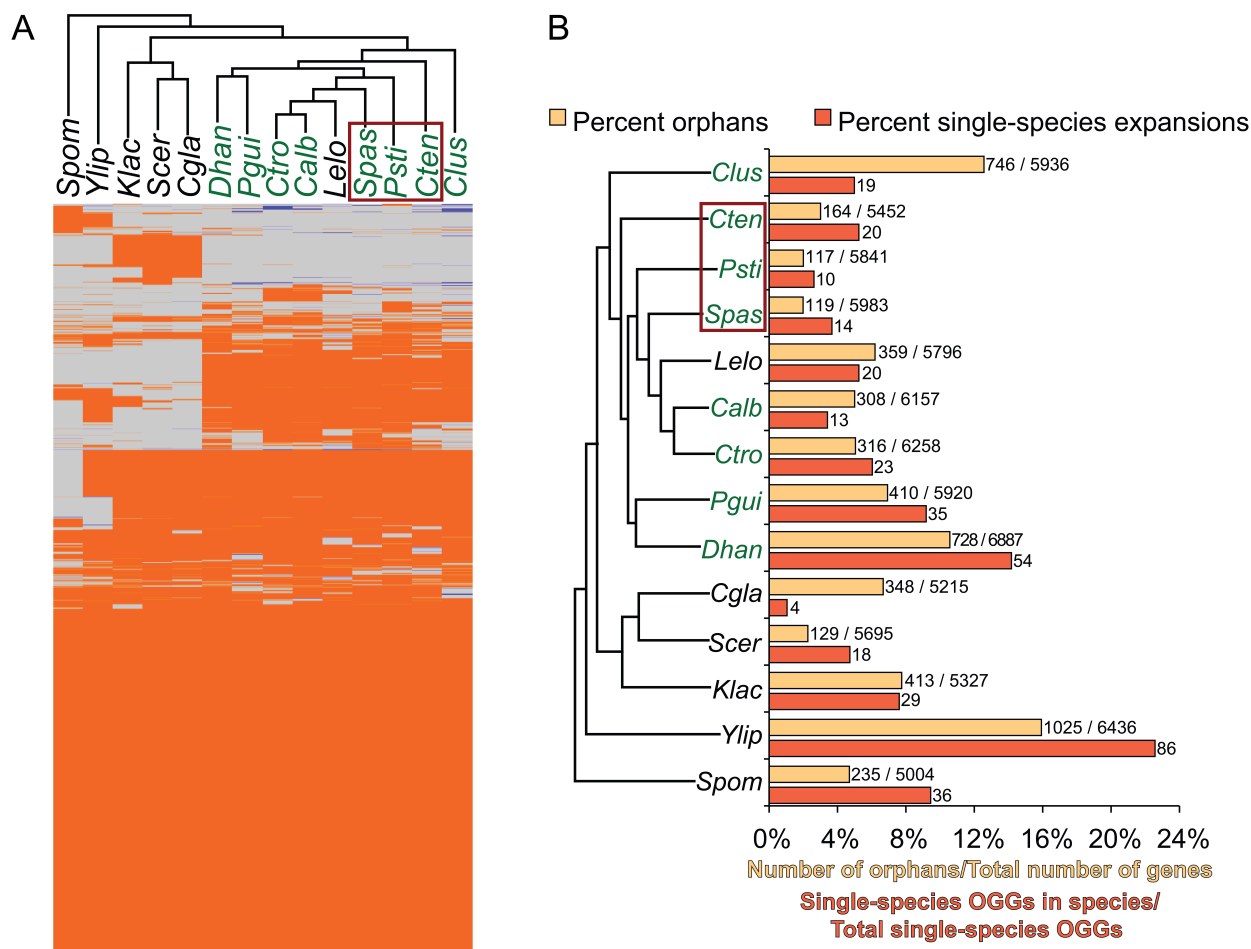


Fig. S3. Ortholog assignment across fourteen Ascomycete yeasts. (A) Patterns of ortholog presence (orange) or absence (grey) for all 5,749 multi-species OGGs, as revealed by hierarchical clustering of OGGs. Blue indicates BLAST homology despite no ortholog call. **(B)** Patterns of single-species OGGs. For orphan genes (yellow), the total number of orphans is given along with the total number of genes in the genome; the bar represents the number of orphans in that species / the total number of genes in the genome within that species. For expansions (red), the total number of single-species OGGs is given; the bar represents the number of single-species expansions in that species / the total number of single-species OGGs in the entire dataset. Species abbreviations as in Table 1. Green text, xylose-growing species; red box, xylose-fermenting species.

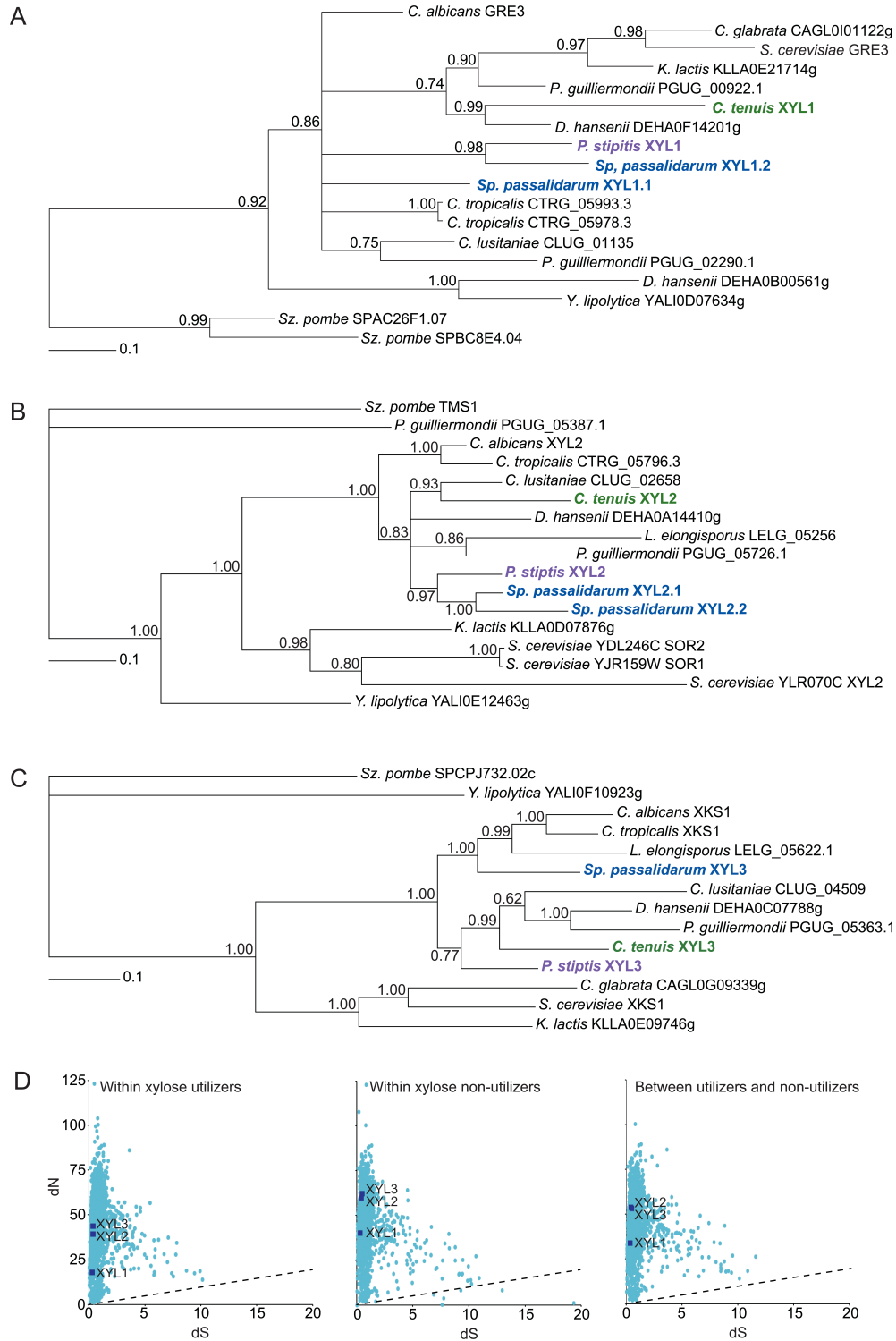


Fig. S4. Known xylose-utilization genes are present throughout the Ascomycetes. Bayesian gene trees were reconstructed for the *XYL1* (A), *XYL2* (B), and *XYL3* (C) OGGs using MrBayes (21, 22). Posterior probabilities are indicated on all branches. (D) Plots of dS versus dN for 2664 OGGs present in both xylose utilizers and non-utilizers. Measurements of dN and dS were computed between all pairs of all genes in each OGG with PAML v4.3 (25). The average dN and dS was calculated for all pairs of genes within the xylose utilizers, within the xylose non-utilizers, and between the xylose utilizers and non-utilizers. The dashed line indicates dN/dS = 1.

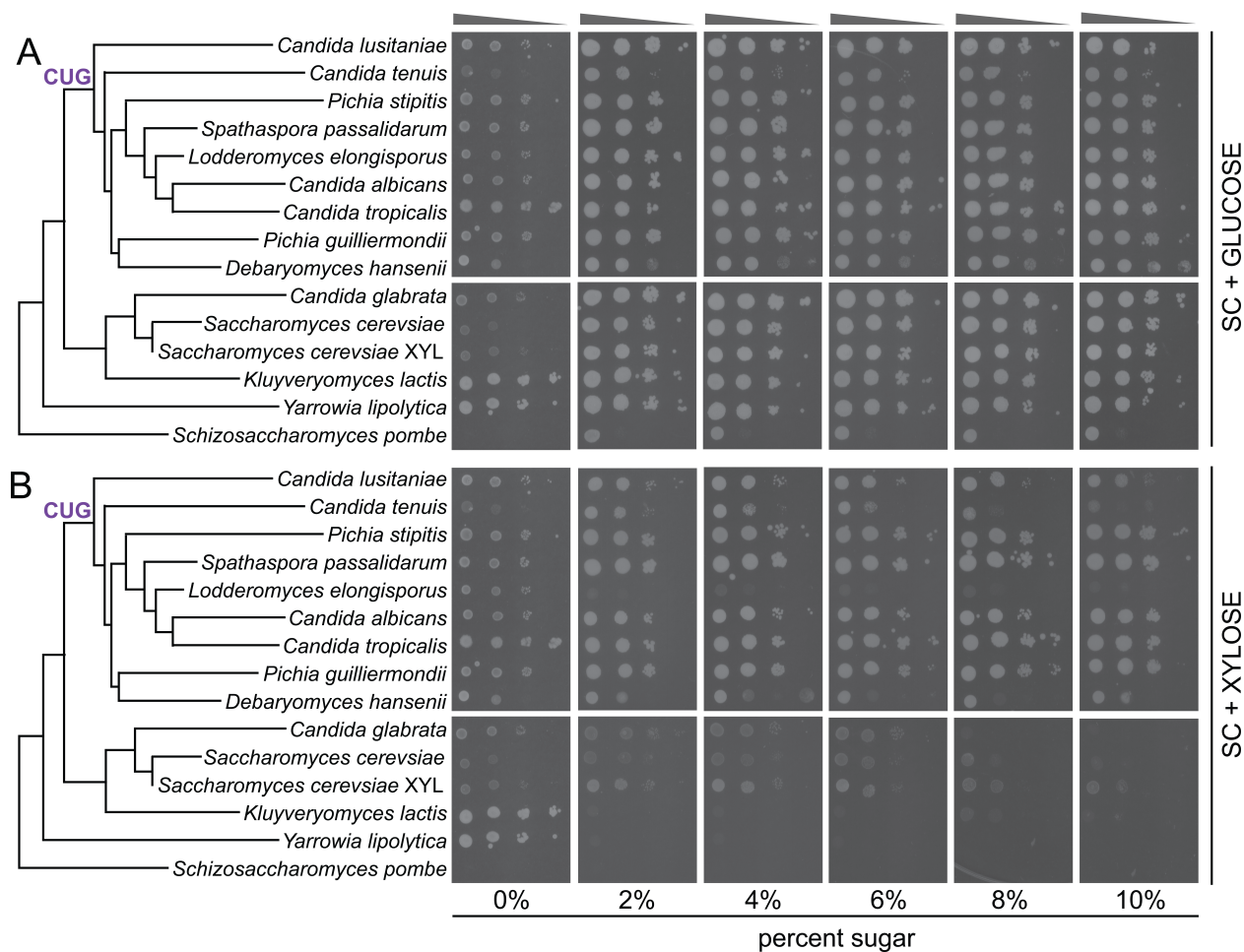


Fig. S5. Xylose growth phenotypes of fourteen Ascomycete yeasts. Cultures were initially grown in liquid YPD (1% yeast extract, 2% peptone, 2% glucose). Cultures were washed once and spotted onto plates containing 0%, 2%, 4%, 6%, 8%, or 10% glucose (**A**) or xylose (**B**) in minimal media. Growth was scored after three days at 30°C. Serial dilutions of cultures are indicated by grey triangles. *Scer* XYL, engineered strain with *PsXYL123*; SC, Synthetic Complete.

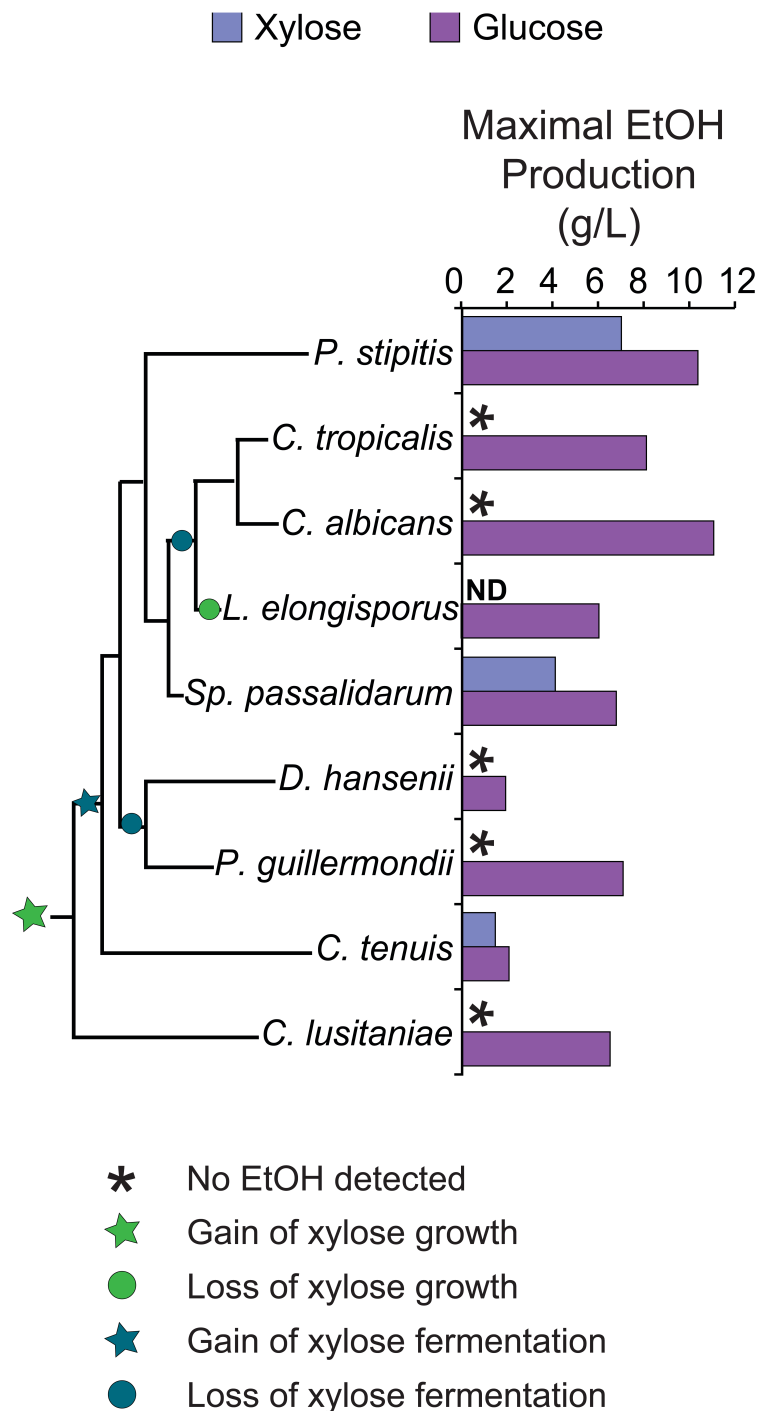


Fig. S6. EtOH fermentation in Hemiascomycete yeasts. Maximal EtOH production from 8% xylose (blue) or 8% glucose (purple) over 55 hours. High-density cultures were grown in a microaerobic environment (minimal shaking at 30°C), and EtOH concentration was measured by gas chromatograph every eight hours. Values represent the average (n = 3). Limit of EtOH detection is 0.2 g/L. The most parsimonious explanation for evolution of xylose growth and fermentation is indicated with green and blue symbols, respectively. ND, no data was measured for *Lelo* in xylose, as it does not grow in this condition.

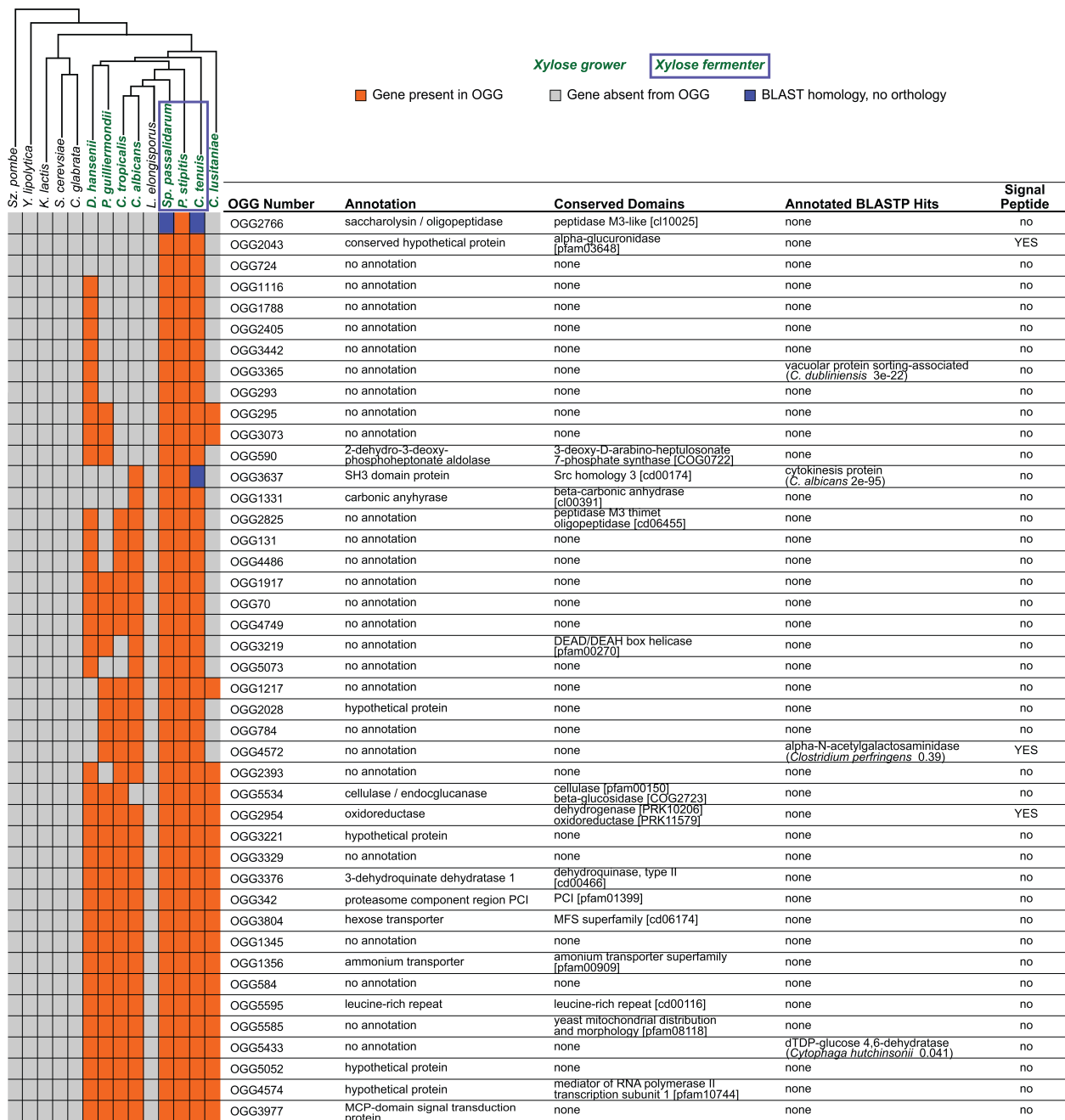


Fig. S7. Comparison of the 43 OGGs present in all xylose-growing species and absent from all species unable to grow on xylose. The amino acid sequence of each gene in *Psti*, *Spas*, and *Cten* was examined by InterProScan (8) to identify conserved protein domains, by BLAST (9) to identify homologous proteins, and by SignalP v3.0 (26) to identify signal peptide sequences. The summary of these analyses is given in the table adjacent to the image showing patterns of OGG presence (orange), absence (grey), and BLAST homology despite no ortholog call (blue) across the phylogeny. For annotated BLASTP hits, the species in which the BLAST hit occurred and the E-value is given. Sequences of these 43 OGGs are available in *SI Appendix*, Dataset S1.

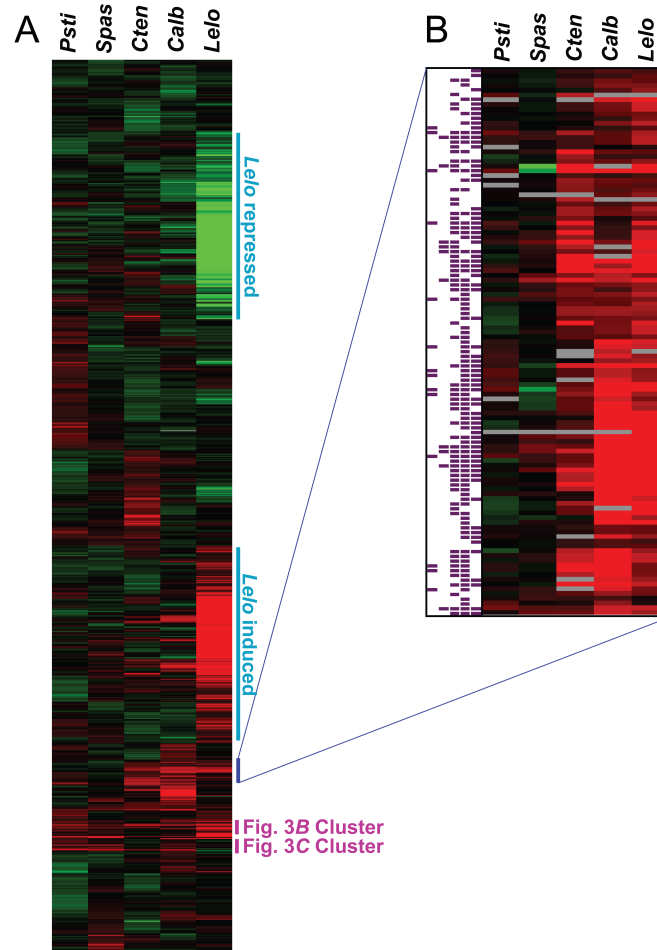


Fig S8. Transcriptome analysis of xylose growing cultures. Three xylose-fermenting species (*Psti*, *Spas*, *Cten*), one xylose-growing, non-fermenting species (*Calb*) and one non-xylose-growing species (*Lelo*) were grown for three generations in 2% xylose or 2% glucose. Expression levels of three biological replicate samples were measured and hierarchically clustered. The averaged $\log_2$ fold expression change of xylose versus glucose is shown for all OGGs present in three or more species. Red boxes, higher expression in xylose; green boxes, lower expression on xylose; Grey boxes, no ortholog present. (A) Hierarchical clustering of all 6777 rows of expression data. Five relevant clusters are indicated. For *Lelo* induced/repressed clusters, see *SI Appendix*, Table S9. (B) Zoom-in of cluster of genes commonly induced in *Cten*, *Calb*, and *Lelo*. Purple blocks indicate statistically significant measurement (t-test, FDR < 0.05) in *Psti*, *Spas*, *Cten*, *Calb*, and *Lelo*. See also *SI Appendix*, Table S10.

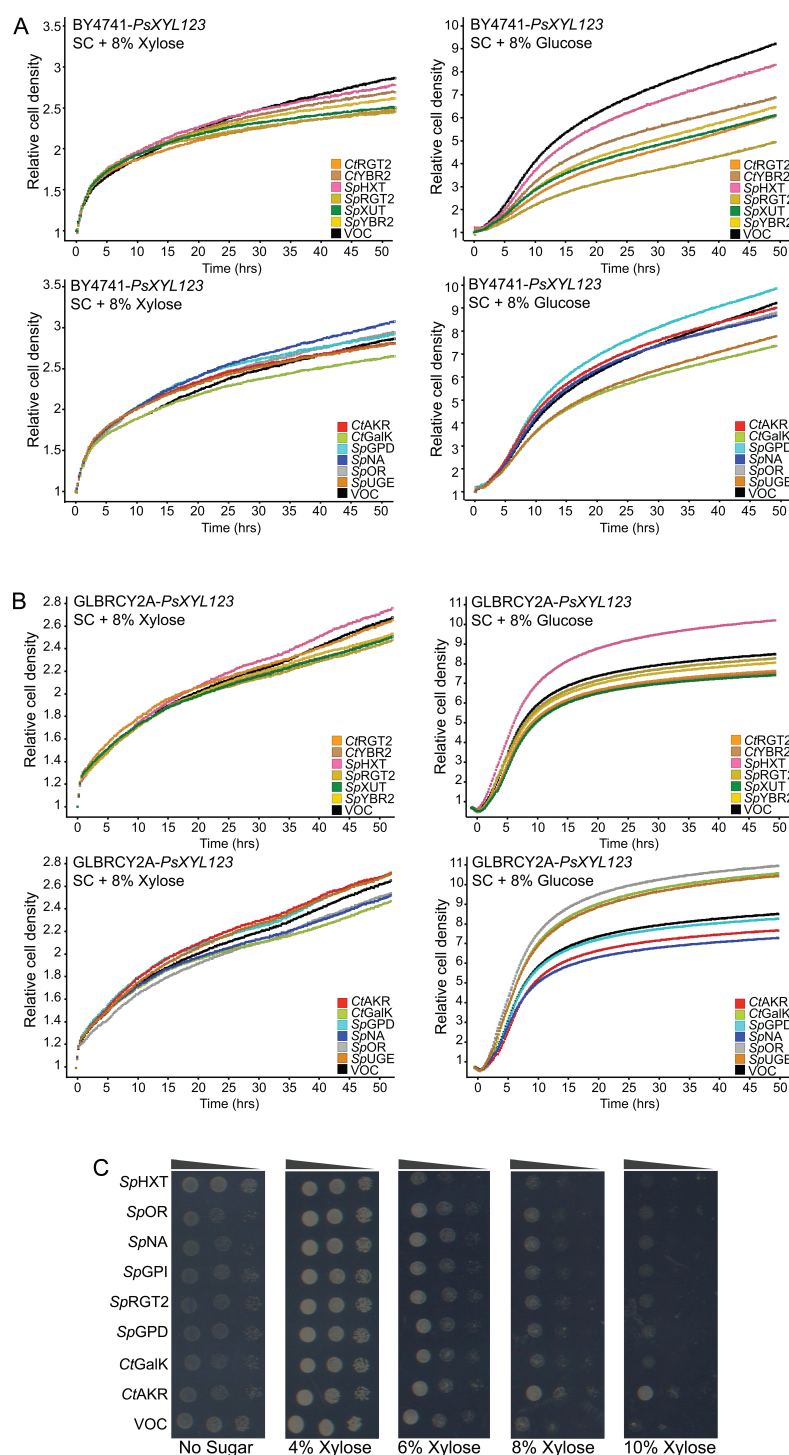


Fig. S9. Screening of candidate genes in engineered *Scer* grown in medium containing 8% xylose or 8% glucose. BY4741+*PsXYL123* (A) or GLBRCY0A+*PsXYL123* (B) strains transformed with multi-copy plasmids expressing the indicated genes were grown in the indicated media, and cell densities were measured every 5-10 minutes for 50 hours. Data represent the average ($n = 4$). Several genes had a negative effect on growth, likely due to the increased burden protein overexpression places on cells. (C) GLBRCY0A+*PsXYL123* cells overexpressing the indicated genes were spotted onto synthetic complete (SC) solid media with the indicated concentrations of xylose. Images were taken after 3 d growth at 30°C. Gene abbreviations as in *SI Appendix*, Note S1. VOC, vector only control.

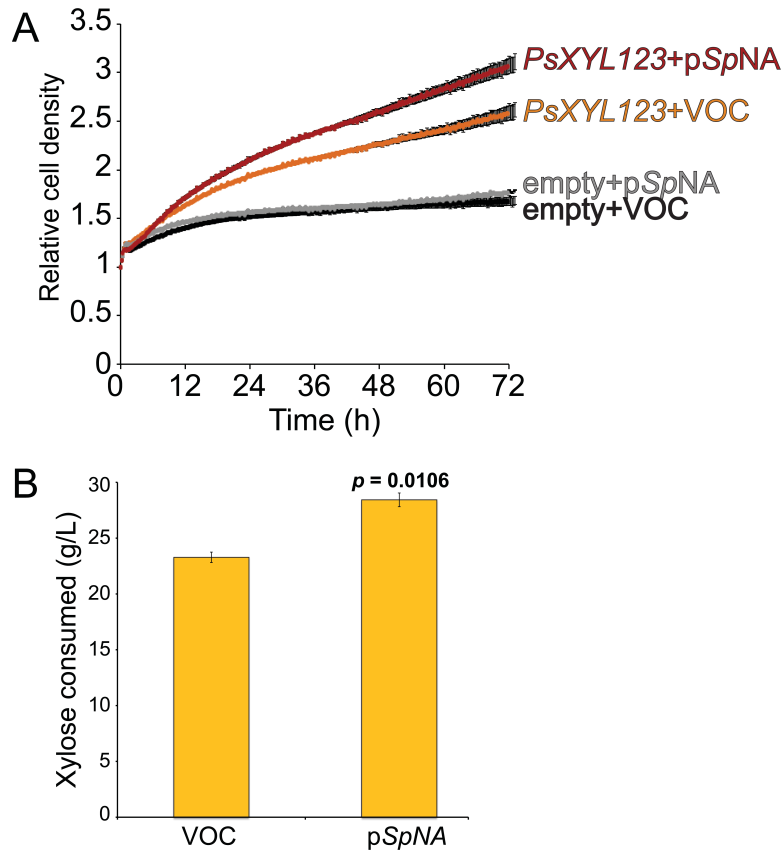


Fig. S10. A *Spas* unannotated protein (*SpNA*) improves *Scer* xylose consumption. (A) Average $\pm$ SD ($n = 4$) growth on 8% xylose of *Scer* strains BY4741+*PsXYL123*+p*SpNA* (red), BY4741+*PsXYL123*+VOC (vector only control; orange), BY4741+p*SpNA* (grey), and BY4741+VOC (black). (B) Average $\pm$ SD ($n = 3$) xylose consumed after 72 hours growth for BY4741+*PsXYL123*+p*SpNA* ('p*SpNA*') or BY4741+*PsXYL123*+VOC ('VOC').

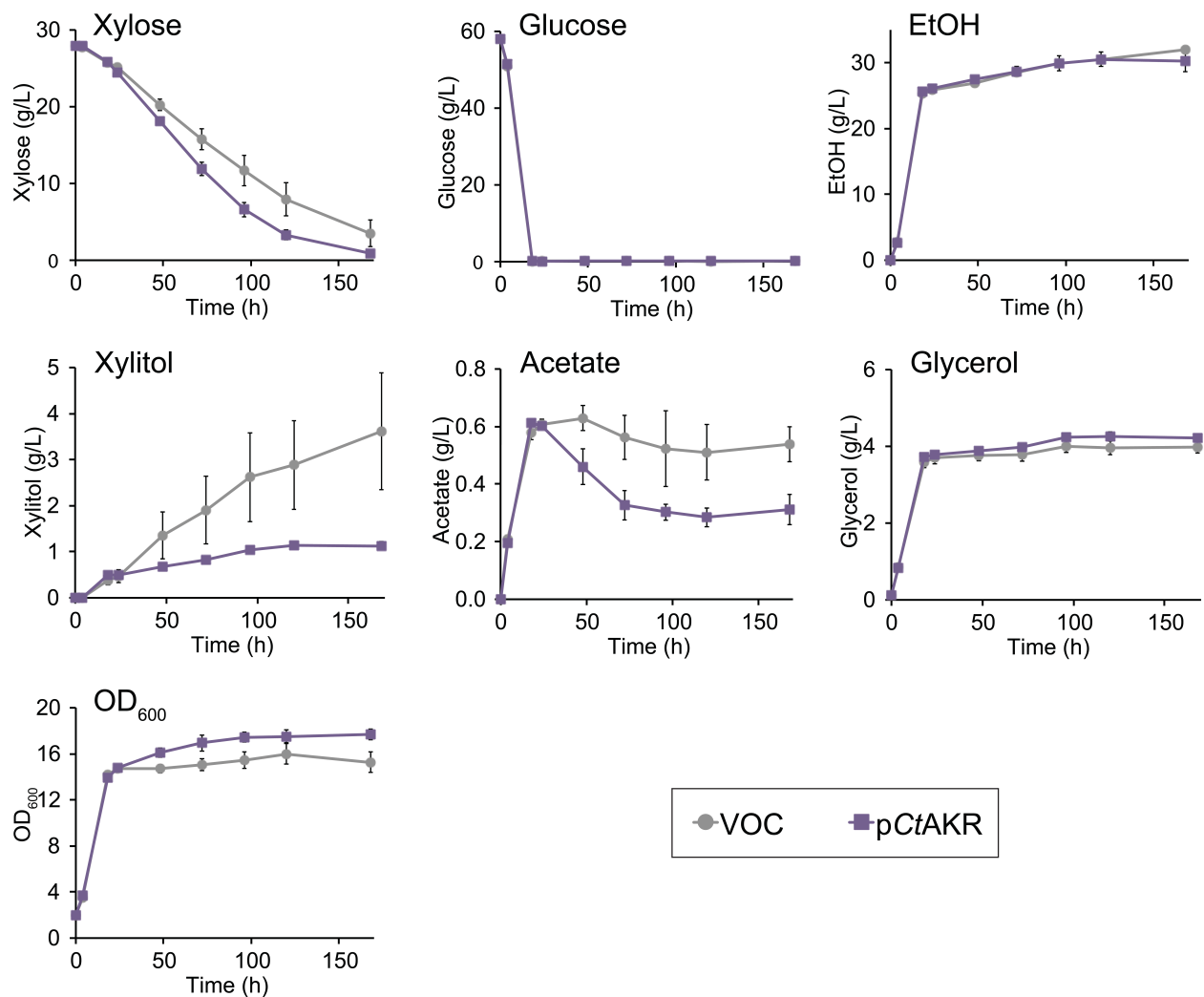
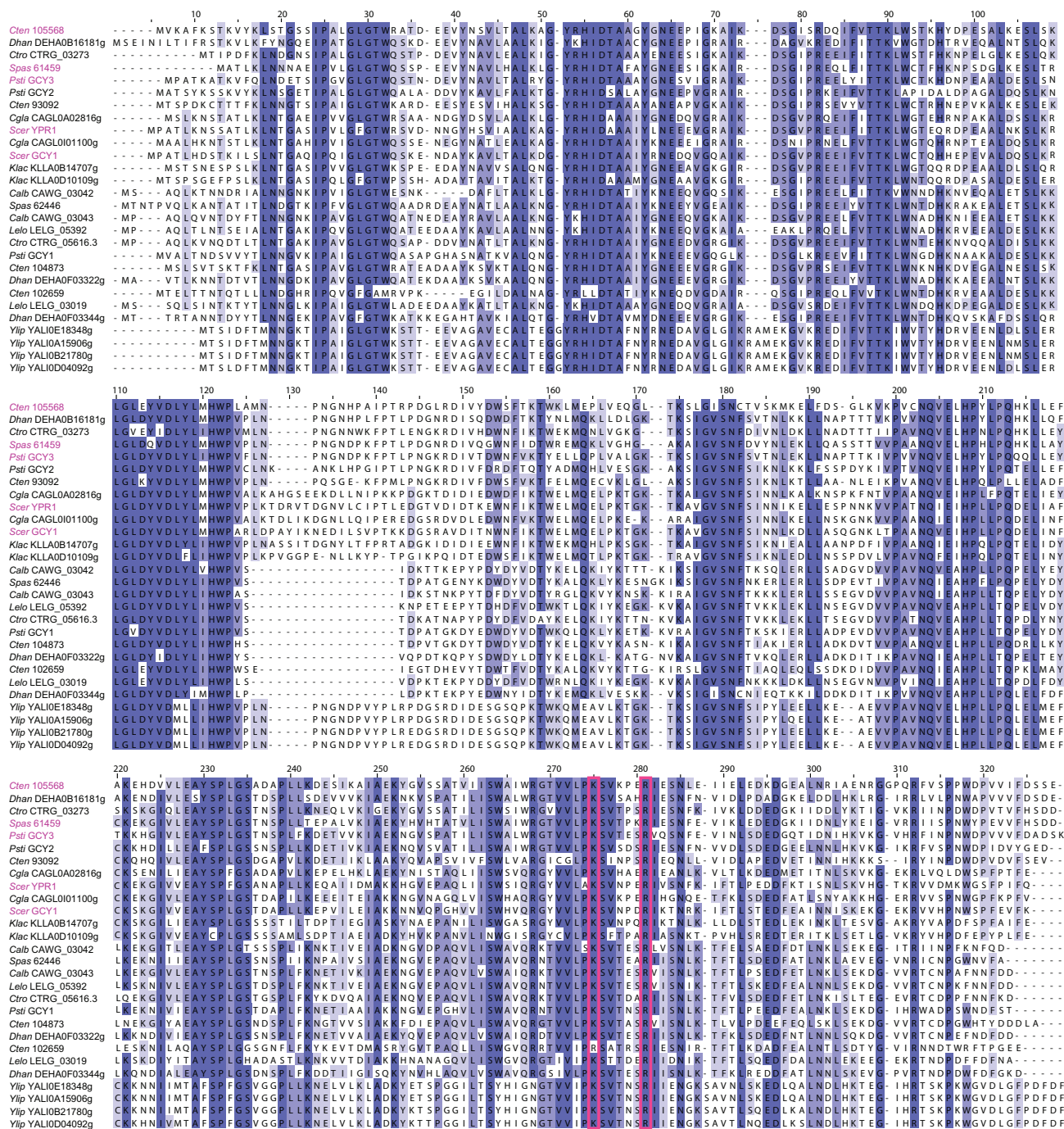


Fig. S11. *CtAKR* improves *Scer* xylose consumption. *Scer* strains GLBRCY0A+*PsXYL123*+pCtAKR (purple) and GLBRCY0A+*PsXYL123*+VOC (vector only control; grey) were grown anaerobically for 168 h. Average $\pm$ SD (n = 3) xylose, glucose, EtOH, xylitol, acetate, glycerol, and OD₆₀₀ are shown.

B



K R

Fig. S12B. Computational analysis of the fungal aldoketo reductases (AKRs). A protein alignment of the fungal AKRs was created using ClustalW (13). Residues are shaded blue according to conservation, with darker blue indicating more conservation. The lysine and arginine residues shown to confer NADP⁺ specificity are outlined in pink (27). Genes in pink text are those tested in *SI Appendix*, Fig. S13. Species abbreviations as in Table 1.

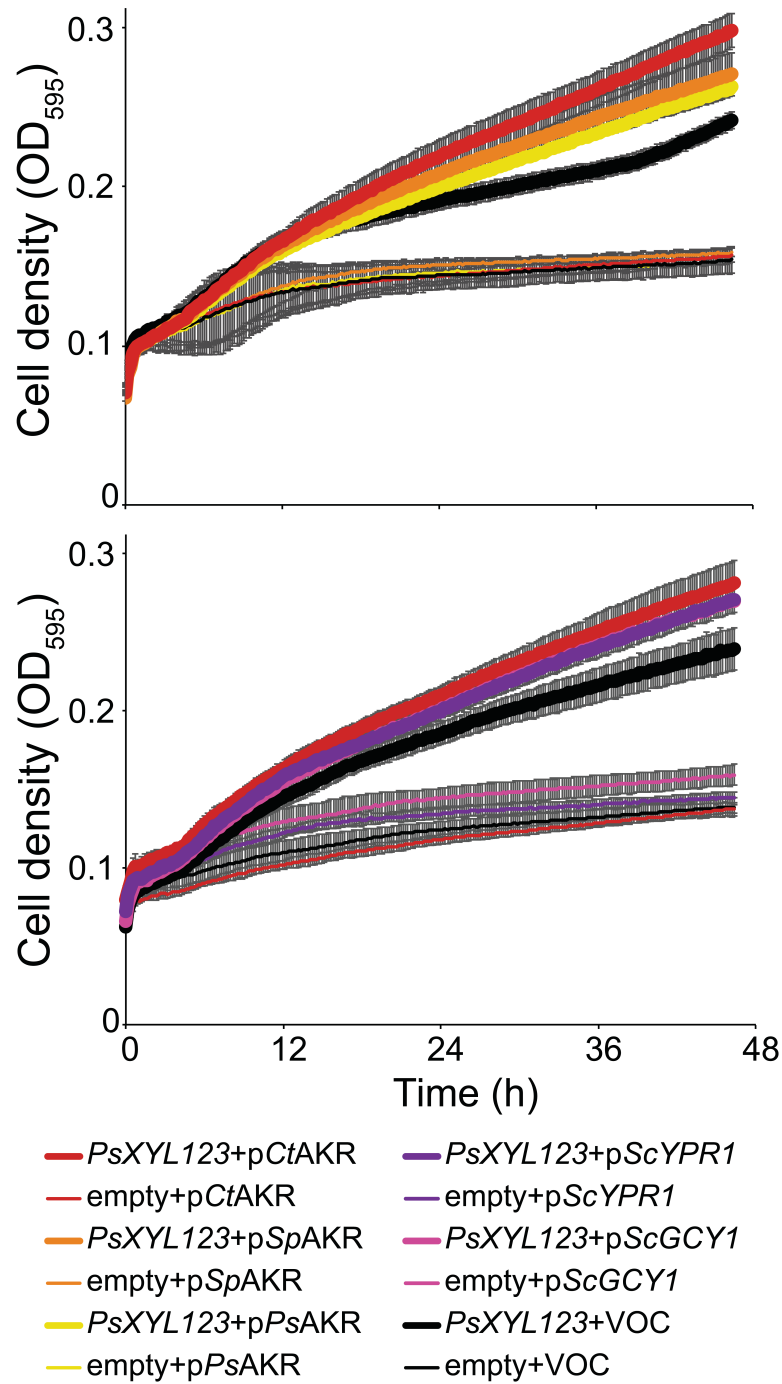


Fig S13. *CtAKR* orthologs improve xylose growth. Average $\pm$ SD ($n = 4$) growth on 8% xylose of GLBRCY0A harboring *PsXYL123* (thick lines) and GLBRCY0A lacking *PsXYL123* ('empty'; thin lines), and carrying p*CtAKR* (red), the *Spas* AKR ortholog (p*SpAKR*; orange), the *Psti* AKR ortholog (p*PsAKR*; yellow), *Scer YPR1* (p*ScYPR1*; purple), *Scer GCY1* (p*ScGCY1*; pink), or a vector-only control (VOC; black).

SI Tables

Table S1. Genome statistics for the xylose-fermenting fungi

	<i>Spas</i>	<i>Cten</i>	<i>Psti</i>
Strain	NRRL Y-27907	NRRL Y-1498	CBS 6054
Genome size (Mb)	13.1	10.6	15.4
Number of chromosomes	8	8	8
Total scaffolds	8	61	9
N_{50} scaffold length (Mb)	3	1.2	2.3
Percent GC	42.0	42.9	42.3
Coding genes	5983	5533	5841
Gene density (per Mb)	453.9	514.8	378.3
Avg. gene length (nt)	1786	1650	1627
Avg. transcript length (nt)	1720	1614	1568
Avg. protein length (aa)	451	447	493
Avg. exon length (nt)	1428	1332	1086
Avg. intron length (nt)	321	171	135
Number of genes with introns	994 (17%)	974 (18%)	1637 (28%)

N_{50} represents the scaffold size N at or above which 50% of all nucleotides are contained.

Table S2. Genome statistics for sequenced Hemiascomycete species

Organism	Strain	Genome Size (Mb)	Total ORFs	Data Source	Reference
<i>Ashbya gossypii</i>	ATCC 10895	8.7	4717	NCBI	(28)
<i>Candida albicans</i>	WO-1	14.4	6157	Broad Institute	(29)
<i>Candida dubliniensis</i>	CD36	14.6	5758	Wellcome Trust	(30)
<i>Candida glabrata</i>	CBS 138	12.3	5215	Genolevures	(31)
<i>Candida lusitanae</i>	ATCC 42720	12.1	5936	Broad Institute	(32)
<i>Candida parapsilosis</i>	CDC317	13.1	5733	Wellcome Trust	(32)
<i>Candida tenuis</i>	NRRL Y-1498	10.7	5533	DOE JGI	this work
<i>Candida tropicalis</i>	MYA-3404	14.6	6258	Broad Institute	(32)
<i>Debaryomyces hansenii</i>	CBS767	12.2	6887	Genolevures	(31)
<i>Kluyveromyces lactis</i>	NRRL Y-1140	10.7	5327	Genolevures	(31)
<i>Kluyveromyces waltii</i>	NCYC 2644	10.6	5214	(33)	(33)
<i>Lodderomyces elongisporus</i>	NRRL YB-4239	15.5	5796	Broad Institute	(32)
<i>Pichia guilliermondii</i>	ATCC 6260	10.6	5920	Broad Institute	(32)
<i>Pichia pastoris</i>	GS115	9.4	5313	NCBI	(34)
<i>Pichia stipitis</i>	CBS 6054	15.4	5841	DOE JGI	(35)
<i>Saccharomyces bayanus</i>	MCYC 623	11.5	4492	SGD	(36)
<i>Saccharomyces cerevisiae</i>	S288c	12.1	5695	SGD	(37)
<i>Saccharomyces mikatae</i>	IFO 1815	12.1	4525	SGD	(36)
<i>Saccharomyces paradoxus</i>	NRRL Y-17217	11.8	4788	SGD	(36)
<i>Spathaspora passalidarum</i>	NRRL Y-27907	13.2	5983	DOE JGI	this work
<i>Yarrowia lipolytica</i>	CLIB122	20.5	6436	Genolevures	(31)

Comparison of genome size and number of ORFs for all sequenced Hemiascomycete yeasts. The xylose-fermenting fungi are highlighted in bold text. DOE JGI, Department of Energy Joint Genome Institute; SGD, Saccharomyces Genome Database.

Table S3. Summary of orthologous groups of gene (OGG) statistics

Type of OGG	Number of OGGs	Number of Genes
<u>Multi-species OGGs</u>	5749 (47.8%)	74633 (91.1%)
High-confidence	5601	65916
Unresolved	148	8648
<u>Single-species OGGs</u>	6289 (52.2%)	7274 (8.9%)
Expansions	381	1366
Orphans	5908	5908
Total dataset	12038	81907

For more details on the types of OGGs, see *SI Appendix, Methods*. Unresolved OGGs contain genes for which there is not sufficient phylogenetic information within amino acid sequences to determine if the genes are derived from a single ancestral gene, or if there are multiple ancestral gene signatures in the OGG. Expansion OGGs are a group of paralogous genes from a single species. Orphan OGGs are single genes with no recognizable homolog in our data set. Numbers in parentheses represent the proportion of total OGGs or genes in that category.

Table S4. Top 50 Pfam domain gene families in the xylose-fermenting species

Pfam Domain	<i>Spas</i>	<i>Cten</i>	<i>Psti</i>	Description
PF07690.7	92	130	145	Major Facilitator Superfamily (MFS)
PF00069.16	83	85	90	Protein kinase domain
PF00400.23	79	67	71	WD domain, G-beta repeat
PF00271.22	59	49	56	Helicase conserved C-terminal domain
PF00172.9	58	85	86	Fungal Zn(2)-Cys(6) binuclear cluster domain
PF00076.13	43	37	41	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00083.15	38	49	53	Sugar (and other) transporter
PF00106.16	35	45	32	Short chain dehydrogenase
PF00096.17	34	37	37	Zinc finger, C2H2 type
PF00270.20	34	30	34	DEAD/DEAH box helicase
PF00153.18	33	32	34	Mitochondrial carrier protein
PF00005.18	31	24	30	ABC transporter
PF00560.24	31	16	30	Leucine Rich Repeat (LRR)
PF04082.9	30	33	48	Fungal specific transcription factor domain
PF01370.12	30	17	20	NAD-dependent epimerase/dehydratase family
PF00324.12	29	30	36	Amino acid permease
PF00004.20	29	30	31	ATPase family associated with various cellular activities
PF00646.24	29	14	14	F-box domain
PF08477.4	28	30	33	Miro-like protein
PF02985.13	28	27	30	HEAT repeat
PF00071.13	27	25	28	Ras family
PF00097.16	27	25	25	Zinc finger, C3HC4 type (RING finger)
PF01073.10	26	13	16	3-beta hydroxysteroid dehydrogenase/isomerase family
PF07993.3	26	12	14	Male sterility protein
PF05792.4	24	1	6	<i>Candida</i> agglutinin-like protein (ALS)
PF00226.22	23	22	24	DnaJ domain
PF00018.19	21	19	22	SH3 domain
PF08241.3	19	16	15	Methyltransferase domain
PF07719.8	19	13	15	Tetratricopeptide repeat
PF00702.17	18	21	19	Haloacid dehalogenase-like hydrolase
PF08240.3	18	19	23	Alcohol dehydrogenase GroES-like domain
PF00248.12	18	17	17	Aldo/keto reductase family
PF08242.3	18	13	16	Methyltransferase domain
PF00515.19	18	11	14	Tetratricopeptide repeat
PF00149.19	17	19	18	Calcineurin-like phosphoesterase
PF00176.14	17	18	18	SNF2 family N-terminal domain
PF00561.11	17	16	22	Alpha/beta hydrolase fold
PF07728.5	17	14	15	ATPase family associated with various cellular activities
PF01794.10	17	8	10	Ferric reductase like transmembrane component
PF07653.8	16	14	16	Variant SH3 domain
PF00023.21	16	13	14	Ankyrin repeat
PF08030.3	16	7	9	Ferric reductase NAD binding domain
PF00107.17	15	17	21	Zinc-binding dehydrogenase
PF00300.13	15	15	12	Phosphoglycerate mutase family
PF01423.13	15	14	12	LSM domain
PF08022.3	15	5	8	FAD-binding domain
PF01266.15	14	19	15	FAD-dependent oxidoreductase
PF00227.17	14	14	14	Proteasome A-type and B-type
PF00443.20	14	14	14	Ubiquitin carboxyl-terminal hydrolase

For each Pfam domain gene family, the total number of genes in each species is shown.

Table S5. Gene families with ≥ 10 members that are expanded ≥ 3 -fold in one of the xylose-fermenting species

ClusterID	<i>Spas</i>	<i>Cten</i>	<i>Psti</i>	Predominant Pfam domain description
11	24	1	3	<i>Candida</i> agglutinin-like (ALS)
22	1	0	18	None
23	1	14	4	Major Facilitator Superfamily
32	3	3	10	Major Facilitator Superfamily
42	1	2	11	Sugar (and other) transporter
47	12	1	1	Glycosyltransferase sugar-binding region containing DXD motif
161	0	0	12	None
62	10	1	1	Leucine Rich Repeat
81	11	0	0	None
83	10	0	0	None
86	0	0	10	Leucine Rich Repeat

ClusterID refers to the cluster number as found on JGI web portal (<http://www.jgi.doe.gov/>). For each clusterID, the total number of genes in each species is shown.

Table S6. Summary of functional enrichment of clusters of species-specific OGGs

Description	Number of OGGs in Cluster	Significant Annotation	<i>p</i> -value	Fold Enrichment
1 Unique to <i>Spom</i> and <i>Ylip</i>	114	No significant enrichment		
2 Unique to <i>Scer</i> , <i>Cgla</i> , and <i>Klac</i>	341	Meiosis	3.379e-8	3.5X
		M phase	7.234e-13	3.0X
		Cell cycle phase	7.308e-13	2.8X
		Unclassified	4.271e-21	1.7X
3 Unique to CUG yeasts	247	<i>de novo</i> NAD biosynthetic process	0.00891	22.3X
		Lipase activity	1.306e-6	9.2X
		Extracellular region	6.665e-6	6.2X
		Unclassified	4.274e-21	1.5X
4 Absent in <i>Spom</i> only	363	α -1,3-mannosyltransferase activity	0.000249	12.0X
		Lipid/fatty acid catabolic process	0.000438	4.4X
		Peroxisome	0.00014	3.2X
		Unclassified	0.001	1.3X
5 Absent in <i>Spom</i> and <i>Ylip</i>	150	No significant enrichment		

Cluster numbers correspond to those in Fig. 2A. *p*-values are Bonferroni-corrected, calculated from the hypergeometric distribution.

Table S7. Number of significantly differentially expressed genes in each species

	Number of Significant Genes		
	Induced	Repressed	Total
<i>P. stipitis</i>	170	219	389
<i>Sp. passalidarum</i>	198	143	341
<i>C. tenuis</i>	427	508	935
<i>C. albicans</i>	499	554	1053
<i>L. elongisporus</i>	952	869	1821

Significance determined with Limma (38) by paired t-tests within each species; FDR < 0.05.

Table S8. Fourteen significantly differentially expressed genes common to all three xylose fermenters

Annotation	<i>Psti</i>	<i>Spas</i>	<i>Cten</i>	<i>Calb</i>	<i>Lelo</i>
<i>EGC2</i> endo-1,4-beta-glucanase (cellulase)	6.77	6.47	1.04		
<i>BGL7</i> beta-glucosidase	2.38	0.46	0.74	0.30	
<i>BGL5</i> beta-glucosidase	0.86	1.19	2.17	0.30	
beta-glucosidase family 3	0.77	1.19	0.74	0.30	
vacuolar transporter chaperone 1	0.58	0.38	0.78	0.24	0.00
transcription regulatory protein	-0.59	0.24	-0.59	0.56	-0.09
protein kinase	0.32	-0.16	-0.90	0.42	-0.20
chromatin remodeling protein	0.55	-0.54	-0.41	-0.33	0.04
<i>XYL1</i> NAD(P)H-dependent D-xylose reductase	7.38	4.92	3.97	5.22	0.61
<i>RGT2</i> high-affinity glucose transporter	3.73	3.20	3.69	5.33	0.33
<i>XYL3</i> D-xylulokinase	3.59	4.41	1.22	3.16	0.88
<i>GAL10</i> UDP glucose-4-epimerase	3.09	3.13	2.29	1.63	2.52
<i>XYL2</i> xylitol dehydrogenase	4.97	6.80	3.89	4.47	5.21
oxidoreductase	1.67	2.16	0.65	3.75	2.30

Values given are log₂ fold-change of xylose versus glucose expression. Red text indicates statistically significant measurement (Limma t-test (38); FDR < 0.05). Blank cell indicates no ortholog present.

Table S9. *Lelo*-specific clusters are enriched for *Scer* stress response genes

Cluster	Num. <i>Lelo</i> Genes in Cluster	<i>Scer</i> class	Enrichment	<i>p</i> -value
Induced in <i>Lelo</i>	1137	Induced Stress Response	2.9X	1.34e-30
Repressed in <i>Lelo</i>	1168	Repressed Stress Response	4.0X	2.98e-168

Bonferroni-corrected *p*-values of enrichment (hypergeometric distribution) are given for each class of environmental stress response genes (39).

Table S10. Summary of functional enrichment of *Cten-Calb-Lelo* expression cluster

GO Term	Frequency in Cluster	Frequency in Genome	Enrichment	p-value
Fatty acid metabolic process	9/88	38/6848	17X	1.39e-7
Carboxylic acid metabolic process	17/88	250/6848	5.2X	2.35e-6
Lipid catabolic process	6/88	19/6848	22.7X	1.64e-5

Calb GO terms were used to identify functional enrichment of genes in the cluster using GO Term Finder (40). Bonferroni-corrected *p*-values of enrichment (hypergeometric distribution) are given for each GO term.

Table S11. Statistics for species-specific custom tiled microarrays

Species	Total Probes	Mean $\pm$ SD Probe Length (nt)	Mean $\pm$ SD Probe T _m (°C)	Median Probe Spacing (nt)
<i>P. stipitis</i>	374100	53.6 $\pm$ 4.1	76.3 $\pm$ 2.1	33
<i>Sp. passalidarum</i>	362487	54.5 $\pm$ 4.0	75.2 $\pm$ 2.6	29
<i>C. tenuis</i>	363196	53.1 $\pm$ 3.8	76.8 $\pm$ 2.2	24
<i>C. albicans</i>	373067	55.2 $\pm$ 4.0	73.7 $\pm$ 3.0	31
<i>L. elongisporus</i>	371451	54.1 $\pm$ 4.1	75.0 $\pm$ 3.0	33

Median probe spacing is determined by measuring the distance between 5' ends of adjacent probes, which are located on opposite strands.

Table S12. Genes predicted by automated annotation, classified by method

Method	<i>Spas</i>	<i>Cten</i>
<i>ab initio</i>	919 (15%)	1185 (21%)
Seeded by proteins in NR	2258 (38%)	2984 (54%)
Seeded by EST isotig	2806 (47%)	1364 (25%)
Total Models	5983 (100%)	5533 (100%)

NR, NCBI non-redundant protein set

Table S13. Quality of and supporting evidence for genes

Number of gene models	<i>Spas</i>	<i>Cten</i>	<i>Psti</i>
with start and stop codons	5524 (92%)	5358 (97%)	4991 (86%)
with EST support	5832 (97%)	5485 (99%)	ND
with NR support	5715 (96%)	5283 (95%)	ND
with Swiss-Prot support	5297 (89%)	4914 (89%)	5156 (88%)
with Pfam domain	4075 (68%)	3921 (71%)	3645 (62%)
with transmembrane domain	1124 (19%)	1063 (19%)	1161 (20%)
in multi-gene family	2921 (49%)	2542 (46%)	2880 (49%)
Total Models	5983 (100%)	5533 (100%)	5841 (100%)

NR, NCBI non-redundant protein set; ND, no data

Table S14. Functional annotation of proteins

Number of proteins assigned	<i>Spas</i>	<i>Cten</i>	<i>Psti</i>
to a KOG	4376 (73%)	3989 (72%)	4417 (76%)
a GO term	3685 (62%)	3465 (63%)	3477 (60%)
an EC number	1823 (31%)	1572 (28%)	1705 (29%)

Numbers in parentheses indicate percentage of total proteins from that species.