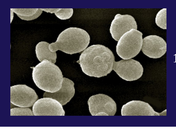
Comparative Analysis of the Complete *Pichia stipitis* Genome



Andrea Aerts[1*], Alan Kuo[1], Asaf Salamov[1], Paramvir Dehal[1], Robert P. Otillar[1], Jane Grimwood[2], Thomas W. Jeffries[3], Paul Richardson[1], Igor Grigoriev[1]
[1]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. [2]Stanford Human Genome Center, 975 California Ave, Palo Alto, CA 94304, USA. [3]USDA Forest Service, Forest Products Laboratory, Madison, WI 53726-2398, USA.
*Correspondence: ALAerts@lbl.gov

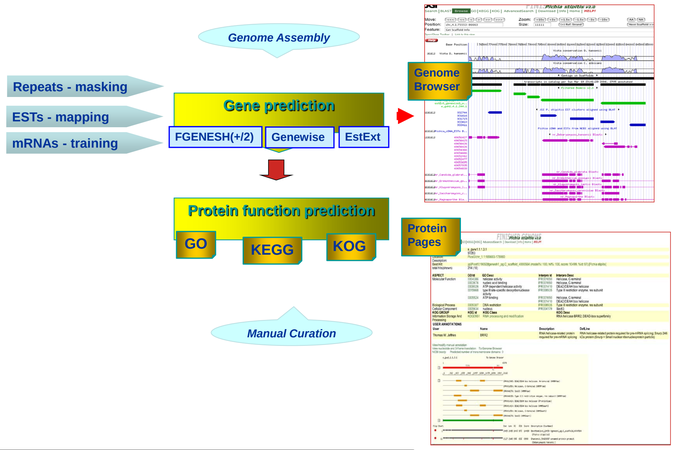
Abstract

The genome of *Pichia stipitis* is estimated to be approximately 15 million base pairs contained in 8 chromosomes. The genome assembly was annotated using the JGI Annotation Pipeline, which combines various gene prediction, annotation, and analysis tools. The *P. stipitis* genome portal can be found at www.jgi.doe.gov/pichia. Gene models and associated transcripts/proteins are predicted based on cDNA, protein homology and ab initio methods. A gene catalog set is chosen from candidate models based on homology and EST support for intron/exon boundary structure, ORFs and presence of UTR. Finally, each predicted model is analyzed for domain content/structure and functionally annotated. The release v2.0 includes a total of 5841 gene models supported by available EST and cDNA evidence and protein homology. Average gene, transcript and CDS lengths are 1.6kb, 1.5kb and 493 a.a., respectively. Average gene density is 56% with 4204 single exon genes. The genome size, number of genes and CDS lengths are comparable to the numbers found in other sequenced yeast genomes.

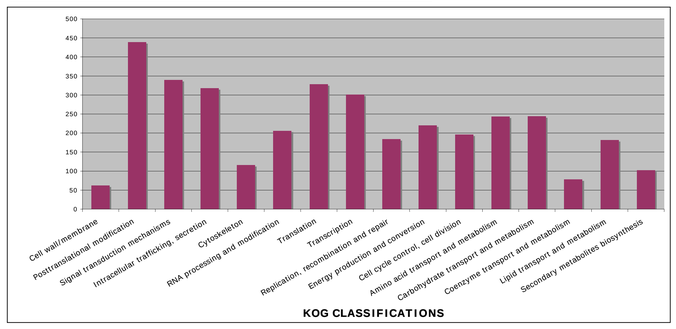
A set of 19,635 ESTs was sequenced and clustered at the JGI. 3839 (94%) of the EST clusters mapped to the genome. 2252 (39%) genes are supported by ESTs. An absolute majority of predicted genes are supported by protein homology including 4879 (84%) with strong homology in other fungi (alignment score > 1000). 4083 (70%) of all predicted genes have a predicted protein domain. Manual curation of this genome is ongoing using the JGI Genome Portal Tools. 2731 genes have been manually curated and there appears to be major differences between *P. stipitis* and other yeasts in oxidative phosphorylation, fatty acid metabolism and fatty acid synthesis.

Pichia stipitis is of fundamental biological interest and important from an applied perspective because it has the highest native capacity for xylose fermentation of any known microbe. We have compared the gene set of *P. stipitis* with the gene sets of five yeasts whose genomes have also been sequenced and assembled. Using comparative methods we have determined a core set of genes common to the six yeasts, as well as the set unique to *P. stipitis*.

Annotation Pipeline Schematic



Functional Classifications



Genome Statistics

Table 1. Sequencing and Assembly (before finishing)

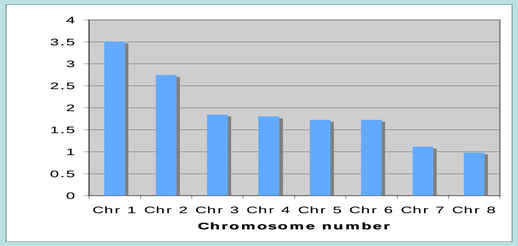
Total Reads	261,986
Coverage (sequence)	8.77 x
Coverage (clones)	54.7 x
Number of Contigs	341
Number of Scaffolds	93
Assembly Size	15.4 Mb
N50 Contigs	20 Kb
N50 Scaffolds	5 Mb

Table 2. Supporting Data

P.stipitis mRNAs & ESTs from GenBank	866
mRNAs & ESTs Aligned to Assembly	672 (66.1%)
EST Clusters	4085
EST Clusters Aligned to Assembly	3839 (94.0)%

Table 3. Gene Catalog Overview

Predicted Genes	5841
Coding region of genome	55.9%
Genes with Homology	5687 (97.4%)
Genes with Predicted Protein Domain	4083 (69.9%)
Genes with >= 90% Conservation with *Debaromyces hansenii* Genomic Sequence	489 (8.4%)
Genes with >= 10% Conservation with *Debaromyces hansenii* Genomic Sequence	3940 (67.5%)
Genes with 2 or more exons	1637 (28.0%)
Genes manually curated	2731 (46.8%)
Genes Supported by 2 Prediction Methods	2526 (43.2%)
Genes Supported by ESTs	2252 (38.6%)



The *Pichia stipitis* genome was finished at the Stanford Genome Center in Palo Alto, CA. The completed genome, which can be found at www.jgi.doe.gov/pichia, contains 8 high-quality chromosomes with 13 unresolved repeat regions. Chromosomes 1, 4 and 8 have one gap each.

Why Sequence *Pichia stipitis*?

Pichia stipitis Pignal (1967) is a predominantly haploid, heterothallic yeast related to *Candida shehatae* and other pentose metabolizing ascomycetous species. It belongs to a clade that uses an alternative nuclear genetic code in which CUG codes for serine rather than leucine. Its closest relatives are found in the intestines of passalid beetles, which in turn are found grazing on white-rotted hardwood. *P. stipitis* has the highest native capacity for xylose fermentation of any known microbe. Strains of *P. stipitis* are among the best xylose-fermenting yeasts in type culture collections. Fed batch cultures of *P. stipitis* produce up to 47 g/L of ethanol from xylose at 30° C under low aeration conditions. In addition to fermenting D-xylose to ethanol, *P. stipitis* can assimilate cellobiose and will oxidize the lignin-related compounds veratraldehyde and vanillin to their respective alcohols and acids. As such it is an ideal organism for lignocellulose bioconversion.

Comparative Analysis

We have compared the gene set of *P. stipitis* with the gene sets of five yeasts whose genomes have also been sequenced, assembled and reported in [1]Dujon, B. et al. Genome evolution in yeasts. *Nature* **430**, 35-44 (2004). Refer to Table 4.

Table 4. General Characteristics of the 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	1
C. glabrata	12.3	38.8	5283	65.0	41.0	493	4881	1
K. lactis	10.6	38.7	5329	71.6	40.1	461	4916	1
D. hansenii	12.2	36.3	6906	79.2	37.5	389	4190	1
Y. lipolytica	20.5	49.0	6703	46.3	52.9	476	6539	1

Table 5. Phyletic Patterns of Yeast Protein Families

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

Pattern	Families	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

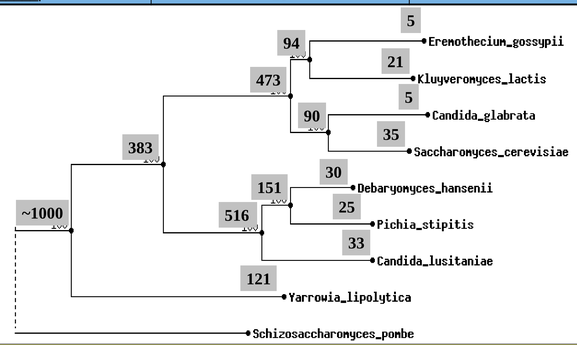


Figure 1. This data for this tree were generated using the PhIGs tool. The number at each node represents the unique lineage-specific gene families that arose at that particular node and are not found at any previous node. There are 25 *Pichia stipitis* specfic gene families found in the gene set. These 25 families include 72 individual genes, or 1.2% of the gene catalog.

Summary of Manual Curation

More than 2700 genes have been manually curated using the JGI Genome Portal Tools, and there appear to be major differences between *P. stipitis* and other yeasts in oxidative phosphorylation, fatty acid metabolism and fatty acid synthesis. The sequenced genome (CBS 6054) includes seven family 3 β-glucosidases, a family 10 endo xylanase, one β-mannanase, two exo-1,3 β-glucanases, five cinnamyl alcohol dehydrogenases and 57 transporters in the major facilitaor superfamily – including 5 putative xylose transporters. Aside from genes for assimilating a wide variety of lignocellulosic polymers, *P. stipitis* codes for several proteins that enable it to ferment xylose with the production of very little xylitol. These include four NADP-dependent alcohol dehyrogenases, two primary alcohol dehydrogenases and three pyruvate decarboxylases.

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.