Evolutionary History of Ascomyceteous Yeasts

Sajeet Haridas[1], Robert Riley[1], Asaf Salamov[1], Markus Göker[2], Hans-Peter Klenk[2], Cletus P. Kurtzman[3], Meredith Blackwell[4], Igor Grigoriev[1], Thomas W. Jeffries[5]

[1]DOE Joint Genome Institute. [2]Leibniz Institute DSMZ. [3]USDA-ARS. [4]Louisiana State University. [5]University of Wisconsin.

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

Abstract

Yeasts are important for many industrial and biotechnological processes and show remarkable diversity despite morphological similarities. We have sequenced the genomes of 16 ascomycete yeasts of taxonomic and industrial importance including members of Saccharomycotina and Taphrinomycotina. A comparison of these with several other previously published yeast genomes have added increased confidence to the phylogenetic positions of previously poorly placed species including *Saitoella complicata*, *Babjeviella inositovora* and *Metschnikowia bicuspidata*.

Phylogenetic analysis also showed that yeasts with alternative nuclear codon usage where CUG encodes serine instead of leucine are monophyletic within the Saccharomycotina. Most of the yeasts have compact genomes with a large fraction of single exon genes with *Lipomyces starkeyi* and the previously published *Pneumocystis jirovecii* being notable exceptions. Intron analysis suggests that early diverging species have more introns. We also observed a large number of unclassified lineage specific non-simple repeats in these genomes.

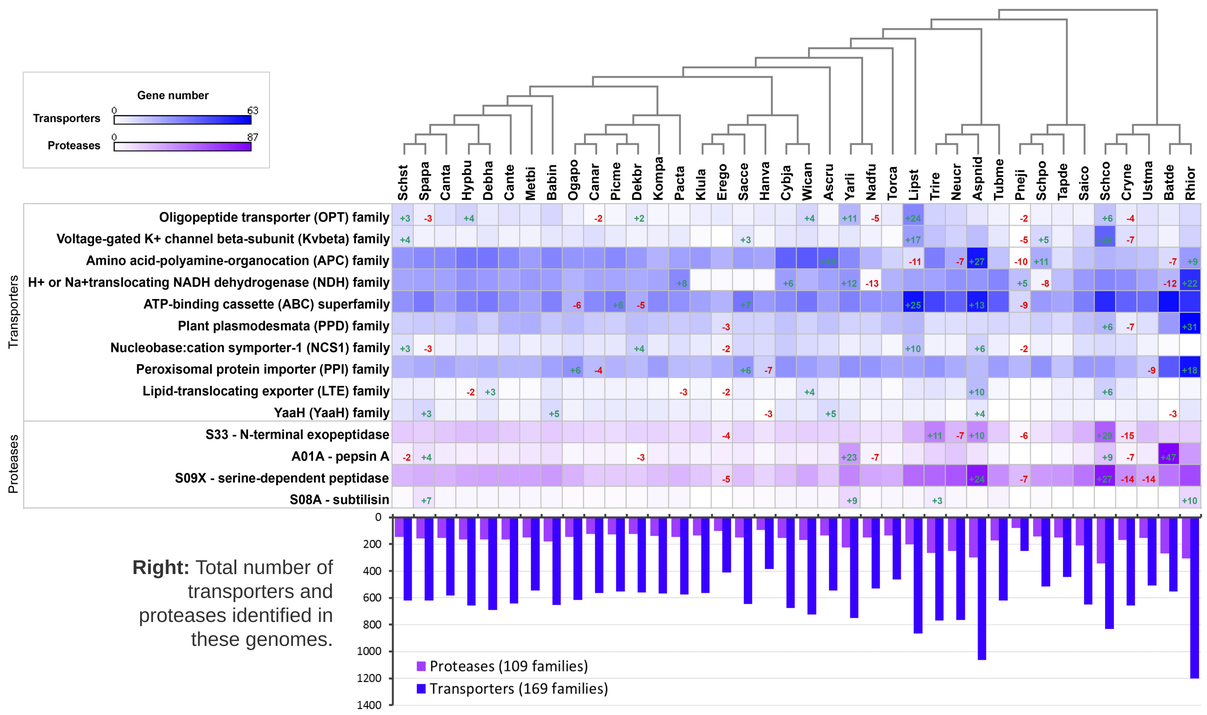
Significance

The largest fungal phylum, Ascomycota (ascomycetes), contains more than 60,000 described species and includes the budding yeasts (in Saccharomycotina) and fission yeasts (in Taphrinomycotina). Many of these yeasts have biotechnological, taxonomic and physiological interest.

We present the genomes of 16 newly sequenced yeasts along with the genomes of several other previously published fungal genomes. We are mining these genomes to elucidate the biochemical, physiological, biotechnological, and bioconversion potential of an entirely new group of yeasts, which would expand our knowledge of the phylogenetic relationships of taxa in understudied lineages. Many of these understudied taxa are likely to have novel genes with biotechnological value.

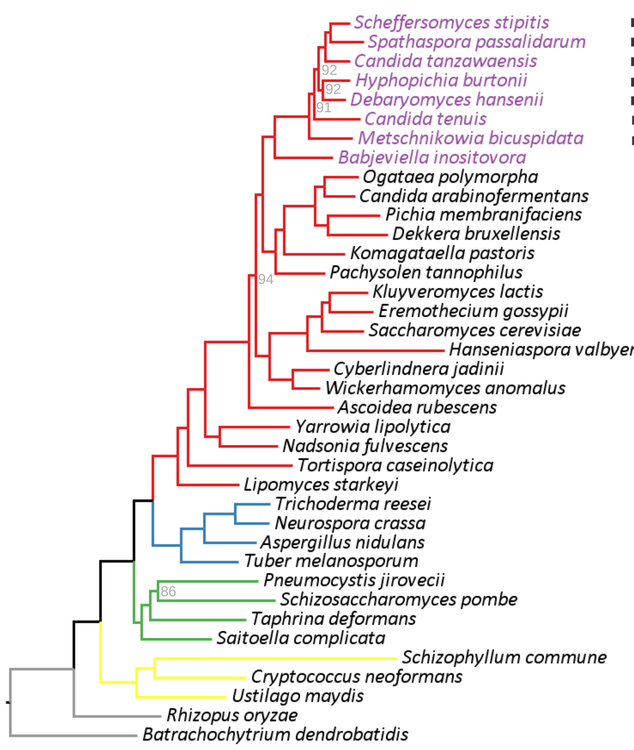
Conclusions

- The ascomycete yeasts are polyphyletic.
- Ascomycete yeasts tend to have small compact genomes with few repeats, but some, like *Lipomyces starkeyi,* have a higher number of repeats, many of which are lineage specific.
- The budding yeasts and fission yeasts are independent clades. The "CUG yeasts", where $tRNA_{CAG}$ encodes serine instead of leucine are budding yeasts and form a monophyletic clade within the Saccharomycotina.
- Among the genomes analyzed, early diverging organisms have more introns.
- Lifestyle-specific expansion and contraction of various gene families is evident. eg., the near absence of the NDH family of transporters in 'fermentation' yeasts.
- These 16 genomes along with over 300 other fungal genomes are available on MycoCosm (jgi.doe.gov/fungi).



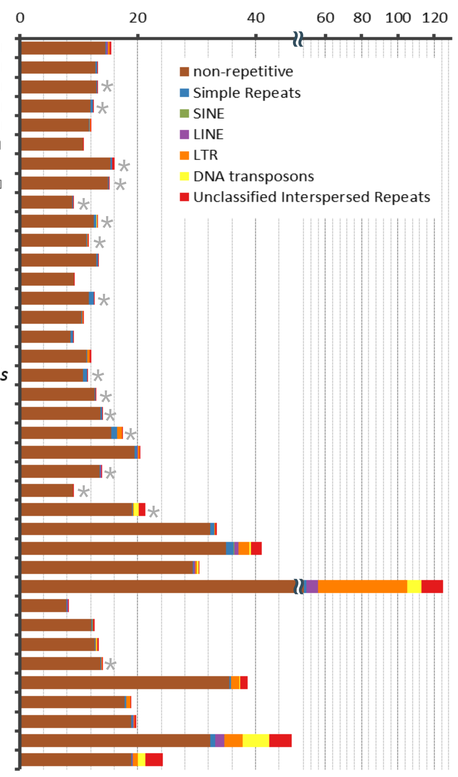
Right: Total number of transporters and proteases identified in these genomes.

Evolution of transporter and protease genes in ascomycete yeasts. Using the CAFE program and a family-wide significance threshold of 0.05, ten transporter families and four protease families were found to differ significantly from a random model of gene birth and death. Statistically significant gains and losses (those with a Viterbi p-value < 0.05) at each organism's most recent node of divergence are shown (gains green; losses red). The organism short names are the first three letters of the genus followed by the first two or three letters of the species names.



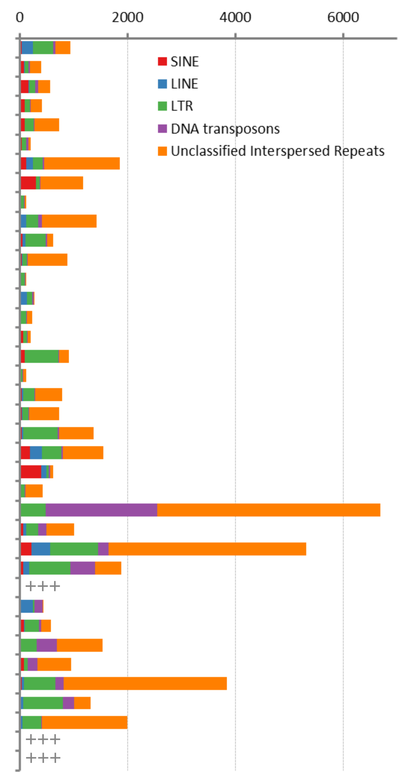
A phylogenetic tree of organisms used in this study. Phylogeny inferred using maximum-likelihood criteria from a matrix of concatenated alignments of amino-acid sequences from orthologous proteins. All genes present in at least four organisms were initially taken into account and then filtered for coverage and phylogenetic informativeness. Only maximum-likelihood bootstrap branch support values less than 100 are shown. The Saccharomycotina (red), Pezizomycotina (blue) and Taphrinomycotina (green) cluster into distinct clades. Basidiomycota (yellow) and non-Dikarya (grey) were used as outgroups.

An analysis of $tRNA_{CAG}$ in the Saccharomycotina showed that some of these had the Ser identity element (□) and the discriminator base (■), two characteristics that suggest that the $tRNA_{CAG}$ in these genomes code for serine rather than leucine. *Metschnikowia bicuspidata* and *Babjeviella inositovora* showed transitional patterns and appear at the base of this monophyletic "CUG clade" (purple text).

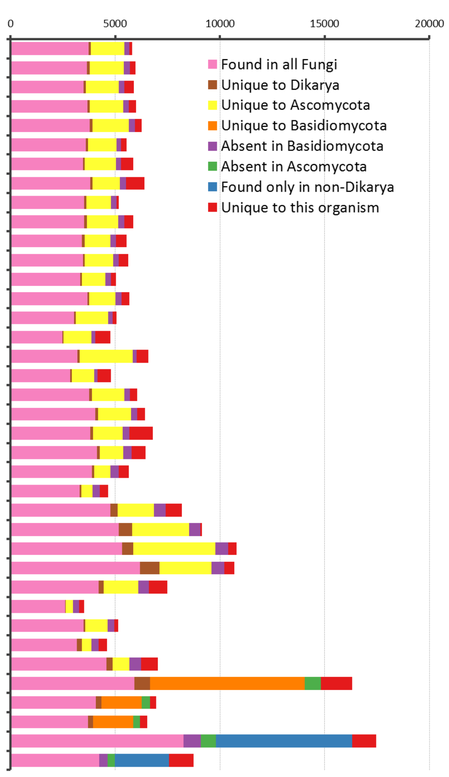


Genome sizes and repeat content. Yeast genomes are small with only a small fraction of the genome made up of repeats when compared to the Pezizomycotina (e.g., *Tuber melanosporum*).

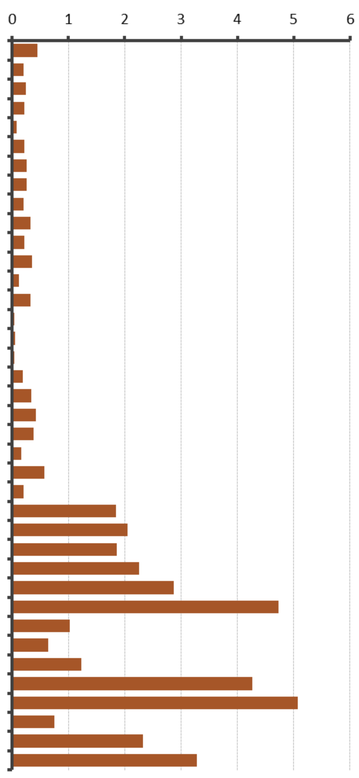
* represents newly sequenced yeast genomes.



Types and number of interspersed repeat elements. Repeatmasker was used to identify repeats in each genome using a custom database of repbase + repeatmodeler identified repeats. The number of unique interspersed repeat elements in each genome is represented. (+++) The data for *Tuber melanosporum* was omitted because this genome is about 10 times larger than the others in this study and had a large number of repeats. The two early diverging fungal species *Rhizopus oryzae* and *Batrachochytrium dendrobatidis* also had a large number of repeat elements and is not shown.



Gene Content. Gene clusters representing potential orthologous genes from 342 species across the fungal tree were used to identify genes belonging to the major taxonomic clades.



Average number of introns. Most yeast genomes are compact with a large fraction of single exon genes. The average number of introns per gene shows that early diverging species analyzed have more introns.

Acknowledgments:

MB acknowledges funding from NSF DEB-0417180 and 0732671.

Questions/Comments : sharidas@lbl.gov

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

DISCLAIMER

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