Comparative genomics of xylose-fermenting fungi for enhanced biofuel production

Dana J. Wohlbach[a,b], Alan Kuo[c], Trey K. Sato[b], Katlyn M. Potts[a], Asaf A. Salamov[c], Kurt M. LaButti[c], Hui Sun[c], Alicia Clum[c], Jasmyn L. Pangilinan[c], Erika A. Lindquist[c], Susan Lucas[c], Alla Lapidus[c], Mingjie Jin[d,e], Christa Gunawan[d,e], Venkatesh Balan[d,e], Bruce E. Dale[d,e], Thomas W. Jeffries[b], Robert Zinkel[b], Kerrie W. Barry[c], Igor V. Grigoriev[c], and Audrey P. Gasch[a,b,1]

[a]Department of Genetics, University of Wisconsin, Madison, WI 53706; [b]Great Lakes Bioenergy Research Center, Madison, WI 53706; [c]US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598; [d]Biomass Conversion Research Laboratory, Department of Chemical Engineering and Materials Science, Michigan State University, Lansing, MI 48910; and [e]Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI 48824

Edited by Chris R. Somerville, University of California, Berkeley, CA, and approved July 5, 2011 (received for review February 24, 2011)

Cellulosic biomass is an abundant and underused substrate for biofuel production. The inability of many microbes to metabolize the pentose sugars abundant within hemicellulose creates specific challenges for microbial biofuel production from cellulosic material. Although engineered strains of *Saccharomyces cerevisiae* can use the pentose xylose, the fermentative capacity pales in comparison with 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 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.

bioenergy | genome sequencing | transcriptomics

Biofuel production from cellulosic material uses 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. Native *Saccharomyces cerevisiae* (*Scer*) does not consume xylose but can be engineered for xylose consumption with a minimal set of assimilation enzymes, including xylose reductase (Xyl1) and xylitol dehydrogenase (Xyl2) from the xylose-fermenting *Pichia stipitis* (*Psti*) (Fig. 1*A*) (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. 1*A*) (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 and also reflected 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 with the existing *Psti* genome (*Materials and Methods* and *SI Appendix, Materials and Methods* and Table S1) (13). The *Spas* genome was sequenced to 43.77× coverage over 13.1 Mb arranged in eight scaffolds. The *Cten* genome was sequenced to 26.9× coverage, generating 10.7 Mb in 61 scaffolds representing eight chromosomes. Compared with 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 with *Spas*, *Cten*, and *Psti* (Fig. 1 *B* and *C*). Whole-genome phylogenetic analysis placed both *Spas* and *Cten* within the CUG clade of yeasts (Fig. 1*B* and *SI Appendix, Materials and Methods*) 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

Author contributions: D.J.W. and A.P.G. designed research; D.J.W., A.K., T.K.S., K.M.P., A.A.S., K.M.L., H.S., A.C., J.L.P., E.A.L., S.L., A.L., M.J., C.G., V.B., and R.Z. performed research; D.J.W., A.K., A.A.S., M.J., B.E.D., T.W.J., K.W.B., and I.V.G. analyzed data; and D.J.W., and A.P.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The assemblies and annotations reported in this paper have been deposited in the GenBank database [accession nos. AEIK00000000 (*Spathaspora passalidarum*) and AEIM00000000 (*Candida tenuis*)]. Microarray data have been deposited in the Gene Expression database (accession no. GSE24858).

[1]To whom correspondence should be addressed. E-mail: agasch@wisc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1103039108/-/DCSupplemental.

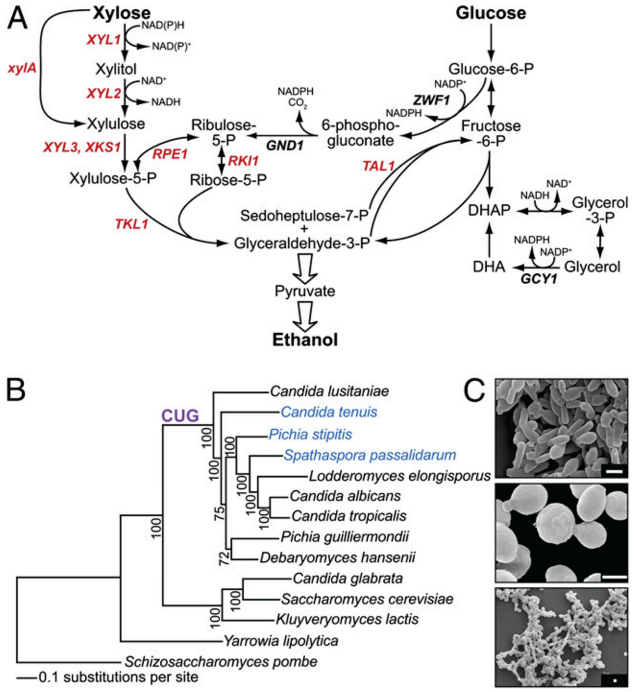

Fig. 1. Overview of xylose assimilation and phylogeny of xylose-fermenting fungi. (*A*) The simplified pathway includes genes that have been engineered in *Scer* via overexpression (red text) for improved xylose fermentation. *GND1*, 6-phosphogluconate dehydrogenase; *RKI1*, ribose-5-phosphate ketol-isomerase; *RPE1*, ribulose-5-phosphate 3-epimerase; *XKS1*/*XYL3*, xylulokinase; *xylA*, xylose isomerase. (*B*) Maximum likelihood phylogeny from concatenated alignment of 136 universal orthologs, with bootstrap values. (*C*) Electron microscopy images of *Cten* (*Top*), *Psti* (*Middle*), and *Spas* (*Bottom*). (Scale bars: 2 μm.)

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. S2 *A* and *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. S2*C*). 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 in the 14 Ascomycetes in our phylogeny by assigning orthology and paralogy relationships among the metaset of 81,907 predicted fungal protein-coding genes (*SI Appendix*, *Materials and Methods*). More than 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. S3*A*). In contrast, the other OGGs (52% of all OGGs representing 9% of all genes) are species-specific paralogs that are distributed nonrandomly throughout the phylogeny (*SI Appendix*, Fig S3*B*). Within the CUG clade, *Debaryomyces hansenii* and *Pichia guilliermondii* have the most single-species expansions, and 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, possibly related to their unique sugar environment (*SI Appendix*, Tables S4 and S5).

We analyzed conservation patterns of the 5,749 multispecies OGGs through a clustering approach, which identified clade-specific OGGs enriched for different functional properties (Fig. 2*A* and *SI Appendix*, Table S6). Approximately half of the multispecies 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 for 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 interpreted previously as being important for pathogenicity (19), their presence in beetle symbionts suggests they may be relevant

Table 1. **Strain sources and genome statistics**

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

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

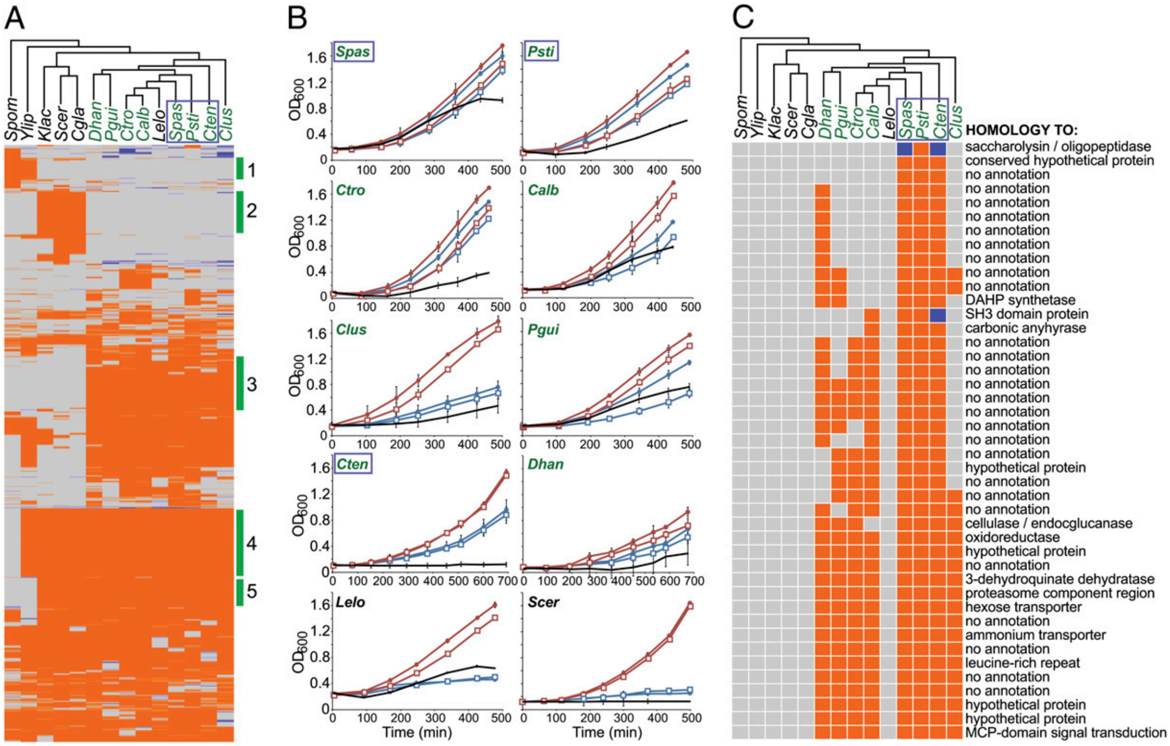

Fig. 2. Mapping of phenotype and genotype onto phylogeny. (*A*) Hierarchical clustering based on ortholog presence (orange) or absence (gray) for 3,073 nonubiquitous multispecies OGGs. Blue indicates BLAST homology despite no ortholog call. Functional enrichment in indicated clusters is described in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

to commensalism rather than to pathogenicity per se. Additionally, many genes unique to CUG yeasts are involved in de novo NAD^+ biosynthetic processes ($p = 0.00891$), suggesting 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 use xylose (6). This group includes orthologs of *Psti XYL1* (11), *XYL2* (12), and xylulokinase (*XYL3*) (10), the minimal set of genes required to engineer *Scer* for xylose assimilation (Fig. 1*A*) (2, 3, 5). However, these genes show no evolutionary signatures of selection or constraint to suggest functional modification in the xylose-using species (*SI Appendix*, Fig. S4). Thus, factors other than the mere presence of this minimal gene set must contribute to phenotypic differences in xylose consumption.

Conservation of Orthologous Gene Groups Points to 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. 2*B* and *SI Appendix*, Figs. S5 and S6). *Psti*, *Spas*, and *Cten* were the only species able to ferment xylose measurably in our assay (*SI Appendix*, Fig. S6). Intriguingly, these species also are the yeasts associated with beetles, many of which are attracted to fermentation byproducts (20). Only three genes are found uniquely 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). Although 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. 2*B*), 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 nongrowers but common to all xylose fermenters, with varying conservation across species that could assimilate xylose (Fig. 2*C* and Dataset S1). Fifteen showed presence and absence patterns that correlated strictly with xylose assimilation. These genes 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; 10 also are 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 potentially could be useful for biomass degradation (see *SI Appendix*, Fig. S7 for protein domain and signal peptide analysis). Although the conservation of these genes suggests 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. 3*A*). In particular, *Lelo* altered the expression of thousands of genes, including orthologs of the yeast environmental stress response that are induced when *Scer* is stressed (23) or experiences xylose (*SI Appendix*, Fig. S8*A* and Table S9) (24). This massive expression pattern in *Lelo* likely represents a starvation response to carbon limitation and demonstrates that the environmental stress response 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. S8*B* 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. 3 *B* and *C*). These clusters include genes whose expression is required for optimal xylose utilization in engineered *Scer* [e.g., *XYL1*, *XYL2*, *XYL3*, transketolase (*TKL1*), and transaldolase (*TAL1*)] (Fig. 1*A*). Strikingly, several of these genes were strongly induced in *Lelo*, even though it cannot use 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 in xylose growers specifically. Genes encoding β-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 up-regulated also. 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 [glucose-6-phosphate dehydrogenase (*ZWF1*) and phosphoglucose isomerase (*PGI1*)] 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.

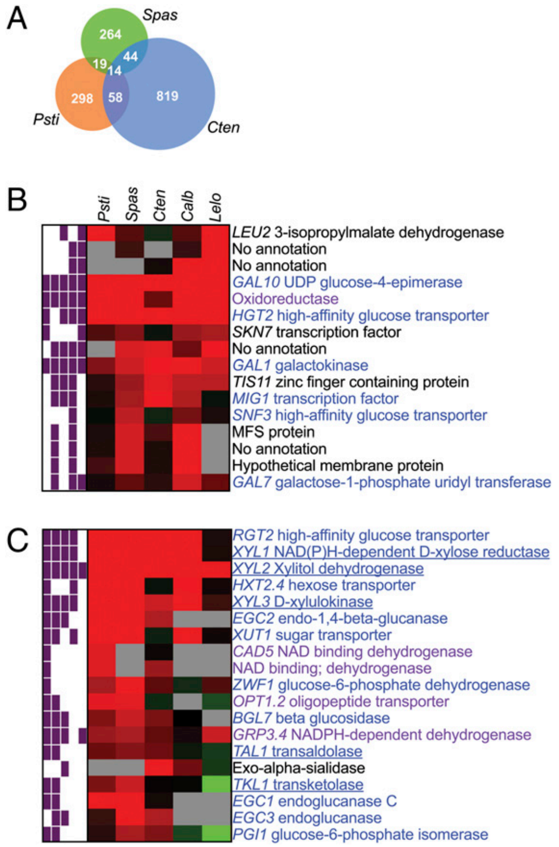

Fig. 3. Transcriptome analysis of xylose-growing cultures. (*A*) Overlap between significantly differentially expressed genes within the xylose fermenters (FDR <0.05). (*B* and *C*) Two clusters of genes, identified by hierarchical clustering, that are highly induced in most species responding to xylose.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

Candidate Genes Improve Xylose Utilization. We tested 10 of the implicated genes 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, *Sp*HXT, and a glucose-6-phosphate dehydrogenase, *Sp*GPD. Importantly, two genes had a specific positive effect on xylose utilization in one or both strain backgrounds: a *Cten* aldo/keto reductase, *Ct*AKR, and a *Spas* unannotated protein, *Sp*NA, with homology to uncharacterized fungal-specific proteins (Fig. 4 and *SI Appendix*, Fig. S10).

Expression of plasmid-born *Ct*AKR 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 con-

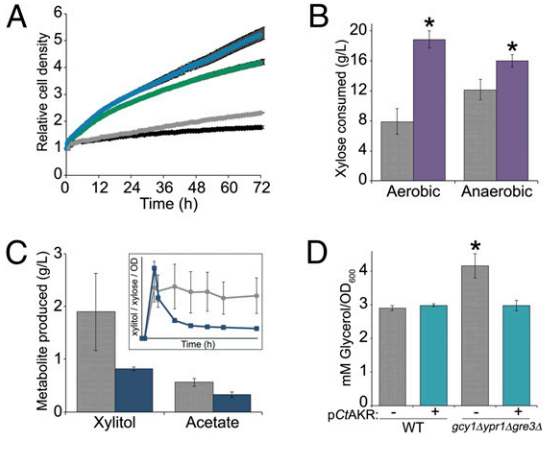

Fig. 4. *Ct*AKR improves *Scer* xylose utilization. (*A*) Average ± SD (*n* = 4) growth on 8% xylose of *Scer* strain GLBRCY0A carrying *PsXYL123*+p*Ct*AKR (blue), *PsXYL123*+VOC (vector only control; green), p*Ct*AKR only (gray), or VOC only (black). (*B*) Average ± SD (*n* = 3) xylose consumed after 72 h growth for GLBRCY0A carrying *PsXYL123*+p*Ct*AKR (purple) or *PsXYL123*+VOC (gray). Asterisks indicate statistically significant measurements ($p < 0.05$, *t* test). (*C*) Average ± SD (*n* = 3) xylitol or acetate produced after 72 h anaerobic fermentation for GLBRCY0A carrying *PsXYL123*+p*Ct*AKR (blue) or *PsXYL123*+VOC (gray). (*Inset*) Time course of average ± SD (*n* = 3) anaerobic xylitol production relative to xylose consumed. (*D*) Average ± SD (*n* = 3) glycerol produced in wild-type (BY4741) or mutant strains carrying p*Ct*AKR (aqua) or VOC (gray).

ditions (25, 26), was not increased significantly (*SI Appendix*, Fig. S11). However, acetate production was reduced 42% (Fig. 4*C*). 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 limited by factors other than carbon availability. Nonetheless, the significant effect of p*Ct*AKR on anaerobic xylose assimilation and concomitant reduction in xylitol represents a major advance in cellulosic biomass conversion by *Scer.*

*Ct*AKR is a member of the large protein family that includes xylose reductases (*SI Appendix*, Fig. S12*A*). However, *Ct*AKR is most similar to the NADP$^+$-dependent glycerol dehydrogenase Gcy1 from *Scer*, which functions in an alternative pathway for glycerol catabolism (Fig. 1*A*) (27). Notably, *Ct*AKR contains residues known to establish NADP$^+$ binding (*SI Appendix*, Fig. S12*B*) (reviewed in ref. 28), suggesting *Ct*AKR also may function in a NADP$^+$-specific manner. We examined the effect of p*Ct*AKR expression on glycerol metabolism in a *Scer* mutant lacking three functionally redundant aldo/keto reductases (*GCY1*, *YPR1*, and *GRE3*) (*Materials and Methods*). Glycerol levels increased in the mutant strain but were restored to wild-type levels by p*Ct*AKR (Fig. 4*D*). Together, these data suggest that *Ct*AKR functions as a NADP$^+$-dependent glycerol dehydrogenase in *Scer.* Indeed, like *Ct*AKR, overexpression of *Scer GCY1* or *YPR1* had a positive effect on xylose utilization (*SI Appendix*, Fig. S13), further supporting our hypothesis concerning *Ct*AKR function.

Conclusions

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 used 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 but 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 largely have 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 *Ct*AKR 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 change significantly in strains engineered with p*Ct*AKR 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 alleviate cofactor imbalance. Decreased acetate levels also may result from increased glycerol cycling, because otherwise acetate is generated as a fermentation byproduct to alleviate cofactor imbalance (4). Although 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 (40-kb fosmid library) and 454 (standard and paired-ended libraries) sequencing platforms. Newbler (v. 2.3; Roche) 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). mRNA was purified using the Absolutely mRNA purification kit (Stratagene) and was reverse transcribed with SuperScriptIII using $dT_{15}VN_2$ primer. cDNA was synthesized with *Escherichia 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 Methods*) and can be accessed through the JGI Genome Portal (http://www.jgi.doe.gov/spathaspora/ and http://www.jgi.doe.gov/tenuis/). *SI Appendix*, Tables S1, S12, S13, and S14 list genome-sequencing statistics.

Species Phylogeny and Orthology. We estimated the phylogeny using protein sequences of 136 single-copy orthologs present in all species (*SI Appendix, Materials and Methods*). Phylogenies were constructed with MrBayes v. 3.1.2 (34, 35). We created OGGs using a modified reciprocal smallest distance (RSD) (36) and OrthoMCL (37) method with the RSD parameters significance threshold, 10^{-5}; alignment threshold, 0.3; and the 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*, or *Schizosaccharomyces pombe*. Results from the two methods were compared and combined using a custom Practical Extraction and Reporting Language (perl) script to maximize high-confidence assignments (true positives) and minimize low-confidence assignments (false positives) (*SI Appendix, Materials and Methods*).

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 (Genscript) containing the *Scer PGK1* promoter (Pr), *TDH3* terminator (t), Pr*TDH3*, t*TEF2*, Pr*TEF2*, and t, followed by the kanamycin (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-μm pRS426 vector (39) modified with an 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 hygromycin (HygMX) drug-resistance cassettes. All constructs were confirmed by diagnostic PCR and/or DNA sequencing.

Phenotypic Assays. For all assays, cultures were grown in 1% yeast extract, 2% peptone, 2% glucose or in 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 to either liquid or solid medium containing 2–10% glucose or xylose. For solid growth assays, plates were scored after 2 d at 30 °C. For liquid growth assays, OD_{600} 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 before analysis by HPLC with a Bio-Rad Aminex HPX-87H column (40). Concentrations of ethanol also were determined using an Agilent Technologies 7890A gas chromatograph with a 7693 autosampler and flame ionization detector (*SI Appendix, Materials and Methods*).

Microarrays. Cells were collected at OD_{600} 0.5–0.6 after growth for three generations in 2% glucose or xylose. Cell lysis and total RNA isolation were performed as previously described (41). RNA was purified further with LiCl and the 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 $\log_2$ signal of RNA versus a 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 v. 2.9.8 (45) with a false-discovery rate (FDR) correction of 0.05 (46). For cross-species comparisons, genes within OGGs were evaluated for differences in expression. When an OGG contained more than one gene from a particular species, genes with the smallest phylogenetic distance [determined with PAML v. 4.3 (47)] were compared directly. Hierarchical clustering of gene expression across species was performed with Cluster 3.0 using the uncentered Pearson correlation as the distance metric (48).

ACKNOWLEDGMENTS. 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 (DOE) and was funded in part by the DOE Great Lakes Bioenergy Research Center (DOE Biological and Environmental Research Office of Science Grant DE-FC02-07ER64494), National Library of Medicine Training Grant 5T15LM007359 (to D.J.W.), and National Science Foundation Career Award 0447887 (to A.P.G.). The work conducted by the US DOE Joint Genome Institute is supported by the Office of Science of the US DOE under Contract No. DE-AC02-05CH11231.

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 Technol* 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. Scientific note. *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 Biotechnol* 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 USA* 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 USA* 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* 11:723–739.
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. (2008) POLISHER: An effective tool for using ultra short reads in microbial genome assembly and finishingEleventh Annual Meeting on Advances in Genome Biology and Technology, February 6–9, 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* 19:1572–1574.
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, Pöhlmann 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(Web Server issue, 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:Article 3.
46. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 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 USA* 95:14863–14868.
49. Jones T, et al. (2004) The diploid genome sequence of *Candida albicans*. *Proc Natl Acad Sci USA* 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* 274:546–567.
52. Wood V, et al. (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature* 415:871–880.