LBNL-62457

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Published Title: Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis

Working Title: Genomic sequence of the xylose fermenting, insect-inhabiting yeast, Pichia stipitis

Author(s) Thomas W. Jeffries, Igor Grigoriev, et al

Division Genomics

Journal Name Nature Biotech

Month Year March 1, 2007

Volume 25

Pages 319-326

Genomic sequence of the xylose fermenting, insect-inhabiting yeast, *Pichia stipitis*

Thomas W. Jeffries,[1]* Igor Grigoriev,[2] Jane Grimwood,[3] José M. Laplaza,[1,4] Andrea Aerts,[2] Asaf Salamov,[2] Jeremy Schmutz,[3] Erika Lindquist,[2] Paramvir Dehal,[2] Harris Shapiro,[2] Yong-Su Jin,[5] Volkmar Passoth,[6] and Paul M. Richardson[2]

[1]USDA, Forest Service, Forest Products Laboratory, One Gifford Pinchot Drive, Madison, WI 53705 and Department of Bacteriology, University of Wisconsin-Madison; [2]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, [3]Stanford Human Genome Center, 975 California Ave, Palo Alto, CA 94304, USA; [4]BioTechnology Development Center, Cargill, PO Box 5702, Minneapolis, MN 55440-5702 [5]Department of Food Science and Biotechnology Sungkyunkwan Univeristy, Suwon, Korea; [6]Swedish University of Agricultural Sciences (SLU), Dept. of Microbiology, Uppsala, Sweden,

*To whom correspondence should be addressed (twjeffri@wisc.edu)

Prepared for submission to: *Nature Biotechnology*

Keywords: yeast, genome, xylose, EST expression, glycoside hydrolase, transcription factories

Version of Thursday, June 28, 2007

ABSTRACT

Xylose is a major constituent of angiosperm lignocellulose, so its fermentation is important for bioconversion to fuels and chemicals. *Pichia stipitis* is the best-studied native xylose fermenting yeast. Genes from *P. stipitis* have been used to engineer xylose metabolism in *Saccharomyces cerevisiae*, and the regulation of the *P. stipitis* genome offers insights into the mechanisms of xylose metabolism in yeasts. We have sequenced, assembled and finished the genome of *P. stipitis*. As such, it is one of only a handful of completely finished eukaryotic organisms undergoing analysis and manual curation. The sequence has revealed aspects of genome organization, numerous genes for biocoversion, preliminary insights into regulation of central metabolic pathways, numerous examples of co-localized genes with related functions, and evidence of how *P. stipitis* manages to achieve redox balance while growing on xylose under microaerobic conditions.

INTRODUCTION

Xylose is a five-carbon sugar that makes up about 15 to 25% of all hardwoods and agricultural residues.[1] Its fermentation is therefore essential for the economic conversion of lignocellulose to ethanol.[2-4] *Pichia stipitis* Pignal (1967) is a predominantly haploid, homothallic, hemiascomycetous yeast[5-7] that has the highest native capacity for xylose fermentation of any known microbe.[8, 9] Fed batch cultures of *P. stipitis* produce up to 47 g/L of ethanol from xylose at 30°C[10] with ethanol yields of 0.35 to 0.44 g/g xylose (Fig. 1),[11] and they are capable of fermenting sugars from hemicellulosic acid hydrolysates with a yield equivalent to about 80% of the maximum theoretical conversion efficiency.[12]

P. stipitis Pignal (1967) was originally isolated from insect larvae. It is closely related to several yeast endosymbionts of passalid beetles[13] that inhabit and degrade white-rotted hardwood.[14, 15] It forms yeast-like buds during exponential growth, hat-shaped spores, and pseudomycelia (Fig. 2). The genomic sequence reveals numerous features such as cellulases, xylanase, and other degradative enzymes that would enable survival and growth in a wood-inhabiting, insect-gut environment.[13] *P. stipitis* has the capacity to grow on and ferment xylan[16, 17], and to use all of the major sugars found in wood. In addition, it has been reported to use low-molecular weight lignin moieties.[18]

P. stipitis has been a source of genes for engineering xylose metabolism in *Saccharomyces cerevisiae*.[19] Although metabolic engineering and adaptive evolution of *S. cerevisiae* for xylose

fermentation has been successful to varying degrees,[20-22] it does not possess the regulatory mechanisms that coordinate ethanol production with xylose.[23] Unlike *S. cerevisiae*, which regulates fermentation by sensing the presence of glucose, *P. stipitis* induces fermentative activity in response to oxygen limitation.[24-26] *P. stipitis* shunts most of its metabolic flux into ethanol, and produces very little xylitol, but its xylose fermentation rate is low relative to *S. cerevisiae* on glucose. Increasing the capacity of *P. stipitis* for rapid xylose fermentation could therefore greatly improve its usefulness in commercial xylose fermentations.

We have sequenced the *P. stipitis* genome to better understand the biology, metabolic machinery, and regulatory networks in this native xylose- fermenting yeast. The *P. stipitis* genome sequence, predicted genes, and annotations are available through the JGI Genome Portal at www.jgi.doe.gov/pichia. The results reveal a versatile lower eukaryote that has unusual genetic and regulatory features for converting lignocellulosic feedstocks into ethanol and other useful chemicals.

RESULTS

General genome features and comparative genomics

The 15.4 Mbp genome of *P. stipitis* genome was sequenced using a whole-genome shotgun approach and finished to high quality (< 1 error in 100,000). The JGI assembler, JAZZ[27] was used to assemble 261,986 reads into 96 scaffolds with 8.8x coverage and 4.4% gaps. The assembly was then finished, gaps were closed, and the scaffolds were linked into 8 chromosomes ranging from 3.5 to 0.97 Mbp, which is similar to results from pulsed field electrophoresis with various other strains of *P. stipitis*.[28] The finished chromosomes have no gaps except one in the centromere region of chromosome 1,

The JGI Annotation Pipeline predicted 5,841 genes. A majority (4,204, or 72%), have a single exon, which is typical for a yeast genome (Table 1). Average gene density, which is similar on all 8 chromosomes, is 56%. Average gene, transcript, and protein lengths are 1.6 kb, 1.5 kb and 493 amino acids, respectively. ESTs support 2,252 (40%) of the predicted genes, and an absolute majority is supported by protein homology; including 4,879 (84%) with strong homology in other fungi. Best bi-directional BLAST analysis of the gene models against the *D. hansenii* genome identified putative orthologs for 4,912 (84%) of the *P. stipitis* genes. These had an average identity of 58% at the amino acid level and average coverage of 91% in alignments between the orthologs. No data base match was found for 154 ORFs. Additionally, analysis of conservation between the genomes of *P. stipitis* and *D. hansenii* at the DNA level using VISTA

tools[29] provided support for exons in 3,940 (67.5%) of the *P. stipitis* genes. Approximately half (2,750) of the gene models had been manually curated at the time of publication.

Functional portrait

Protein function can be tentatively assigned to about 70% of the genes according to KOG (clusters of orthologous groups) classifications.[30] They are roughly equally split between 3 major categories: cellular processes and signaling, information storage and processing, and metabolism (Fig. 3). Protein domains were predicted in 4,083 (70%) of gene models. These include 1,712 distinct Pfam domains.

We used the PhIGs tool (Phylogenetically Inferred Groups,[30] http://phigs.org) to compare the gene set of *P. stipitis* with the gene sets of five other yeasts - *Saccharomyces cerevisiae, Candida glabrata, Kluyveromyces lactis, Debaromyces hansenii and Yarrowia lipolytica* - whose genomes have also been sequenced, assembled, and reported (Fig. 4).[31, 32] This analysis revealed 25 gene families representing 72 proteins that are specific to *P. stipitis* (Table 2). These show no significant homology to any known proteins; neither do they have any predicted domains. *P. stipitis* and *D. hansenii* share 151 gene families that are not found in the other 3 genomes used in this comparison. At the same time the *P. stipitis* gene set was missing 81 gene families relative to the other 5 yeast genomes in the analysis, which represents 442 individual proteins.

The most frequent domains in the *P. stipitis* genome include protein kinases, helicases, transporters (sugar and MFS), and domains involved in transcriptional regulation (fungal specific transcription factors, RNA recognition motifs and WD40 domains). A majority of domains are shared with other hemiascomycota. These range from 1,534 domains common with *S. pombe* to 1,639 with *D. hansenii*. One of the few *P. stipitis*-specific domains (Table S1) belongs to glycosyl hydrolase Family 10, a subgroup of cellulases and xylanases. The only Family 10 glycosyl hydrolase in the *P. stipitis* genome is *XYN1*. Among the domains consistently present in hemiascomycetous yeasts, more than twenty were not found in the *P. stipitis* genome including transposon-related domains removed from *P. stipitis* gene set by masking genomic sequence. These include the integrase core domain, rve, which integrates a DNA copy of a viral genome into the host chromosome,[33] RUT_2, which is indicative of a mobile element such as a retrotransposon,[34] and the HHH domain, which is found in non-sequence specific DNA binding proteins. Several gene families expanded in *P. stipitis* show some sequence similarity to hyphally regulated cell wall proteins, cell surface flocculins, agglutinin-like proteins, and

cytochrome p450 non-specific monooxygenases, Members of these expanded families, however, are poorly conserved and often occur near chromosome termini (within 35,000 bp) where repeated sequences are prevalent.

Syntenic relationships

Co-linearity between chromosomal blocks has been reported in plants, animals,[35] and closely related yeast genomes, e.g. *Saccharomyces sensu stricto*.[35, 36] Co-linearity is harder to find in a more diverse set of fungal genomes.[32] With the relatively recent divergence between *P. stipitis* and *D. hansenii*, chromosomal segments that retain the ancestral gene groupings can be identified. The set of 3,209 genes determined to be orthologous from the PhIGs analysis were used to link regions between the two genomes that represent orthologous chromosomal segments with a minimum of four linking genes that are uninterrupted by other orthology segments in either genome. A total of 263 orthology segments were found, encompassing 4456 (76.3%) genes and 10,950,900 bp in the *P. stipitis* genome, and 4689 (75.8%) genes and 9,057,788 bp in the *D. hansenii* genome. On average, each block in the *P. stipitis* genome encompasses 16.9 genes and is 41.6 kb in length. The largest of these orthologous chromosomal segments, 125 genes, which is 301.9 kb in length and encompasses 125 genes, is between *P. stipitis* chromosome 6 and *D. hansenii* chromosome F (Fig. 5).

Metabolic functions

Sugar transport: *P. stipitis* possesses genes for a number of transporters that are similar to putative xylose transporters from *Debaromyces hansenii* (NCBI AAR06925)[37] and *Candida intermedia* (*GXF1*, EMBL AJ937350; *GXS1*, EMBL AJ875406).[38] *C. intermedia GXF1* has the closest similarity to the previously described, closely related *SUT1, SUT2 and SUT3* genes of *P. stipitis* and to the *P. stipitis SUT4* gene that was identified in the present genome sequence (supplemental Fig. 1). Notably, *SUT2* and *SUT3* are each located very near one end of chromosomes 4 and 6, respectively, and our EST data has not shown that they are expressed.

Glycolytic and pentose phosphate pathways: All of the genes for xylose assimilation, the oxidative pentose phosphate pathway (PPP), glycolysis, the tricarboxylic acid cycle (TCA) and ethanol production were present in isoforms similar to those found in other yeasts (Fig. 6). The *XYL1, XYL2* and *XKS1* (*XYL3*) genes, which are required for xylose assimilation, were present in a single copy each. There are, however, several aldo/keto reductases homologous to *XYL1* (e.g. *GCY1-3*) and a family of sorbitol dehydrogenases with homology to *XYL2*.

Glucose 6-P-dehydrogenase (*ZWF1*), and 6-phosphogluconate dehydrogenase (*GND1*) generate NADPH necessary for cell growth and xylose assimilation by their roles in the oxidative phase of the PPP. Transcripts of the latter are strongly induced by growth on xylose under both aerobic and oxygen limiting conditions (Fig. 6). Transketolase (*TKT1*) is used twice in the non-oxidative phase of the PPP. It is strongly induced on xylose, and is one of the most abundant transcripts in the cell under those conditions. A gene for a second transketolase-like protein is present, but it is closer in structure to dihydroxyacetone synthase (*DHA1*) or formaldehyde transketolase.

P. stipitis has a gene for a bacterial-like ribose-5-phosphate isomerase B (*RPI1*). This is structurally similar to the *lacB* for galactose-6-P isomerase, which is found in *Streptococcus*, *Staphylococcus*, *Lactococcus,* and other bacteria. Proximal to *RPI1,* is *SPS23*, which codes for a glucose 1-dehydrogenase. A second glucose 1-dehydrogenase (*DHG2*) is also present. *RPI1* is relatively uncommon in yeasts and fungi. All three of these genes are similar to bacterial homologs (S3). The genome also includes a yeast ribose-5-phosphate ketol-isomerase (*RKI1*).

Transcripts for *PGI1*, *PFK1,* and *PFK2* were all induced on xylose under oxygen limitation, but were relatively low under aerobic conditions (Fig. 6). Glyceraldehyde-3-phosphate dehydrogenase isoform 3 (*TDH3*), which generates NADH and is the gateway for glycolysis, was induced by oxygen limitation on both glucose and xylose. Transcript levels for *PDC1* and *ADH1* might not be sufficient for high rates of ethanol production on xylose under oxygen-limited conditions. The genome also codes for five NADP(H)-coupled alcohol dehydrogenases (*ADH3, 4, 5, 6* and *7*), which might be important in maintaining cofactor balance between NADH and NADPH. Transcripts for mitochondrial isocitrate dehydrogenases (*IDH1, IDH2*) are elevated on xylose under oxygen-limited conditions, as are those for malate dehydrogenase (*MDH1*), fumarase (*FUM1*), and succinic dehydrogenase (*SDH1*). The transcript for 2-ketoglutarate dehydrogenase (*KGD1*), which generates NADH in the TCA cycle, was reduced during cultivation on xylose.

Responses of other transcripts to carbon sources and oxygen limitation: *P. stipitis* possesses an NAD-specific glutamate dehydrogenase (*GDH2*), a glutamate decarboxylase (*GAD2*), and two NADP-dependent succinate semialdehyde dehydrogenases (*UGA2*, *UGA22*), which constitute a bypass that can convert α-ketoglutarate into succinate and NADH into NADPH when cells are growing on xylose. The NADH-specific *GDH2* is elevated on xylose under oxygen limitation, while the NADPH-linked glutamate dehydrogenase 3 (*GDH3*) is not.

The increased level of *GDH2* could also account for the decreased level of *KGD2* when cells are growing on xylose.

Distinctly different sets of genes are strongly induced under oxygen-limited growth on glucose and xylose (Table S2). On xylose, the transcript for fatty acid synthase 2 (*FAS2*) and the stearoyl-CoA desaturase, (*OLE1),* are strongly induced under oxygen limitation. This induction corresponds with the onset of ethanol production. The *FAS2* transcript is about 1/3 as abundant under the other three conditions tested. *OLE1* is about five fold higher under oxygen limitation when growing on either carbon source. Transcripts for the Ca^{++}-transporting P-type ATPase, *PCM1*, are about 5-fold higher than the aerobic level when cells are grown under oxygen limiting conditions. Transcript levels for the high-affinity inorganic phosphate transporter, *PHO84*, are induced about 10-fold under oxygen limiting conditions.

Genes for polysaccharide degradation: Aside from its capacity for xylose fermentation, *P. stipitis* has several genes and gene families that make it particularly suitable for bioconversion of lignocellulosics. These include an unusual xylanase, several endoglucanases, and numerous β-glucosidases. A blast analysis of the genome with *Trichoderma reesei*, *Bacillus* Family 10 and Family 11 xylanases, and the xylanase (XynA) previously reported as cloned from *P. stipitis* NRRL Y-11543[39] did not turn up any homologous proteins in the *P. stipitis* CBS 6054 genome, and a *P. stipitis* xylanase (*XYN1*) became apparent only during manual annotation. It appears to be a Family 10 glucosidase, but it is not closely related to any other known yeast glycosidases. Domain analysis found this protein to be one of only four Pfam domains unique to *P. stipitis* among the eight fungi examined. It is, however, highly similar to six Family 10 glycoside hydrolases found in *Phanerochaete chrysosporium*. Physically, *XYN1* is found near one terminus of chromosome 4. Our EST data did not provide evidence for its expression.

Three endo, and three exo glucanases (glycoside hydrolases) are represented in the *P. stipitis* genome. The endo-1,4-β-glucanases (*EGC1*, *EGC2*, and *EGC3*) are fairly closely related and all belong to glycoside hydrolase Family 5. *ECG2* is strongly expressed in cells growing on xylose (Table S2). The three exoglucanases (*EXG1*, *EXG2*, *EXG3*) are somewhat more diverse. Two of these appear to be glucan 1,3-β-glucosidases but the function of the third is less certain. The presence of active 1,3-β-glucosidases (laminarinases) can be expected since passalid beetles are known to digest wood containing fungal hyphae, which have large 1,3-β–glucan components.[40] These glycoside hydrolases belong to a family that has relatively low substrate specificity. In addition, *P. stipitis* has three Family 17 soluble cell wall glucosidases

(*SCW4.1*, *SCW4.2* and *SCW11*) along with two Family 17 exo-1,3-β-glucanases (*BGL2*, *BOT2*), all of which are most likely involved in cell wall expansion and growth.

The *P. stipitis* genome includes sequences for seven β-glucosidases (*BGL1-7)* belonging to glycosyl hydrolase Family 3. Enzymes in this family can have activity against cellobiose or xylobiose. Of these seven genes, *BGL4* codes for a protein most similar to classical cellobiases or gentiobiases that have been studied in other yeasts and fungi and *BGL7* is expressed the most when cells are growing on xylose (S2).

The genome contains two sequences for β-mannosidases (*BMS1*, *MAN2*) that belong to glycoside hydrolase Family 2, and which are probably responsible for the capacity of this yeast to grow on and ferment mannan oligosaccharides. Two endo-1,6-α-mannosidases (*DCW1*, *DFG5*) are also present, but these are most likely involved in yeast cell wall expansion during growth, rather than with external polysaccharide degradation, since both are present when cells are growing on either glucose or xylose.

P. stipitis can readily use both glucose and maltose. It has four separate genes for α-glucosidase (*MAL6*, *7*, *8* and *9*). *P. stipitis* also possesses a gene for a putative Family 31 α-glucosidase/α-xylosidase (*YIC1*), of which its closest orthologs are bacterial in origin. Of these, transcripts, only *MAL8* was detected when cells were grown on xylose.

The genome contains almost 60 ORFs that are identified as chitinases according to KOG classification. Only four of these (*CHT1*, *CHT2*, *CHT3*, *CHT4*), however, appear to be true chitinases that might be involved in degradation of insect or fungal cell walls. Many of the remaining models are mucin-like proteins that occur in multiple copies throughout the genome. *MUC1* appears at least four times in nearly identical copies. Segments of *MUC1* proteins exist in approximately 25 copies in the genome, suggesting expansion through frequent duplication.

Respiration system: The respiration system of *P. stipitis* differs from that of *S. cerevisiae* in many aspects. First, as has been documented previously, *P. stipitis* has a SHAM-sensitive terminal alternative oxidase (*AOX1* or *STO1*) that enables the cells to oxidize ubiquinone.[41] *S. cerevisiae* lacks this alternative oxidase. *P. stipitis* has genes coding for the complete proton-translocating NADH dehydrogenase complex (Complex I), which is also lacking in *S. cerevisiae*. Based on these differences, Transcript levels for *AOX1* are up regulated on xylose under aerobic conditions and on glucose under oxygen limitation, but was not found on xylose under oxygen limitation.

Aromatic catabolism: The *P. stipitis* genome includes a number of genes that appear to be involved in aromatic catabolism. Most conspicuous is a family of salicylate hydroxylases (*NHG1.1*, *NHG 1.2*, *NHG2*, *NHG3*, *NHG4*) that are similar to homologs from *Pseudomonas putida* and a series of plant-related proteins. These are not clustered, but rather are scattered throughout the genome. Only *NHG2* shows conservation relative to *D. hansenii. The rest of the* genes and their surrounding loci have no identity to proteins found in *C. albicans* or *D. hansenii*. These findings suggest that the genes for salicylate hydroxylase are the result of relatively recent introduction and amplification.

Alternative codon usage: *P. stipitis* uses the alternative yeast nuclear codon (12) that substitutes serine for leucine when CUG is specified.[42] To understand this feature better we examined whether or not CUG codon usage was evenly distributed in the genome. A count of CUG usage showed 15,265 occurrences in 4238 ORFs, or about 72% of all gene models (S4). Nine out of the 21 ORFs having 18 or more CUGs in the gene model occurred at or near a terminus of chromosomes 4, 8, 7 or 1. All gene models having a large number of CUGs in the open reading frame were large (>2,500 bp), very large (>5,000 bp), repetitive, hypothetical, or poorly defined. A plot of expression level vs. CUG usage for 94 annotated ORFs that contained CUG codons generally showed higher expression levels with lower CUG frequency. Two exceptions were the conserved sequences *ENA5* and *SEC31*, which were both highly expressed and which contained 4 and 14 CUGs, respectively (SF2).

Adjacent and proximal genes with related functions: This study found numerous intriguing instances of adjacent and proximal genes with related functions. These included genes for pentose phosphate metabolism, glycolysis, urea metabolism, sugar assimilation and possibly aromatic catabolism.

XYL1 is adjacent to a putative gene for *MIG1* (CREA), which is a transcription factor involved in glucose repression. This is a complex locus that includes two other transcriptional regulators (*SPT8* and *STB4*) and sorbitol dehydrogenase (*SOR4*) within about 19.8 kbp. The putative sugar transporter, *XUT2* is adjacent to *SOR3*, which appears to be L-arabinitol 4-dehydrogenase that is highly similar to *XYL2*, and *SOR3* is in turn is adjacent to formaldehyde transketolase, *DHA1*, which is a homolog to transketolase, *TKT1*. This latter gene is immediately adjacent to one of the two principal genes for NADH-coupled alcohol dehydrogenase activity, *ADH2*. *OLE1*, which converts fatty acids into unsaturated fatty acids, is also in this locus.

A gene for *DUR1* (*DUR1,2,* urea amidolyase) - which codes for both urea carboxylase, and allophanate hydrolase activities - is immediately adjacent to *DUR3.1*, which codes for urea transport, on chromosome 1. This latter protein shares strong similarity with the second gene for urea transport, *DUR3.2,* which is located on one terminus of chromosome 6, and *DUR5.1*, which is elsewhere on chromosome 6. Multiple copies of urea transporters (e.g. *DUR4*, *DUR5.2*, *DUR5.3*, *DUR8*) are found throughout the genome, which suggests that this function might be required at a high level.

β-Glucosidases were often found adjacent or proximal to genes with related functions. For example, on either side of the Family 5 β-1,4 endoglucanase *EGC2* ,one finds *BGL5* and the probable hexose transporter, *HXT2.4*. *BGL6* is adjacent to *EGC1,* and *BGL3* is adjacent to the sugar transporter, *SUT3*. *BGL1* is adjacent to *SUT2* on chromosome 4. Both of the putative β-mannosidases (*BMS1*, *MAN2*) are adjacent or proximal to putative lactose permeases (*LAC3* and *LAC2*, respectively).

One of the most conspicuous examples of tandem genes with related functions was found in a putative *MAL3* locus (Fig. 7). This site extends over approximately 16 kbp on chromosome 6. Two out of the six genes appear to be conserved in *C. albicans,* and four out of the six are conserved in *D. hansenii.* The site contains the putative maltose permease *MAL3*, and the α-glucosidase, *AGL1*. Adjacent but in an opposite orientation to *MAL3,* is the putative maltose permease, *MAL5*, which is adjacent to *YIC1*, a putative α-glucosidase belonging to glycosyl hydrolase Family 31. Most of its closest orthologs appear to be bacterial genes (S3). Flanking this complex of four genes are the putative fungal transcriptional regulatory protein, *SUC1.2*, which is similar to MAL-activator proteins in the complex MAL3 locus of S. cerevisiae,[43] and a second putative fungal-specific regulatory protein, *SUC1.4*. Elsewhere in the genome, on chromosome 6, the α-glucosidase, *MAL8*, is immediately adjacent to the maltose permease, *MAL4*.

The putative salicylate hydroxylases also appear to have permeases, oxidases or genes coding for aromatic degradation proximal to them on the chromosome. For example, *NHG4* is flanked by two acetyl coenzyme A oxidases (*POX1* and *ACOX2*), and *NHG1.1* and *NHG1.2* are each adjacent to the transporters *HOL41* and *HOL42*, respectively. Adjacent to *NHG3* is the putative allantoate permease, DAL10 and nearby is an aromatic ring hydroxlase, SAL1. Proximal to *NHG1.1* is a putative cinnamyl Co-A reductase (CAD1) and a gene for 5-carboxymethyl-2-hydroxymuconate delta-isomerase, (*UMH1*), both of which could have roles in aromatic *catabolism.* Also proximal to *NHG1.2* is the fumarylacetoacetate hydralase, *FML1*, which is

similar to genes for proteins involved in aromatic degradation. Finally *NHG2*, the only gene in this family that has any conservation in *D. hansenii*, is flanked on either side by the E1 component of α-ketoglutarate dehydrogenase, *KGD1*, and a probable oxidoreductase.

A few other examples of tandem gene structures were noted. Two *MUC1*-like models (*MUC1.7* and *MUC1.10*), segments of which also occur in multiple copies, are adjacent to one another in chromosome 8. Two copies of similar, but not identical ESS1 genes (*ESS1.1*, *ESS1.2*), which code for peptidyl-prolyl cis-trans isomerase, exist in tandem adjacent to a hypothetical protein that occurs in multiple copies (e.g. HMC1). Two *MUC1*-like models (*MUC1.7* and *MUC1.10*), segments of which also occur in multiple copies, are adjacent to one another in chromosome 8.

Viral and transposon elements

We identified a number of transposable elements using a composite library of fungal repeats.[44] The most abundant elements include LTR retrotransposons Tdh5, Tdh2, Tse5, pCal, most of which were previously reported in hemiascomycetes including the *D. hansenii* genome,[45] and single copies of DNA mediated elements Ty1-I, Mariner-5, and Folyt1 were reported earlier in fungi.[46] We have identified multiple copies of a highly variable element that appears to be similar to the transposons Tdh5 and Tdh2, which we have termed Tps5. These are scattered throughout the genome with one well-defined locus on each chromosome (S4). Portions of these elements are actively transcribed and can be detected as ESTs (S2). Certain genes in proximity of these repeat elements appear in multiple copies throughout the genome (e.g., 10 copies of HMC-related genes).

DISCUSSION

By aligning gene models with expression profiles and vista analyses, we were able to determine gene conservation, expression, and linkage patterns. Domain analysis was more useful in identifying the genes absent from *P. stipitis* than in highlighting those present, because the latter tend to be widespread rather than unique. The high number of homology based gene models (84%), is probably attributable to improved identification resulting from better data sets and the quality of our EST library. The average gene density falls between those of *D. hansenii* and *Y. lipolytica* and is in line with their relative genome sizes.

Codon usage

Three lines of evidence point to *P. stipitis* using alternative yeast nuclear codon system (12), in which CUG codes for serine rather than leucine. The first is that *P. stipitis* appears to be closely

related to other yeasts that use this system.[62] Second, the *Sh ble* gene can impart resistance to Zeocin in *P. stipitis* after its CUG codons are engineered into different leucine codons, but the native gene does not.[42] Third, the genome contains the characteristic tRNA(Ser)CAG gene that is used to transfer serine to the nascent polypeptide.[63, 64] The high frequency of CUG usage in large putative ORFs occurring at chromosome termini has not been previously reported.

Syntenic relationships

P. stipitis chromosomes are evolving through both translocations within the genome and local inversion. Translocations within any one chromosome do not appear to be favored over sites in other chromosomes. The large number of genome rearrangements in yeasts seemingly obliterates any meaningful syntenic relationships except between the most closely related yeast species. In the present study only one strain was sequenced, so we cannot draw conclusions about the frequency of translocations within the species, however, we used MAUVE[47] to compare the synteny of fully assembled yeast genomes over greater taxonomic distances (*P. stipitis* vs. *D. hansenii*, *C. albicans*, and *S. cerevisiae*), and we observed increasing fragmentation with taxonomic divergence (data not shown). This technique, however, is based on nucleotide sequence not protein identity, and it could not show whether local assemblages of genes with related function were conserved over groups retained by chance. The high rates of genomic rearrangement observed here between *P. stipitis* and *D. hansenii* are consistent with previously reported rates of rearrangement for the closely-related species *D. hansenii* and *C. albicans*.[48]

Regulation

Fermentation requires coordinated regulation of the central metabolic pathways because the substrate is being converted into more reduced and more oxidized portions at the same time. This process is complicated during the conversion of xylose, since some oxygen is necessary to enable cell growth. The EST analysis gave clear evidence of transcript levels in response to carbon source and aeration. The ESTs also produced a high-quality genomic sequence and annotations for *P. stipitis* to provide insights into the biology of this organism.

Genes for xylose assimilation were found only in the absence of glucose. *GND1* and *TKT1* were significantly elevated on xylose, which reflects the increased activity of the PPP for xylose metabolism. *PGI1, PFK1* and *PFK2* were elevated most with cells growing on xylose under oxygen-limited conditions. Presumably elevated *PGI1* is necessary to cycle F6P through the

oxidative PPP while *PFK1* and 2 take F6P into glycolysis. *GLK1* was elevated in cells growing on xylose aerobically, which could reflect carbon catabolite de-repression.

The *P. stipitis* genome has many traits that suit it well for the fermentation of xylose and other sugars from lignocellulose. The CBS 6054 strain was isolated from insect larvae, and other yeast strains closely related to *P. stipitis* have been isolated from the guts of wood-inhabiting passalid beetles,[14] which suggests that this yeast has evolved to inhabit an oxygen- limited environment rich in partially digested wood. The presence of numerous genes for endoglucanases and β-glucosidases, along with xylanase, mannanase, and chitinase activities suggests that these yeasts could be metabolizing polysaccharides in the beetle gut. No clear evidence was found for enzymes capable of degrading lignin-related compounds, but many genes were present for salicylate catabolism. Various strains of *P. stipitis* previously have been reported to ferment cellobiose to ethanol,[49-51] so it is likely that these are active during growth and fermentation. Exo-1,4-cellobiohydrolases, which are responsible in part for the degradation of cellulose, produce cellobiose from cellulose and most endo-1,4-xylanases produce a mixture of xylose, xylobiose and xylotriose. β-glucosidases and β-xylosidase activities are therefore very useful traits because cellobiose and xylobiose fermentation can increase cellulose saccharification when combined with cellulose saccharification.

Respiration and redox balancing:

Excess NADH is generated during growth on xylose,[52] which necessitates some mechanism to balance cofactor oxidation. *KGD2*, which generates NADH in the TCA cycle, was three times higher in cells growing on glucose over those on xylose. Gdh2 consumes NADH while generating NAD^+, and leads into a pathway that eventually consumes NADH while generating NADPH. A similar pathway was previously engineered in *S. cerevisiae* to reduce cofactor imbalances when cells are growing on xylose,[53] but it appears to exist naturally in *P. stipitis*.

P. stipitis has a complete mitochondrial respiration system including NADH dehydrogenase Complex I. *S. cerevisiae* lacks Complex I, so it has less capacity for ATP generation through oxidative phosphorylation. The presence of *AOX1* suggests that this yeast can scavenge for oxygen when it is present in trace amounts, but the exact role of this enzyme in xylose metabolism is not clear since *AOX1* transcripts were present at a lower level when cells were growing on xylose under oxygen limiting conditions.

The abundance of genes for NADP(H) oxidoreductase reactions suggests that *P. stipitis* is capable of various strategies for balancing NAD and NADP-specific cofactors under oxygen

limiting conditions. Not least among these is *FAS2*, which appears to be highly active when cells are growing under oxygen limited conditions on xylose, and which could be a redox sink for the cell.

Fas2 synthesizes long chain acyl-CoA precursors of fatty acids from malonyl-CoA, Acetyl-Co-A, NADH and NADPH. As such, it could serve as a reductant sink when cells are growing under oxygen limitation on xylose. Genes were present for the other activities in glutamate dehydrogenase shunt, but transcripts were not detected, so further transcriptional and metabolite studies are required to determine how this bypass might function. Transcripts for fatty acid synthesis including *OLE1* and, particularly, *FAS2* were elevated in oxygen limited, xylose-grown cells (XOL), indicating that substantial amounts of reductant might be channeled into lipid synthesis under oxygen limitation. More reductant can be stored for each gram of carbon in lipid than in ethanol, so this might enable the cells to consume excess reductant when growing on xylose under oxygen limiting limited conditions.

Functional localization

Co-location of a gene from an expanded family with a gene having different but related function (e.g. a permease with a hydrolase for maltose) seems to occur with high frequency in *P. stipitis*. As we show here, co-location occurs between genes that have totally different origins – and different members of the same closely related gene family are found co-located with various genes having functions that are each related to members of that family in different ways. For example, this was observed for the salicylate hydroxylases and the *SUT* family of sugar transporters.

Similar examples are known in yeast. Members of multi-gene families are often found near *S. cerevisiae* telomers and are repeated elsewhere in the genome. Zakian has proposed that the concentration of multigene families in the telomere-adjacent regions may reflect a recombination mediated dispersal mechanism. [54] The fact that some *P. stipitis* genes at chromosome termini are found proximal to genes with related functions deeper within the chromosomes suggests that duplication or translocation might confer a survival advantage.

Genes in telomeric regions might be under less selective pressure due to silencing. In *S. cerevisiae* the COMPASS histone methyltransferase carries out telomeric silencing of gene expression,[55] and the *P. stipitis* genome contains a homolog (*SET1*). Without selective pressure, genes in the telomeric regions might diverge more rapidly. We noted that genes

occurring at chromosome termini often had a high frequency of CUG usage, which might be indicative of genetic drift.

The proximal co-location of glucosidases to corresponding sugar transporters and urea amidolyase adjacent to urea permease, suggests that these loci might be co-regulated. In *S. cerevisiae*, genes for α-glucosidase and maltose permease are adjacent. Each complete MAL locus consists of maltose permease, maltase, and a transcription activator. [59, 60] The MAL loci each map to the telomeric region of a different chromosome.[61] The observations reported here extend functional co-location to endoglucanase, β-glucosidase, and urea metabolism.

Co-regulated genes distal from one another are physically co-localized in nuclear "transcriptional factories". Osborne et al. have proposed that linked genes are more likely to occupy a transcriptional factory than genes in trans. In the human transcriptional map, genes occur in gene dense regions with increased gene expression.[56] Adjacent eukaryotic genes are more frequently co-expressed than is expected by chance and co-expressed neighboring genes are often functionally related. For example, in *Arabidopsis*, 10% of the genes occur in 266 groups of large-co-expressed chromosomal regions distributed throughout the genome.[57] The model advanced by Bartlett et al.[58] encapsulates the advantages of proximal co-location of actively transcribed genes: The concentration of RNA polymerase II is 1000-fold higher in a transcription factory than in the whole nucleus; modifications occurring during transcription leave the promoter open to new transcript initiation; after being released at the termination, promoters in the vicinity of a transcription factory are more likely to encounter machinery for transcriptional initiation again.

The adjacency of *DUR1,2* and *DUR3.1* in a single locus is notable because *DUR1,2* has merged the functions for urea carboxylase and allophanate hydrolase activities into a single protein, urea amidolyase. In bacteria, genes for sequential reactions in biochemical pathways are often found in operons. In higher eukaryotes evolution tends to favor the fusion of proteins coding for sequential related biochemical functions. In yeasts for example, separate genes code for sequential steps in uracil synthesis. *URA3* codes for orotidine-5'-phosphate (OMP) decarboxylase while two isozymes, *URA5* and *URA10*, code for orotate phosphoribosyltransferase. In *A. niger a URA3 homolog, PYRF* is present and two isozymes code for both uridine 5'- monophosphate synthase and orotate phosphoribosyltransferase. In *Xenopus tropicalis* and *Populus trichocarpa* only genes for the fused proteins are present.

Conclusions

Clearly the *P. stipitis* genome is endowed with numerous genes and physiological features that enable it to ferment a wide variety of sugars derived from lignocellulose. Surprisingly it also seems to have a high capacity for cellobiose degradation. Evidence for lignin degradation is less clear, but also present.

Because this is a completely finished genome, we have been able to discern structural features that suggest evolutionary aspects: When genes with related functions are found proximal to one another, the combined gene activities enhance survival. The separate genes can occur in different regions of the genome, but proximal location could affect their mutual function and the probability of co-inheritance. Duplicated genes might persist in the genome because activities of their gene products are limiting and an increased copy number confers a selective advantage. Following duplication, co-location with various other related genes could further increase their functions and perhaps contribute to differentiation. In this model regulation of expression is not just a function of transcriptional activators on individual promoters, but also the product of the coding and non-coding elements in the locus.

One implication of this study is that expression, and perhaps regulated co-expression, may depend greatly upon location in the genome. Aside from co-location, other chromosomal elements such as transcriptional activators may be important for migration of promoters to transcriptional factories. Alternately, such factories might arise dynamically by the co-location of multiple genes under control of similar cis-acting promoters and transcriptional activators. Expression mapping or detailed study of the corresponding cis-acting promoters could provide more insight. If some gene families persist in multiple copies simply from the advantage of higher transcript levels, then evolution toward higher promoter strength would seem sufficient. If they have been acquired from divergent sources, however, codon usage might also limit translational expression.

If chromosomal co-location does affect expression, this would have strong implications with respect to the design and placement of genes for metabolic pathway engineering.

METHODS

Yeast strain

Pichia stipitis Pignal (1967), synonym *Yamadazyma stipitis* (Pignal) Bilon-Grand (1989), (NRRL Y-11545 = ATCC 58785 = CBS 6054 = IFO 10063) was obtained as a lyophilized powder from Dr. Cletus P. Kurtzman of the USDA ARS Culture Collection (NRRL), Peoria IL. It was revived

and streaked on YPD agar to obtain isolated colonies. A single colony was transferred to 150 ml of YPD broth. To test for contamination, the overnight was observed under the microscope and streaked in both YPD and LB plates. For fermentation studies, cells were grown in 125 ml Erlenmeyer flasks containing 50 ml of 1.67 g/l yeast nitrogen base (YNB) with 2.27 g/l urea and 80 g/l xylose. The YNB and urea solutions were filter sterilized in a 20x solution and added to the sugar, which was sterilized separately by autoclaving. For mRNA preparation, cells were growing in yeast extract, peptone, dextrose (YPD), which was prepared as described in Kaiser et al.[65] except that sugars were autoclaved separately from the basal medium. Yeast peptone xylose (YPX) was similar to YPD but replaced dextrose with xylose. Preparation of mRNA was by the method previously described.[42]

DNA preparation

Yeast genomic DNA was prepared following the protocol of Burke et al.[66] Two extra phenol:chloroform/chloroform extractions and ethanol precipitation were carried out. To prevent shredding of the DNA, the sample was never vortexed. The final gDNA concentration was 500 ng/µl as determined by optical density at 260 nm.

cDNA library construction and sequencing:

P. stipitis CBS 6054 was grown at 30 °C in 200 ml of either YPD or YPX in either a 2.8 l flask shaken at 300 rpm or a 500 ml flask shaken at 50 rpm. Aerobic cultures were inoculated with a low cell density (0.025 mg/ml), shaken at 200 rpm and harvested at a cell density of less than 0.5 mg/ml. Oxygen limited cultures were inoculated with a high cell density (2.5 mg/ml), shaken at 100 rpm and harvested at 5 mg/ml. Cells were collected by centrifugation at 4 °C and 9279 x *g*. Cells were suspended in water and centrifuged at 835x*g* for 5 min. Cells were then frozen in liquid N_2. Poly A+ RNA was isolated from total RNA for all four *P. stipitis* samples using the Absolutely mRNA Purification kit (Stratagene, La Jolla, CA). cDNA synthesis and cloning was a modified procedure based on the “SuperScript plasmid system with Gateway technology for cDNA synthesis and cloning” (Invitrogen). 1-2 µg of poly A+ RNA, reverse transcriptase SuperScript II (Invitrogen) and oligo dT primer (5'- GACTAGTTCTA GATCGCGAGCGGCCGCCC TTTTTTTTTTTTTTT -3') were used to synthesize first strand cDNA. Second strand synthesis was performed with *E. coli* DNA ligase, polymerase I, and RNaseH followed by end repair using T4 DNA polymerase. The SalI adaptor (5'- TCGACC CACGCGTCCG and 5'- CGGACGCGTGGG) was ligated to the cDNA, digested with NotI (NEB), and subsequently size selected by gel electrophoresis (1.1% agarose). Size ranges of

cDNA were cut out of the gel (L: 600-1.2kb, M: 1.2kb-2kb, H: >2kb) and directionally ligated into the SalI and NotI digested vector pCMVsport6 (Invitrogen). The ligation was transformed into ElectroMAX T1 DH10B cells (Invitrogen).

Library quality was first assessed by PCR amplification the cDNA inserts of 20 clones with the primers M13-F (GTAAAACGACGGCCAGT) and M13-R (AGGAAACAGCTATGACCAT) to determine insert rate. Clones for each library were inoculated into 384 well plates (Nunc) and grown in LB for 18 hours at 37 C. DNA template for each clone was prepared by RCA and sequenced using primers (FW: 5'- ATTTAGGTGACACTA TAGAA and RV 5' – TAATACGACTCACTATAGGG) using Big Dye chemistry (Applied Biosystems). The average read length and pass rate were 753 (Q20 bases) and 96%, respectively.

EST sequence processing and assembly:

The JGI EST Pipeline begins with the cleanup of DNA sequences derived from the 5'and 3' end reads from a library of cDNA clones. The Phred software[67] is used to call the bases and generate quality scores. Vector, linker, adapter, poly-A/T, and other artifact sequences are removed using the Cross_match software,[67] and an internally developed short pattern finder. Low quality regions of the read are identified using internally developed software, which masks regions with a combined quality score of less than 15. The longest high quality region of each read is used as the EST. ESTs shorter than 150 bp are removed from the data set. ESTs containing common contaminants such as *Escherichia coli*, common vectors, and sequencing standards are also removed from the data set.

EST Clustering is performed ab-initio, based on alignments between each pair of trimmed, high quality ESTs. Pair-wise EST alignments are generated using the Malign software (Chapman, et. al., Unpublished), a modified version of the Smith-Waterman algorithm,[68, 69] which was developed at the JGI for use in whole genome shotgun assembly. ESTs sharing an alignment of at least 98% identity, and 150 bp overlap are assigned to the same cluster. These are relatively strict clustering cutoffs, and are intended to avoid placing divergent members of gene families in the same cluster. However, this could also have the effect of separating splice variants into different clusters. Optionally, ESTs that do not share alignments are assigned to the same cluster, if they are derived from the same cDNA clone.

EST cluster consensus sequences were generated by running the Phrap software[67] on the ESTs comprising each cluster. All alignments generated by malign are restricted such that they will always extend to within a few bases of the ends of both ESTs. Therefore, each cluster

looks more like a 'tiling path' across the gene, which matches well with the genome based assumptions underlying the Phrap algorithm. Additional improvements were made to the phrap assemblies by using the 'forcelevel 4' option, which decreases the chances of generating multiple consensi for a single cluster, where the consensi differ only by sequencing errors.

Genome Assembly

The initial data set was derived from four whole-genome shotgun (WGS) libraries: one with an insert size of 3 KB, two with insert sizes of 8 KB, and one with an insert size of 35 KB. The reads were screened for vector using cross_match, then trimmed for vector and quality.[27] Reads shorter than 100 bases after trimming were then excluded. The data was assembled using release 1.0.1b of Jazz, a WGS assembler developed at the JGI.[27, 70] A word size of 14 was used for seeding alignments between reads. The unhashability threshold was set to 50, preventing words present in more than 50 copies in the data set from being used to seed alignments. A mismatch penalty of -30.0 was used, which will tend to assemble together sequences that are more than about 97% identical. The genome size and sequence depth were initially estimated to be 16.5 MB and 9.3, respectively. The assembly contained 394 scaffolds, with 16.4 MB of sequence, of which 4.5% was gap. The scaffold N/L50 was 5/1.46 MB, while the contig N/L50 was 21/262 KB. The sequence depth derived from the assembly was 8.77 ± 0.05.

Gap closure and finishing

To perform finishing, initial read layouts from the *P. stipitis* whole genome shotgun assembly were converted into our Phred/Phrap/Consed pipeline.[71] Following manual inspection of the assembled sequences, finishing was performed by resequencing plasmid subclones and by walking on plasmid subclones or fosmids using custom primers. All finishing reactions were performed with 4:1 BigDye to dGTP BigDye terminator chemistry (Applied Biosystems). Repeats in the sequence were resolved by transposon-hopping 8kb plasmid clones. Fosmid clones were shotgun sequenced and finished to fill large gaps, resolve large repeats or to resolve chromosome duplications and extend into chromosome telomere regions. Finished chromosomes have no gaps and the sequence has less than 1 error in 100,000 bp.

Gene prediction and annotation

The JGI Annotation Pipeline combines a suite of gene prediction and annotation methods. Gene prediction methods used for analysis of the *P. stipitis* genome include *ab initio* Fgenesh, [72]

570 homology-based Fgenesh+ (www.softberry.com) and Genewise,[73] and an EST-based method
571 estExt [Grigoriev, unpublished]. Predictions from each of the methods were taken to produce
572 'the best' single gene model per every locus. The best model was determined on basis of
573 homology to GenBank proteins and EST support.

574 Every predicted gene was annotated using Double Affine Smith-Waterman alignments
575 (www.timelogic.com) with Swissprot and KEGG proteins. Protein domains were predicted using
576 InterProScan[74, 75] against various domain libraries (Prints, Prosite, PFAM, ProDom, SMART,
577 etc). Individual annotations have been then summarized according to Gene Ontology,[76]
578 eukaryotic orthologous groups (KOGs),[30] and KEGG metabolic pathways.[77]

579 **Phylogenetic tree reconstruction of sequenced fungal genomes**

580 A multiple sequence alignment of 94 single copy genes present in 26 taxa was constructed
581 using the MUSCLE 3.52 program,[78] trimmed using Gblocks 0.91b and was used as input for the
582 maximum likelihood tree reconstruction program PHYML (4 rate categories, gamma +
583 invariants, 100 bootstrap replicates) resulting in a fully resolved tree with all but one node
584 having bootstrap values of 100. Figure 4 represents the portion of the tree describing
585 relationships between the genomes of interest for this analysis.

586 **Comparative analysis of the 6 yeast genomes**

587 Comparisons of the phylic patterns of gene family distributions of *Pichia stipitis* and five hemi-
588 ascomycete yeasts (*P. stipitis*, *S. cerevisiae*, *C. glabrata*, *K. lactis*, *D. hansenii* and *Yarrowia*
589 *lipolytica*) were done using the PhIGs orthology database. The PhIGs resource generated
590 clusters of genes at each node on the evolutionary tree representing the descendents from a
591 single ancestral gene existing at that node. This allows for the comparisons of the
592 presence/absence patterns of gene families across the six species avoiding confusion from
593 paralogous genes. In this analysis, gene families specific to a single species are defined as
594 those having a minimum of two family members.

595 **Expression analysis**

596 To enable complete sampling of the expressed genes, we generated four separate EST libraries
597 by growing cells on glucose or xylose under aerobic or oxygen limited conditions. A set of
598 19,635 *P. stipitis* ESTs was sequenced from the four libraries and clustered into 4,085
599 consensus sequences. Ninety-four percent (3,839) of the clusters were mapped to the genome
600 and the numbers of hits for each consensus cluster was used to estimate EST frequency under

each growth condition. An absolute majority of unplaced ESTs had problems with the sequences so the data indicates completeness and accurateness of genome assembly. Only 44% of the transcripts were represented by more than one EST cluster-hit under any one of the four growth conditions. The cluster-hit enumeration represents only a single biological sample for each off the four conditions, so these observations must be interpreted with care and be limited to the 200 to 400 most abundant gene models in which at least 1 transcript was recovered under each of the four conditions. However the relative abundances of these ESTs under each of the four conditions provided a preliminary expression analysis.

Nucleotide sequence accession

[Note: accession numbers in process]

ACKNOWLEDGEMENTS

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36, and by the USDA, Forest Service, Forest Products Laboratory. The authors are grateful to C. P. Kurtzman of the USDA ARS Culture Collection (NRRL) for providing the *P. stipitis* stock culture, to W.Huang, G.Werner and his group of the JGI for engineering support of annotation, to A. Polyakov and I.Dubchak of the JGI for VISTA analysis, to A. Darling for advice and support in MAUVE analysis, W. R. Kenealy, T. A. Kuster and Mark Davis of the USDA Forest Products Laboratory for carrying out continuous culture studies, providing photomicrographs, and analyzing fermentation products, and to James Cregg, and Lisbeth Olsson and Jennifer Headman Van Vleet for critical readings.

TABLES

Table 1. General characteristics of several yeast genomes

Species	Genome Size (Mb)	Avg G+C Content (%)	Total CDS	Avg Gene Density (%)	Avg G+C in CDS (%)	Avg CDS size (codons)	Maximum CDS size (codons)	Source
P. stipitis	15.4	41.1	5841	55.9	42.7	493	4980	JGI
S. cerevisiae	12.1	38.3	5807	70.3	39.6	485	4911	Dujon [32]
C. glabrata	12.3	38.8	5283	65.0	41.0	493	4881	Dujon[32]
K. lactis	10.6	38.7	5329	71.6	40.1	461	4916	Dujon[32]
D. hansenii	12.2	36.3	6906	79.2	37.5	389	4190	Dujon[32]
Y. lipolytica	20.5	49.0	6703	46.3	52.9	476	6539	Dujon[32]

Table 2 Phyletic patterns of yeast protein families[1]

Pattern[2]	Familes	Proteins
Families universal to all- Genes that occur more than once in each genome and have no matches to any other fungal genomes.		
sckdyp	2343	16,922
Families missing in one species		
_ckdyp	35	184
s_kdyp	54	359
sc_dyp	35	184
sck_yp	106	549
sckd_p	351	1977
sckdy_	81	442
Species-specific families		
s_____	35	92
_c___	5	12
__k___	21	53
___d__	30	87
____y_	121	338
_____p	25	72

[1] Data generated using the PhIGs tool (Phylogenetically Inferred Groups), http://phigs.org

[2] Abbreviations: p, *P. stipitis*; s, *S. cerevisiae*; c, *C. glabrata*; k, *K. lactis*; d, *D. hansenii*; y, *Y. lipolytica*

FIGURES

Figure 1.

Fermentation of xylose by *Pichia stipitis* CBS 6054 in minimal medium

Figure 2.

Morphology under various conditions. (A) *Pichia stipitis* growing exponentially with bud scars; (B) *P. stipitis* hat-shaped spores seen from top and side; (C) Pseudomycelia formed under carbon-limited continuous culture. Photo by Thomas Kuster, USDA, Forest Products Laboratory.

Figure 3.

Distribution of gene models as determined by KOG (clusters of orthologous groups) classification.

Figure 4.

Phylogenetic tree of seven sequenced hemiascomycetous yeast genomes based on multiple alignment of 94 single copy genes conserved in 26 taxonomic groups (see Methods). Numbers next to each branch correspond to the number of families (clusters) specific to a genome or a group of genomes leading to this node.

Figure 5.

Orthologous chromosomal segments observed between *Pichia stipitis* and *Debaryomyces hansenii*.

Figure 6.

Expression of transcripts in the central metabolic pathways of *Pichia stipitis*. Cells were grown batch-wise on minimal defined medium under four conditions: glucose aerobic (GA), xylose aerobic (XA), glucose oxygen limited (GOL) and xylose oxygen-limited (XOL). cDNA was harvested and sequenced.

Figure 7.

The *MAL3* locus of *Pichia stipitis*. Two putative a-glucosidases (*YIC1*, *AGL1*) and two putative maltose permeases (*MAL3*, *MAL5*) are co-located along with two putative fungal transcriptional regulators (*SUC1.2*, *SUC1.4*) within 16 kbp on chromosome 6.

BIBLIOGRAPHY

1. Jeffries, T.W. & Shi, N.Q. Genetic engineering for improved xylose fermentation by yeasts. *Adv Biochem Eng Biotechnol* **65**, 117-161 (1999).
2. Hinman, N.D., Wright, J.D., Hoagland, W. & Wyman, C.E. Xylose fermentation - an economic analysis. *Appl. Biochem. Biotechnol.* **20-1**, 391-401 (1989).
3. Gulati, M., Kohlmann, K., Ladisch, M.R., Hespell, R. & Bothast, R.J. Assessment of ethanol production options for corn products. *Bioresource Technol* **58**, 253-264 (1996).
4. Saha, B.C., Dien, B.S. & Bothast, R.J. Fuel ethanol production from corn fiber - Current status and technical prospects. *Appl Biochem Biotech* **70-2**, 115-125 (1998).
5. Kurtzman, C.P. *Candida shehatae*--genetic diversity and phylogenetic relationships with other xylose-fermenting yeasts. *Antonie Van Leeuwenhoek* **57**, 215-222 (1990).
6. Vaughan Martini, A.E. Comparazione dei genomi del lievito *Pichia stipitis* e de alcune specie imperfette affini. *Ann. Fac. Agr. Univ. Perugia* **38B**, 331-335 (1984).
7. Melake, T., Passoth, V.V. & Klinner, U. Characterization of the genetic system of the xylose-fermenting yeast *Pichia stipitis*. *Curr Microbiol* **33**, 237-242 (1996).
8. van Dijken, J.P., van den Bosch, E., Hermans, J.J., de Miranda, L.R. & Scheffers, W.A. Alcoholic fermentation by 'non-fermentative' yeasts. *Yeast* **2**, 123-127 (1986).
9. du Preez, J.C., Bosch, M. & Prior, B.A. Xylose fermentation by *Candida shehatae* and *Pichia stipitis* - Effects of pH, temperature and substrate concentration. *Enzyme Microb Technol* **8**, 360-364 (1986).
10. du Preez, J.C., van Driessel, B. & Prior, B.A. Ethanol tolerance of *Pichia stipitis* and *Candida shehatae* strains in fed-batch cultures at controlled low dissolved-oxygen levels. *Appl Microbiol Biotechnol* **30**, 53-58 (1989).
11. Hahn-Hägerdal, B. & Pamment, N. Microbial pentose metabolism. *Appl Biochem Biotech* **113-16**, 1207-1209 (2004).
12. Nigam, J.N. Ethanol production from wheat straw hemicellulose hydrolysate by *Pichia stipitis*. *J Biotechnol* **87**, 17-27 (2001).
13. Nardi, J.B. et al. Communities of microbes that inhiabit the changing hindgut landscape of a subsocial beetle. *Arthropod Structure & Development* **35**, 57-68 (2006).
14. Suh, S.O., Marshall, C.J., McHugh, J.V. & Blackwell, M. Wood ingestion by passalid beetles in the presence of xylose-fermenting gut yeasts. *Mol Ecol* **12**, 3137-3145 (2003).
15. Suh, S.O., White, M.M., Nguyen, N.H. & Blackwell, M. The status and characterization of *Enteroramus dimorbhus*: a xylose-fermenting yeast attached to the gut of beetles. *Mycologia* **96**, 756-760 (2004).
16. Ozcan, S., Kotter, P. & Ciriacy, M. Xylan-hydrolyzing enzymes of the yeast *Pichia stipitis*. *Applied Microbiology And Biotechnology* **36**, 190-195 (1991).
17. Lee, H., Biely, P., Latta, R.K., Barbosa, M.F.S. & Schneider, H. Utilization of xylan by yeasts and its conversion to ethanol by *Pichia stipitis* strains. *Appl Environ Microbiol* **52**, 320-324 (1986).

18. Targonski, Z. Biotransformation of lignin-related aromatic-compounds by *Pichia stipitis* Pignal. *Zbl Mikrobiol.***147**, 244-249 (1992).

19. Jeffries, T.W. & Jin, Y.S. Metabolic engineering for improved fermentation of pentoses by yeasts. *Appl Microbiol Biotechnol* **63**, 495-509 (2004).

20. Sonderegger, M., Jeppsson, M., Hahn-Hägerdal, B. & Sauer, U. 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 (2004).

21. Harhangi, H.R. et al. Xylose metabolism in the anaerobic fungus *Piromyces* sp. strain E2 follows the bacterial pathway. *Arch Microbiol* **180**, 134-141 (2003).

22. Karhumaa, K., Hahn-Hägerdal, B. & Gorwa-Grauslund, M.F. Investigation of limiting metabolic steps in the utilization of xylose by recombinant *Saccharomyces cerevisiae* using metabolic engineering. *Yeast* **22**, 359-368 (2005).

23. Jin, Y.S., Laplaza, J.M. & Jeffries, T.W. *Saccharomyces cerevisiae* engineered for xylose metabolism exhibits a respiratory response. *Appl Environ Microbiol* **70**, 6816-6825 (2004).

24. Passoth, V., Cohn, M., Schafer, B., Hahn-Hägerdal, B. & Klinner, U. Analysis of the hypoxia-induced *ADH2* promoter of the respiratory yeast *Pichia stipitis* reveals a new mechanism for sensing of oxygen limitation in yeast. *Yeast* **20**, 39-51 (2003).

25. Passoth, V., Zimmermann, M. & Klinner, U. Peculiarities of the regulation of fermentation and respiration in the crabtree-negative, xylose-fermenting yeast *Pichia stipitis*. *Appl. Biochem. Biotechnol.* **57-58**, 201-212 (1996).

26. Klinner, U., Fluthgraf, S., Freese, S. & Passoth, V. Aerobic induction of respiro-fermentative growth by decreasing oxygen tensions in the respiratory yeast *Pichia stipitis*. *Appl Microbiol Biotechnol* **67**, 247-253 (2005).

27. Chapman, J., Putnam, N., Ho, I. & Rokhsar, D. (Joint Genome Institute, 2005).

28. Passoth, V., Hansen, M., Klinner, U. & Emeis, C.C. The electrophoretic banding pattern of the chromosomes of *Pichia stipitis* and *Candida shehatae*. *Curr Genet* **22**, 429-431 (1992).

29. Mayor, C. et al. VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046-1047 (2000).

30. Koonin, E.V. et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology* **5** (2004).

31. Dehal, P.S. & Boore, J.L. A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics* **7**, 201 (2006).

32. Dujon, B. et al. Genome evolution in yeasts. *Nature* **430**, 35-44 (2004).

33. Dyda, F. et al. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science* **266**, 1981-1986 (1994).

34. Xiong, Y. & Eickbush, T.H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *Embo J* **9**, 3353-3362 (1990).

35. Schmidt, R. Synteny: recent advances and future prospects. *Curr Opin Plant Biol* **3**, 97-102 (2000).

36. Fischer, G., Neuveglise, C., Durrens, P., Gaillardin, C. & Dujon, B. Evolution of gene order in the genomes of two related yeast species. *Genome Res* **11**, 2009-2019 (2001).

37. Nobre, A. & Lucas, C. in unpublished (NCBI, 2003).

38. Leandro, M.J., Goncalves, P. & Spencer-Martins, I. Two glucose/xylose transporter genes from the yeast *Candida intermedia*: first molecular characterization of a yeast xylose/H + symporter. *Biochem J* (2006).

39. Basaran, P., Hang, Y.D., Basaran, N. & Worobo, R.W. Cloning and heterologous expression of xylanase from *Pichia stipitis* in *Escherichia coli*. *Journal Of Applied Microbiology* **90**, 248-255 (2001).

40. Suh, S.O., McHugh, J.V. & Blackwell, M. Expansion of the *Candida tanzawaensis* yeast clade: 16 novel *Candida* species from basidiocarp-feeding beetles. *Int J Syst Evol Microbiol* **54**, 2409-2429 (2004).

41. Shi, N.Q., Cruz, J., Sherman, F. & Jeffries, T.W. SHAM-sensitive alternative respiration in the xylose-metabolizing yeast *Pichia stipitis*. *Yeast* **19**, 1203-1220 (2002).

42. Laplaza, J.M., Torres, B.R., Jin, Y.S. & Jeffries, T.W. *Sh ble* and *Cre* adapted for functional genomics and metabolic engineering of *Pichia stipitis*. *Enzyme Microb Technol* **38**, 741-747 (2006).

43. Chow, T.H., Sollitti, P. & Marmur, J. Structure of the multigene family of MAL loci in *Saccharomyces*. *Mol Gen Genet* **217**, 60-69 (1989).

44. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-467 (2005).

45. Neuveglise, C., Feldmann, H., Bon, E., Gaillardin, C. & Casaregola, S. Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts. *Genome Res* **12**, 930-943 (2002).

46. Daboussi, M.J. & Capy, P. Transposable elements in filamentous fungi. *Annu Rev Microbiol* **57**, 275-299 (2003).

47. Darling, A.C., Mau, B., Blattner, F.R. & Perna, N.T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394-1403 (2004).

48. Fischer, G., Rocha, E.P.C., Brunet, F., Vergassola, M. & Dujon, B. Highly variable rates of genome rearrangements between hemiascomycetous yeast Lineages. *PLoS Genet.* **2**, e32 (2006).

49. Sibirny, A.A. et al. Xylose and cellobiose fermentation to ethanol by the thermotolerant methylotrophic yeast *Hansenula polymorpha* and by xylose fermenting yeast *Pichia stipitis*. *Yeast* **20**, S219-S219 (2003).

50. Parekh, S.R., Parekh, R.S. & Wayman, M. Fermentation of xylose and cellobiose by *Pichia stipitis* and *Brettanomyces clausenii*. *Appl. Biochem. Biotechnol.* **18**, 325-338 (1988).

51. Parekh, S. & Wayman, M. Fermentation of cellobiose and wood sugars to ethanol by *Candida shehatae* and *Pichia stipitis*. *Biotechnol Lett* **8**, 597-600 (1986).

52. Bruinenberg, P.M., de Bot, P.H.M., van Dijken, J.P. & Scheffers, W.A. The role of redox balances in the anaerobic fermentation of xylose by yeast. *Eur J Appl Microbiol Biotechnol* **18**, 287-292 (1983).

53. Grotkjaer, T., Christakopoulos, P., Nielsen, J. & Olsson, L. Comparative metabolic network analysis of two xylose fermenting recombinant *Saccharomyces cerevisiae* strains. *Metab Eng* **7**, 437-444 (2005).

54. Zakian, V.A. Structure, function, and replication of *Saccharomyces cerevisiae* telomeres. *Annn Rev Genet* **30**, 141-172 (1996).

55. Krogan, N.J. et al. COMPASS, a histone H3 (lysine 4) methyltransferase required for telomeric silencing of gene expression. *J Biol Chem* **277**, 10753-10755 (2002).

56. Osborne, C.S. et al. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* **36**, 1065-1071 (2004).

57. Zhan, S., Horrocks, J. & Lukens, L.N. Islands of co-expressed neighbouring genes in *Arabidopsis thaliana* suggest higher-order chromosome domains. *Plant Journal* **45**, 347-357 (2006).

58. Bartlett, O. et al. in Transcription, Vol. 73 67-75 (Portland Press Ltd., London, England; 2006).

59. Naumov, G.I., Naumova, E.S. & Michels, C.A. Genetic variation of the repeated MAL loci in natural populations of *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*. *Genetics* **136**, 803-812 (1994).

60. Needleman, R.B. & Michels, C. Repeated family of genes controlling maltose fermentation in *Saccharomyces carlsbergensis*. *Mol Cell Biol* **3**, 796-802 (1983).

61. Vidgren, V., Ruohonen, L. & Londesborough, J. Characterization and functional analysis of the MAL and MPH loci for maltose utilization in some ale and lager yeast strains. *Appl Environ Microbiol* **71**, 7846-7857 (2005).

62. Sugita, T. & Nakase, T. Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus *Candida*. *Syst Appl Microbiol* **22**, 79-86 (1999).

63. Ueda, T., Suzuki, T., Yokogawa, T., Nishikawa, K. & Watanabe, K. Unique structure of new serine transfer-RNAs responsible for decoding leucine codon CUG in various *Candida* species and their putative ancestral transfer-RNA genes. *Biochimie* **76**, 1217-1222 (1994).

64. Santos, M.A.S., Keith, G. & Tuite, M.F. Nonstandard translational events in *Candida albicans* mediated by an unusual seryl-transfer RNA with A 5'-CAG-3' (Leucine) anticodon. *Embo Journal* **12**, 607-616 (1993).

65. Kaiser, C., Michaelis, S. & Mitchell, A. Methods in Yeast Genetics. (Cold Spring Harbor Laboratory Press, 1994).

66. Burke, D., Dawson, D. & Stearns, T. Methods in yeast genetics: a Cold Spring Harbor Laboratory course manual. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.; 2000).

67. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-194 (1998).

68. Smith, T.F. & Waterman, M.S. Overlapping genes and information theory. *J Theor Biol* **91**, 379-380 (1981).

69. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197 (1981).

70. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301-1310 (2002).

71. Gordon, D., Abajian, C. & Green, P. Consed: A graphical tool for sequence finishing. *Genome Res.* **8**, 195-202 (1998).

72. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res* **10**, 516-522 (2000).

73. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res* **10**, 547-548 (2000).

74. Zdobnov, E.M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848 (2001).

75. Mulder, N.J. et al. InterPro, progress and status in 2005. *Nucleic Acids Res* **33**, D201-205 (2005).

76. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

77. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**, D277-280 (2004).

78. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).

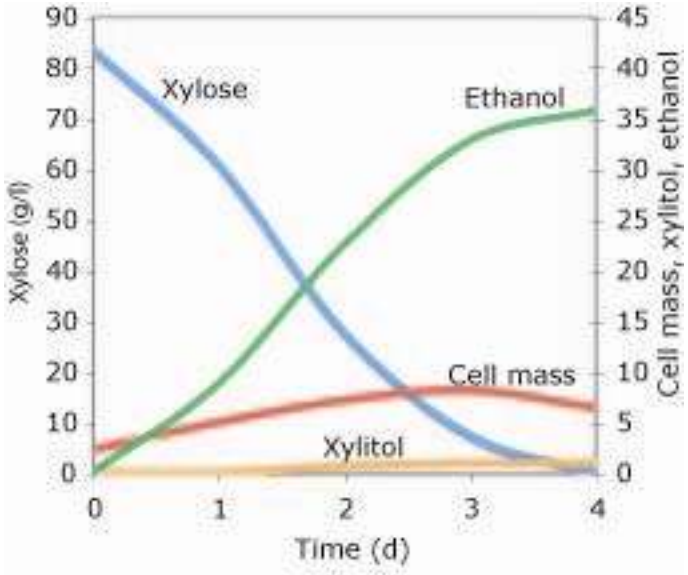


Figure 1 (Jeffries)

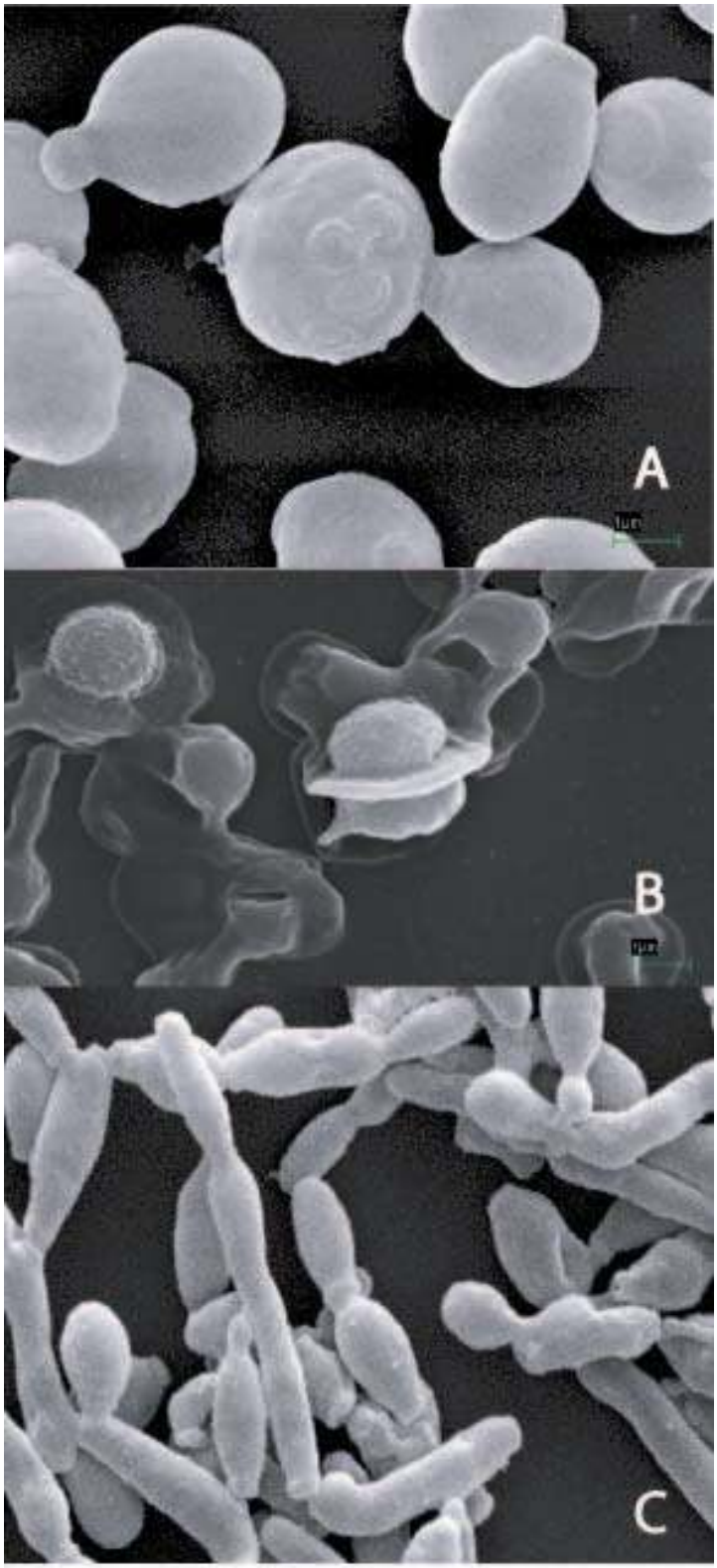


Figure 2 (Jeffries)

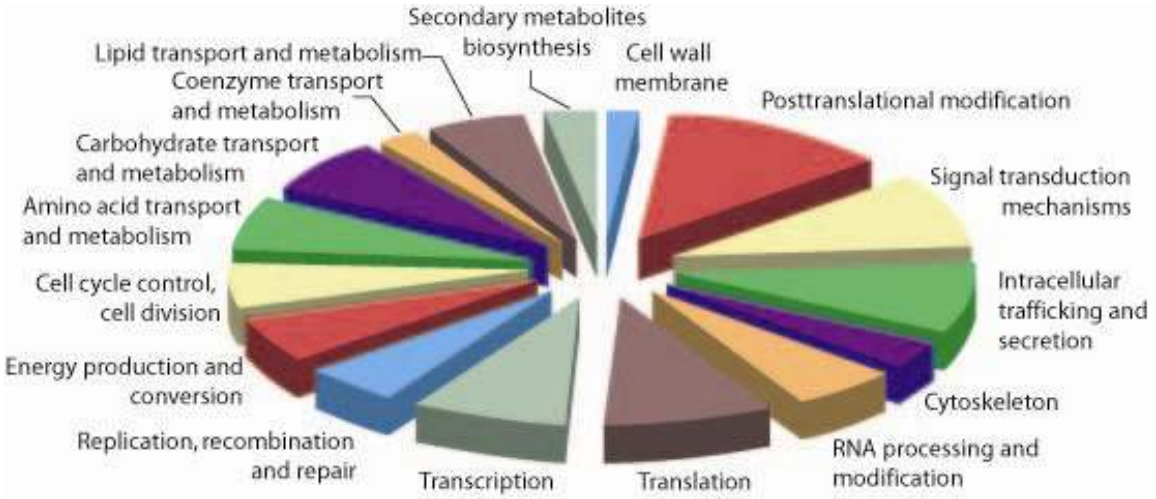


Figure 3 (Jeffries)

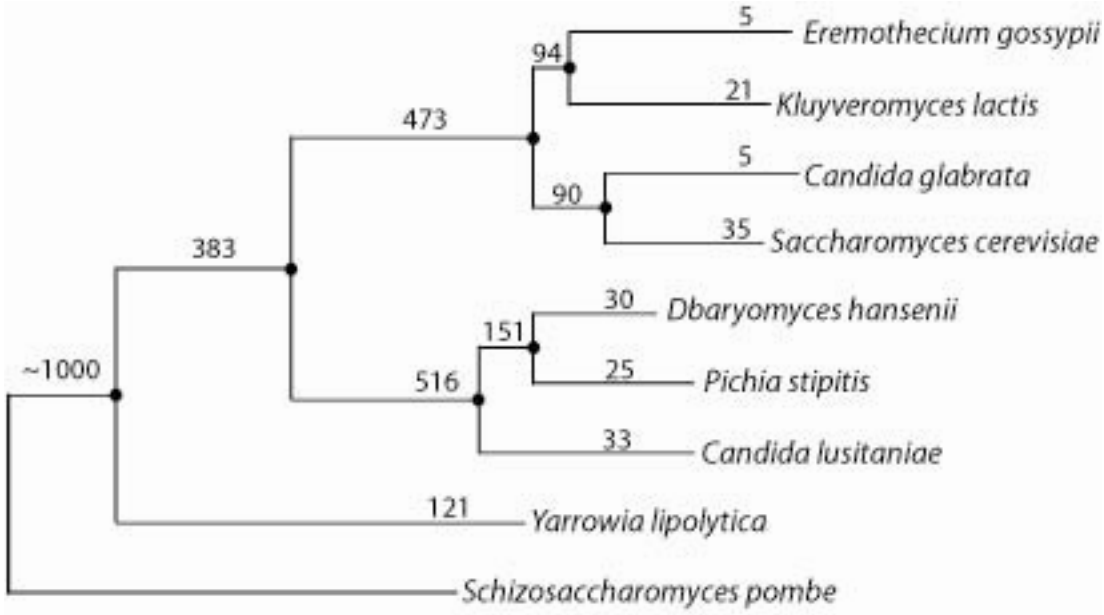


Figure 4 (Jeffries)

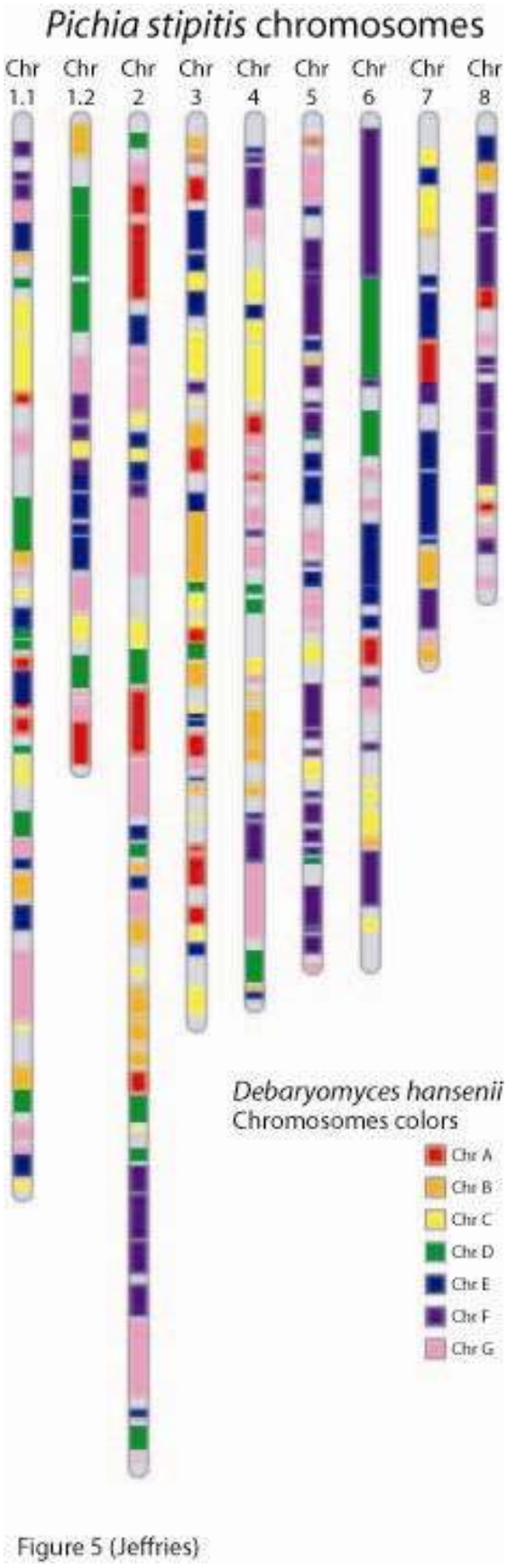


Figure 5 (Jeffries)

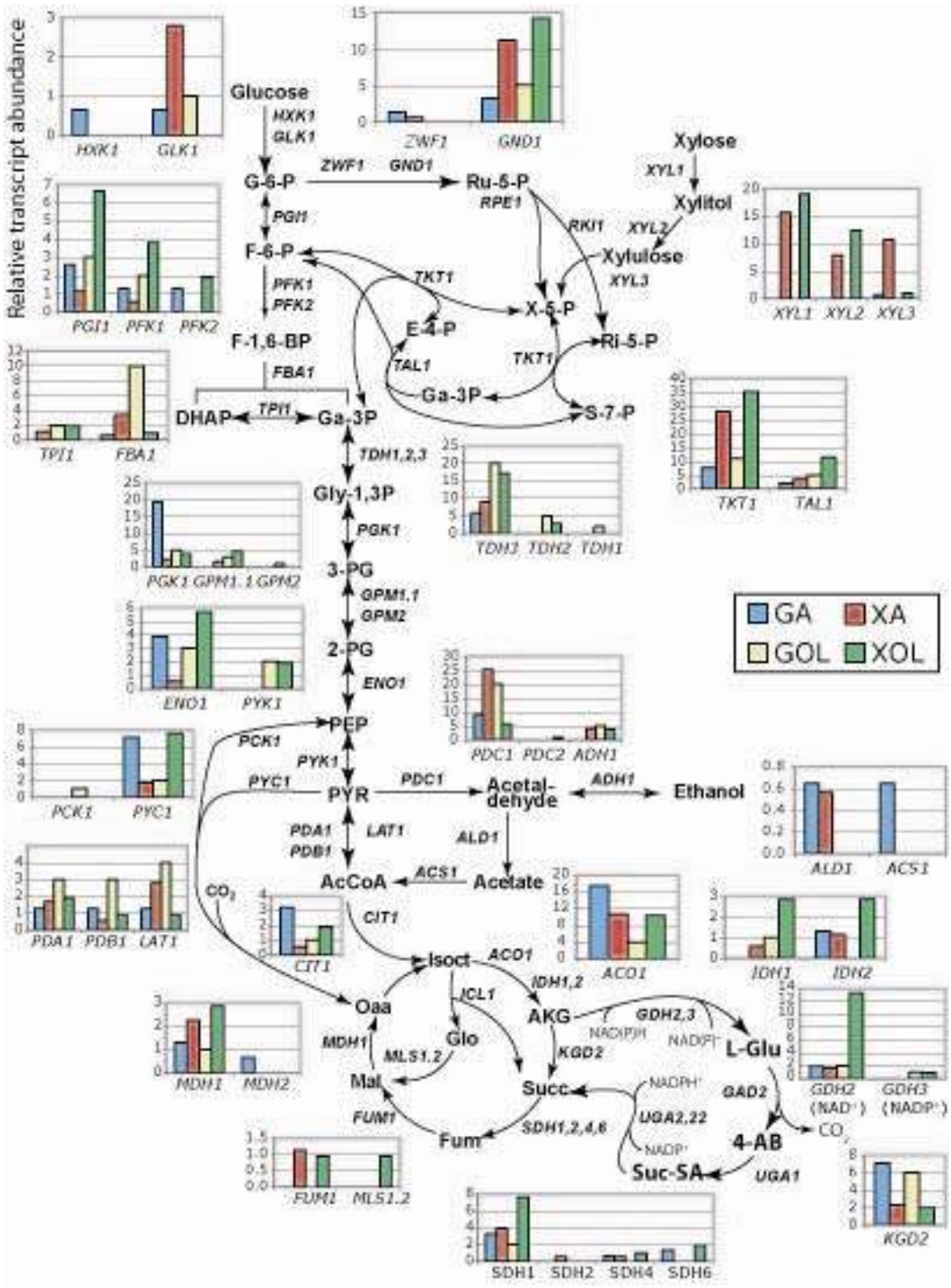
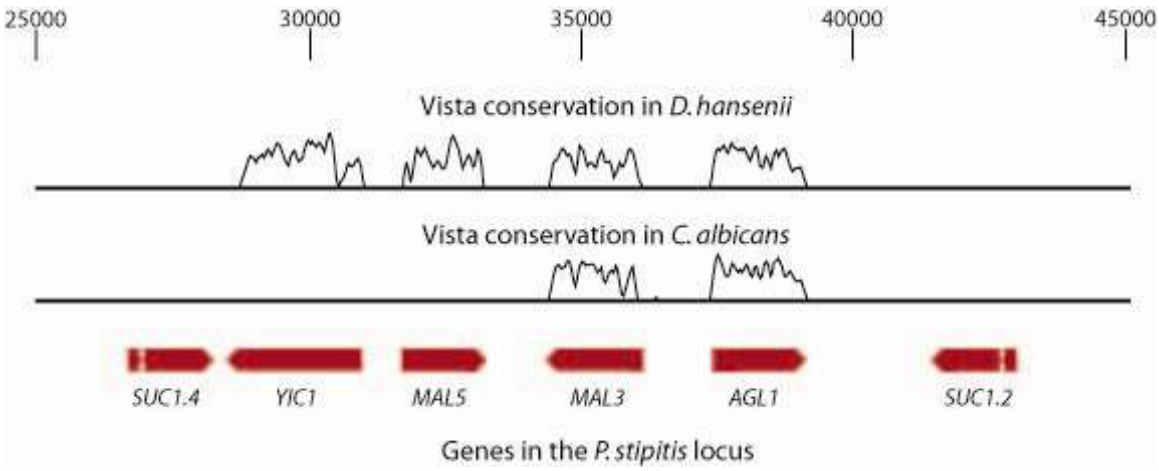


Figure 6 (Jeffries)


25000
30000
35000
40000
45000
Vista conservation in D. hansenii
Vista conservation in C. albicans
SUC1.4
YIC1
MAL5
MAL3
AGL1
SUC1.2
Genes in the P. stipitis locus

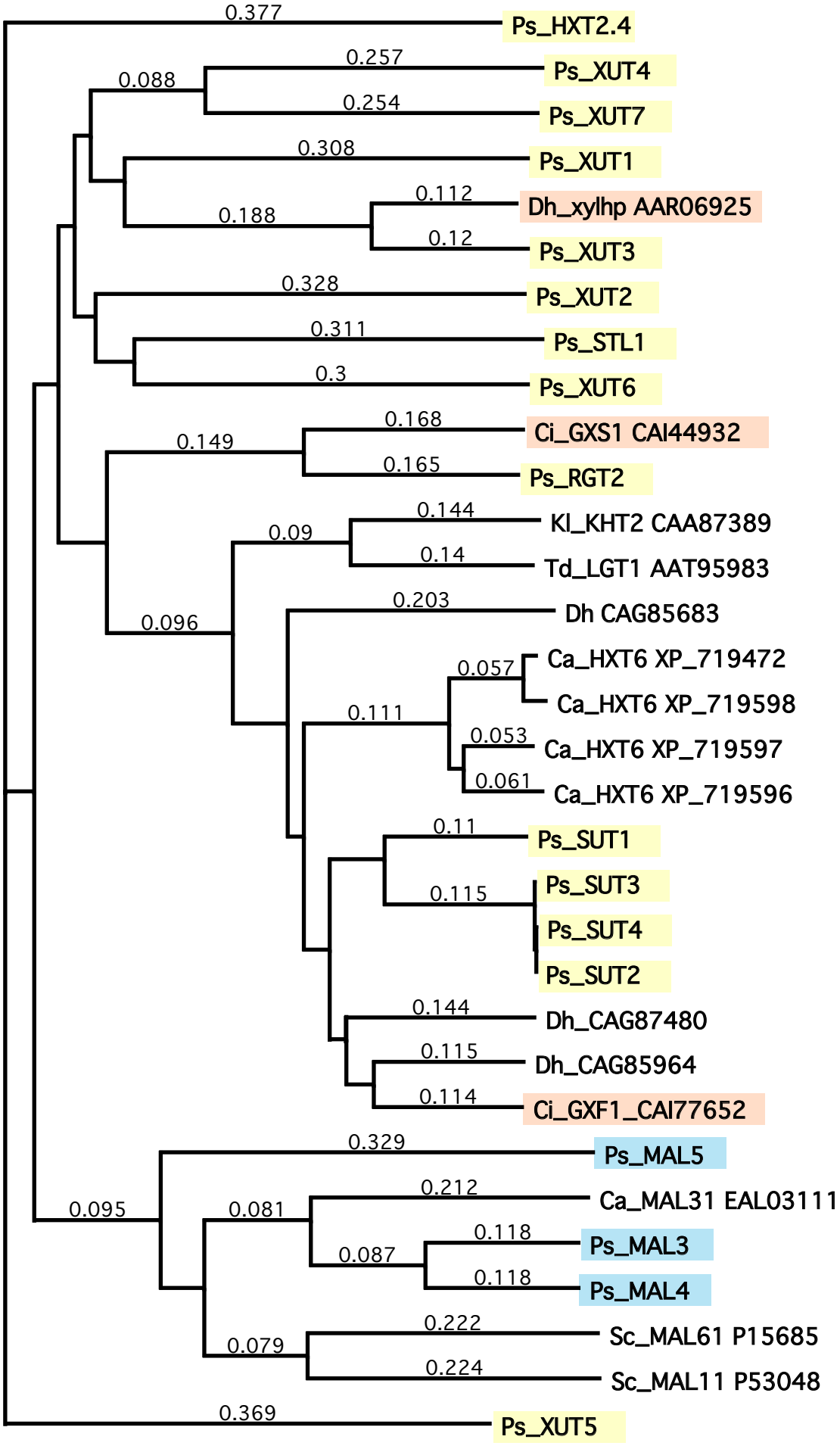


Figure Supplemental 1 (Jeffries)