J Mol Evol (2005) 60:475–483
DOI: 10.1007/s00239-004-0218-1

JOURNAL OF MOLECULAR EVOLUTION

Comparison of the Yeast Proteome to Other Fungal Genomes to Find Core Fungal Genes

Tom Hsiang,[1] David L. Baillie[2]

[1] Department of Environmental Biology, University of Guelph, Guelph, Ontario, NIG 2W1, Canada
[2] Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, B.C., V5A 1S6, Canada

Received: 15 July 2004 / Accepted: 4 October 2004 [*Reviewing Editor:* Prof. David Guttman]

Abstract. The purpose of this research was to search for evolutionarily conserved fungal sequences to test the hypothesis that fungi have a set of core genes that are not found in other organisms, as these genes may indicate what makes fungi different from other organisms. By comparing 6355 predicted or known yeast (*Saccharomyces cerevisiae*) genes to the genomes of 13 other fungi using Standalone TBLASTN at an e-value $< 1E\text{-}5$, a list of 3340 yeast genes was obtained with homologs present in at least 12 of 14 fungal genomes. By comparing these common fungal genes to complete genomes of animals (*Fugu rubripes, Caenorhabditis elegans*), plants (*Arabidopsis thaliana, Oryza sativa*), and bacteria (*Agrobacterium tumefaciens, Xylella fastidiosa*), a list of common fungal genes with homologs in these plants, animals, and bacteria was produced (938 genes), as well as a list of exclusively fungal genes without homologs in these other genomes (60 genes). To ensure that the 60 genes were exclusively fungal, these were compared using TBLASTN to the major sequence databases at GenBank: NR (nonredundant), EST (expressed sequence tags), GSS (genome survey sequences), and HTGS (unfinished high-throughput genome sequences). This resulted in 17 yeast genes with homologs in other fungal genomes, but without known homologs in other organisms. These 17 core, fungal genes were not found to differ from other yeast genes in GC content or codon usage patterns. More intensive study is required of these 17 genes and other common fungal genes to discover unique features of fungi compared to other organisms.

Key words: Bioinformatics — BLAST — Fungi

Correspondence to: Tom Hsiang; *email:* thsiang@uoguelph.ca

Introduction

Fungal comparative genomics is a relatively new field and can provide insights in major areas of biology including drug discovery, phylogenetics, gene function, and evolution (reviewed in Hsiang and Baillie 2004; Piskur and Langkjaer 2004). With the use of powerful and versatile sequence comparison programs such as BLAST (basic local alignment search tool; Altschul et al. 1997), nucleotide and amino acid sequences can be compared, and the identification of homologous sequences allows discovery of core genes shared among organisms. Through gene silencing or gene expression studies, essential genes can then be identified among the core genes.

Eukaryotes share a set of genes that was derived from the common eukaryotic ancestor and conserved within existing lineages. Aside from this core set, every organism should have genes that are exclusive at various taxonomic levels from species to kingdoms, which may help make that taxon different from sister taxa (Chervitz et al. 1998; Rubin et al. 2000). The core set includes common housekeeping genes, while the taxon-specific genes can involve functions which are exclusive to the particular taxon (Mata and Bahler

Table 1. Nonfungal genomes used in this study, with file size, file date, and source

Species name	Taxon	File date	File size	Genome source
Agrobacterium tumefaciens	Bacterium	2001/09/27	5.8 Mb	Univ. Washington (Wood et al. 2001) www.ncbi.nlm.nih.gov
Arabidopsis thaliana	Plant	2000/12/14	120.8 Mb	TIGR (Arabidopsis Genome Initiative 2000)
Caenorhabditis elegans	Animal	2003/06/30	101.5 Mb	Sanger Institute (Harris et al. 2003) ftp.wormbase.org/pub/wormbase/DNA_D
Fugu rubripes	Animal	2003/07/22	336.6 Mb	DOE Joint Genome Institute (unpublished) genome.jgi-psf.org/fugu6
Oryza sativa	Plant	2001/12/27	371.5 Gb	Beijing Genomics Inst. (Yu et al. 2002) 210.83.138.53/rice/download.php
Xylella fastidiosa	Bacterium	2000/03/24	2.8 Mb	ONSA, Brazil (unpublished) aeg.lbi.ic.unicamp.br/xf

2003; Boffelli et al. 2004). Most of the genes that are essential for an organism are also likely to be found in the conserved set of genes (Decottignies et al. 2003).

Studies in comparative genomics utilize data generated on a massive scale, and the amount of data is increasing exponentially. With funding provided from government sources, the Whitehead Institute for Biomedical Research (www-genome.wi.mit.edu/annotation/fungi/fgi/history.html) associated with the Massachusetts Institute of Technology, the Joint Genome Institute (http://www.jgi.doe.gov/) operated by the University of California for the U.S. Department of Energy, and the Stanford Genome Technology Center (http://www-sequence.stanford.edu/) deserve credit for releasing most of the fungal genomes now publicly available. The small size of fungal genomes, averaging less than 30 Mb for publicly accessible genomes, is one reason that fungi represent the largest number of complete or almost-complete genomes sequenced among eukaryotes. As currently described, true fungi include ascomycota, basidiomycota, chytridiomycota, and zygomycota. The former two are commonly referred to as higher fungi, while the latter two are called lower fungi. Although still commonly placed with lower fungi, oomycetes such as *Phytophthora* species are more closely related to protists than to true fungi.

The set of genes common to yeasts and higher fungi should include ones present in fungi prior to the divergence of ascomycota from basidiomycota, which has been estimated to be 1.2 billion years ago (Heckman et al. 2001; Hedges and Kumar 2003) or 550 million years ago (Berbee and Taylor 2001). This set of common fungal genes should comprise ones that are essential housekeeping genes homologous with those in other organisms such as plants and animals, and these genes may have been present since the origin of these major taxa estimated at 1.6 billion years ago (Hedges and Kumar 2003). An even older subset would be those with homologs in bacteria which would have been present since the divergence of eukaryotes from prokaryotes estimated at 2.7 billion years ago (Hedges and Kumar, 2003). In addition to fungal genes with homologs in other organisms, there may be a subset present only in fungi, which have been lost from or never developed in other lineages. The purpose of this research was to test the hypotheses that fungi have a set of evolutionarily conserved core genes that are not found in other organisms and to see whether these genes share characteristics different from other yeast genes.

Materials and Methods

The *S. cerevisiae* genome and proteome were downloaded from the *Saccharomyces* Genome Database (ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence). The files were dated 23 May 2003, and the genomic DNA file size was 12.3 Mb, while the proteome file contained 6355 amino acid sequences. The genomes of six nonfungal organisms including two plants (*Arabidopsis thaliana* and *Oryza sativa*), two animals (*Caenorhabditis elegans* and *Fugu rubripes*), and two proteobacteria (*Agrobacterium tumefasciens* and *Xylella fastidiosa*) were also obtained (Table 1). A search was made for completely sequenced and almost completely sequenced fungal genomes publicly available on the Internet up to March 2004 (Table 2).

Standalone BLAST version 2.2.6 (Altschul et al. 1997) was set up on a Linux system, and a BLAST database was generated for each genome. Every yeast protein sequence was compared separately against each of the 6 nonfungal genomes (Table 1) and the 13 fungal genomes (Table 2) using Standalone TBLASTN. This program translates nucleotide sequences in three forward and three reverse reading frames and then compares the query protein sequence to these predicted protein sequences from the database to find the best match. A score is calculated for each match, the significance of which is represented by an e-value (expect value). The e-value refers to "the number of hits one can expect to see just by chance when searching a database of a particular size" (www.ncbi.nlm.nm.gov/BLAST/blast_FAQs.shtml). All BLAST analyses were run with default parameters in Standalone BLAST version 2.2.6. For most analyses, the threshold e-value of 1E-5 was chosen to indicate homology, and no corrections for database size were made since the fungal genomes did not differ greatly in size. The BLAST output files were parsed using PERL scripts (available at www.uoguelph.ca/~thsiang/est/ or upon request to the senior author). The scripts extracted the following data from BLAST output: query sequence name, query sequence length, top match, length of top match, e-value for top match, number of matches, and a list of the next five top matches. The PERL scripts generated a single line of output from each BLAST output file. For each data set, the single lines were combined into a single file, and the file

Table 2. Fungal genomes used in this study, with file size, file data, and source

Species name	File date	File size	Genome source
Aspergillus fumigatus[a]	2003/04/10	29.9 Mb	Sanger Institute (unpublished) ftp.sanger.ac.uk/pub/pathogens/A_fumigatus
Aspergillus nidulans[a]	2003/03/7	30.5 Mb	Whitehead Institute (unpublished) www.broad.mit.edu/annotation/fungi/aspergillus/
Candida albicans[a]	2002/05/24	15.2 Mb	Stanford Genome Tech. Center (Tzung et al. 2001)
Coprinus cinereus	2003/07/30	36.8 Mb	Whitehead Institute (unpublished) www.broad.mit.edu/annotation/fungi/coprinus_cin
Cryptococcus neoformans D	2003/03/31	18.8 Mb	Stanford Genome Tech. Center (unpublished) www-sequence.stanford.edu/group/C.neoformans
Fusarium graminearum[a]	2003/03/11	36.6 Mb	Whitehead Institute (unpublished) www.broad.mit.edu/annotation/fungi/fusarium/
Magnaporthe grisea[a]	2002/09/17	39.3 Mb	Whitehead Institute (unpublished) www.broad.mit.edu/annotation/fungi/magnaporthe
Neurospora crassa[a]	2002/06/30	39.0 Mb	Whitehead Institute (Galagan et al. 2003) www.broad.mit.edu/annotation/fungi/neurospora/
Phakopsora pachyrhizi	2004/01/21	1.33 Gb	DOE Joint Genome Institute (unpublished) ftp.jgi-psf.org/pub/JGI_data/Phakopsora_pachyrhi
Phanerochaete chrysosporium	2002/02/16	30.6 Mb	DOE Joint Genome Institute (Martinez et al. 2004)
Podospora anserina[a]	2004/01/23	36.4 Mb	Université de Paris (unpublished) podospora.igmors.u-psud.fr/download.html
Saccharomyces cerevisiae[a]	2003/05/23	12.3 Mb	SGD, Stanford (Goffeau et al. 1996; Mewes et al. 1997)
Trichoderma reesei[a]	2003/07/18	35.2 Mb	DOE Joint Genome Institute (unpublished) ftp.jgi-psf.org/pub/JGI_data/Treesei/v1.0
Ustilago maydis	2003/07/29	20.1 Mb	Whitehead Institute (unpublished) www.broad.mit.edu/annotation/fungi/ustilago_ma

[a]These species are placed in the ascomycota, while the other five are placed in the basidiomycota.

imported into a spreadsheet program. The use of commas for delimiters as specified in the PERL script allowed for the spreadsheet file to be parsed within the spreadsheet program into columns of relevant data listed above.

In addition to Standalone BLAST, further analysis of select yeast genes was also carried out with NETBLAST (Blastcl3.exe version 2.2.6) to access various GenBank databases: NR (nonredundant), EST (expressed sequence tags), GSS (genome survey sequences), and HTGS (unfinished high-throughput genome sequences). These TBLASTN analyses at e-value $\leq$ 1E-5 allowed assessment of whether the yeast genes which had no match with any of the six nonfungal genomes were indeed exclusively fungal.

Further study was made of the core fungal genes of unknown function by generating sequence alignments with fungal homologs using CLUSTAL X (Thompson et al. 1997) and selecting the most conserved portions for further analyses. These 20-aa to 140-aa conserved portions were compared to the GenBank NR protein database by searching for short exact matches with a high e-value of 20. For the core fungal genes, GC content was calculated by hand, and codon usage patterns were calculated using the web program available at www.kazusa.or.jp/codon. These were compared to the GC content and the codon usage pattern of the entire yeast genome.

Results and Discussion

Yeast Genomic Database

The yeast proteome dataset contained 6355 predicted or known yeast genes, of which 19 were mitochondrial genes (starting with "Q"). None of these mitochondrial genes were later identified as being part of the common fungal set of genes since several of the fungal genomic databases did not include mitochondrial genes. All the fungal genomic datasets used here except for yeast were reportedly incomplete, and even genomic datasets reported as finished are rarely complete particularly for eukaryotes (Mardis et al. 2002). Furthermore, certain types of sequences such as extended repeats, telomeres, and rRNA clusters are typically excluded for shotgun sequencing of eukaryotes (Martinez et al. 2004).

Estimates of the number of genes in *S. cerevisiae* range from 4800 to 6400 (Kellis et al. 2003). By comparing the genome of *S. cerevisiae* with the genomes of three other *Saccharomyces* species, and examining regulatory motifs and analyzing conservation of predicted gene sequences, Kellis et al. (2003) proposed that the proteome of *S. cerevisiae* could be reduced by approximately 500 previously annotated *S. cerevisiae* genes. We examined these yeast genes suggested for removal and found six with homologs in 12 other fungal genomes at e-values $\leq$ 1E-5: homologs of YNL089C could also be found in the two animal genomes with e-values of 1E-7 and 1E-20; homologs of YGL235W could be found in the bacterial genomes with e-values of 1E-14 and 1E-10; homologs of YLR379 W could be found in the plant genomes with e-values of 1E-11 and 1E-42; homologs

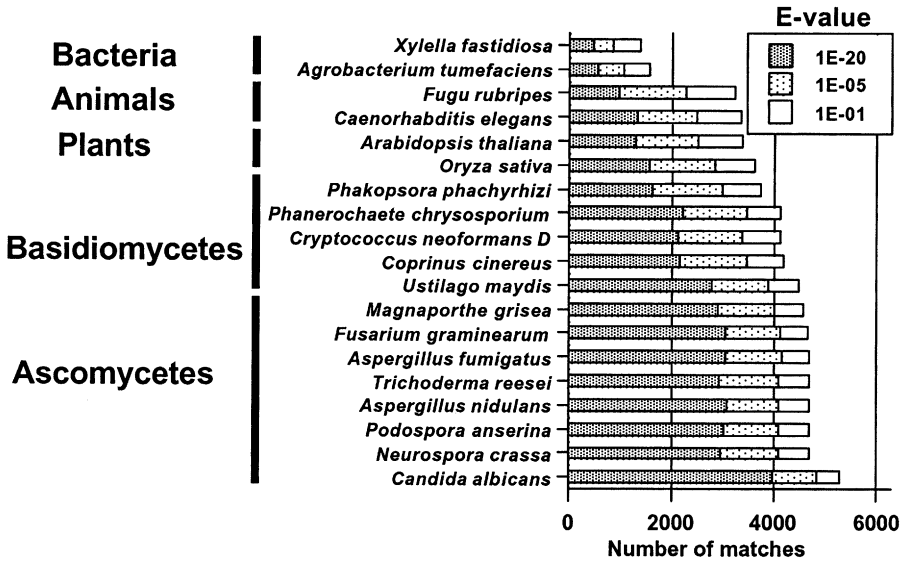

Fig. 1. Bar graph showing the number of matches at increasing e-values for *Saccharomyces cerevisiae* putative genes (6355) against BLAST databases of individual genomes of bacteria, animals, plants, and other fungi.

of YLR076C and YDL228C were present in both animal and plant genomes with e-values ranging from 1E-8 to 1E-60, as well as being present in 5 other fungal genomes with e-values less than 1E-20; and, finally, homologs of YAL004W were present in all 13 other fungal genomes, with 12 showing e-values less than 1E-20, as well as homologs present in plant, animal and bacterial genomes. These genes should be reconsidered for retention in the yeast proteome.

When we compared yeast genes with the 13 fungal and 6 nonfungal genomes separately, the number of matches between the yeast proteome and the various genomes was generally reflective of the phylogenetic relationships of *S. cerevisiae* to the other organisms: the other ascomycetes showed the highest number of matches, followed by basidiomycetes, plants, animals, and bacteria (Fig. 1). Of 6355 putative yeast genes, ascomycetes averaged 4181 homologs; basidiomycetes, 3433 homologs; plants, 2497 homologs; animals, 2374 homologs; and bacteria, 958 homologs at e-values $\leq$ 1E-5. Among ascomycetes, *C. albicans* was found to have the highest number of yeast homologs, and the basidiomycete *U. maydis* had almost as many yeast homologs as some of the ascomycetes. The number of homologs shared between organisms is not an absolute indicator of their phylogenetic relationship, since an analysis of the gene sequences would be required to reveal the finer details of the relationship. Furthermore, some consider the gene set to be a phenetic character which can undergo convergence through selective pressures (Dutilh et al. 2004).

Common Fungal Genes

In addition to yeast, 13 other fungal genomes were used in this study (Table 2). Of the total of 6355 yeast

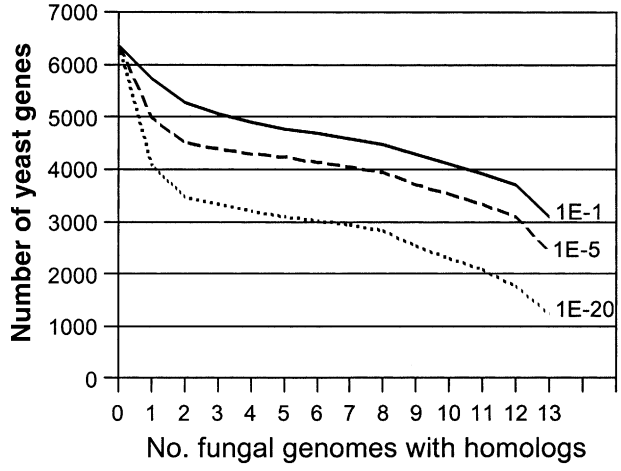

Fig. 2. The number of yeast genes with homologs in the 13 other fungal genomes at threshold e-values of 1E-1, 1E-5, and 1E-20.

genes, 3340 had homologs in at least 11 of 13 other fungal genomes at e-values $\leq$ 1E-5 (Fig. 2). At e-values of 1E-1 or 1E-20, the number with homologs was, respectively, 3907 or 2084 (Fig. 2). The criterion was set to 11 of 13, since the fungal genomes were reported to be incompletely sequenced. This set of yeast genes common to fungi represented just over half of the putative yeast genes, and assuming that horizontal transfer of genes was rare, their origins hypothetically predate the divergence of basidiomycetes and ascomycetes. If only ascomycetes were considered, the number of yeast genes with homologs among seven of eight other ascomycetes was 3879 at e-values $\leq$ 1E-5. The difference between these two numbers, 539, may represent the genes which developed after the basidiomycete/ascomycete divergence and before the divergence of the lineage of *S. cerevisiae* from that of the filamentous ascomycetes.

Among these 3340 yeast genes with homologs in other fungi, 772 (23%) gave nonviable phenotypes in deletion studies (*Saccharomyces* Genome Database). Among the remaining 3015 yeast genes, 342 (11.3%) gave nonviable phenotypes. This indicated that the common fungal sequences comprised a far greater proportion of essential genes than sequences which were not part of the conserved fungal set, and this was previously observed for *Schizosaccharomyces pombe* (Decottignies et al. 2003).

Eukaryotic Homologs

The 3340 yeast genes with homologs in other fungi, hereafter referred to as common fungal genes, were compared to the genomes of four nonfungal eukaryotic taxa. The plants *O. sativa* and *A. thaliana* yielded a total of 3080 matches, while the animals *F. rubripes* and *C. elegans* had 2907 matches (Fig. 3). Some researchers have reported that fungi are phylogenetically closer to animals than to plants using certain morphological and molecular characteristics (Baldauf 2003), while others contend that the major eukaryotic taxa arose at the same time (Philippe et al. 2000). The simple analysis presented here of comparing the number of homologs at an e-value of 1E-5 does not show support for either hypothesis presented above, since the number of matches to plants was slightly but significantly greater (chi-squared test at $p = 0.05$) than the number of matches to animals. Furthermore, the number of yeast genes which had homologs in either of the two plants but not in the animals was 254, while the number in either of the two animals but not in the plants was 151 (Fig. 4). In contrast, Martinez et al. (2004) found that the proteome of the basidiomycete *Phanerochaete chrysosporium* had 1279 predicted genes that had homologs in animals but not plants and 490 predicted genes that had homologs in plants but not animals. This could reflect a difference in the ascomycetous yeast genes compared to the basidiomycetous *P. chrysosporium* genes.

Among these 3340 common fungal genes, 2723 could also be found in an animal and a plant genome; these can be regarded as an eukaryotic core for fungal genes which should predate the divergence of these eukaryotic kingdoms estimated to be at least 1.6 billion years old (Hedges and Kumar 2003). Koonin et al. (2004) analyzed the protein sets from seven different eukaryotes (*C. elegans, Drosophila melanogaster, Homo sapiens, A. thaliana, S. cerevisiae, Schizosaccharomyces pombe*, and *Encephalitozoon cuniculi*) to construct 5873 clusters of predicted orthologs. About 40% or around 2350 were present in six or seven genomes, and these were considered to be the minimal set of essential eukaryotic genes. Jones et al. (2004) found that 3027 of the 6419 predicted

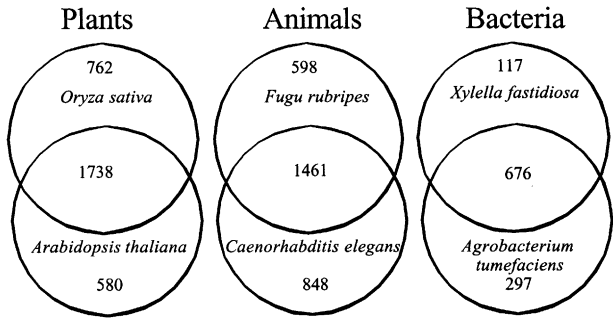

Fig. 3. Venn diagrams of the number of common putative fungal genes (3340) with homologs in plants, animals, and bacteria showing the numbers exclusive and common for each pairing at an e-value ≤ 1E-5.

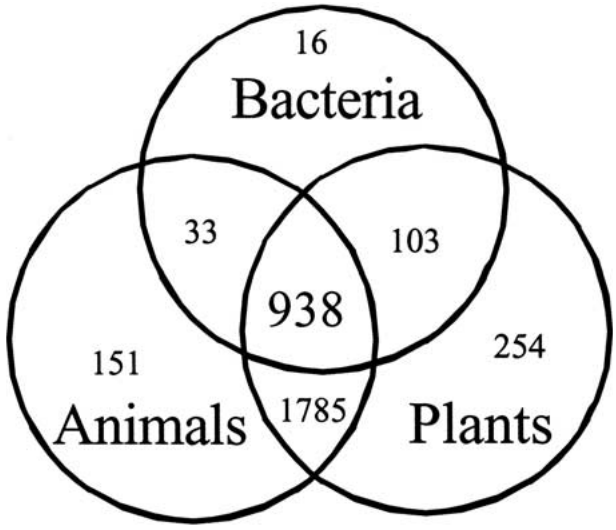

Fig. 4. Venn diagram of the number of common putative fungal genes (3340) with homologs in plants, animals, and bacteria showing the numbers exclusive and common for the different taxa at e-values ≤ 1E-5.

genes in *Candida albicans* had homologs in *S. pombe, S. cerevisiae*, and humans using BLASTP with an e-value ≤ 1E-8. The results of these three studies show similar ratios, with 43, 40, and 47% of the total genes common to eukaryotes of different kingdoms, respectively, for the current study, Koonin et al. (2004), and Jones et al. (2004).

Although just over 40% of the total genes in an eukaryotic genome may be shared with other eukaryotes of different kingdoms, these are not all essential genes. Winzeler et al. (1999) created *S. cerevisiae* deletion mutants for 2026 ORFs and found that 17% were essential for viability on rich media, and 40% showed some growth defect. Decottignies et al. (2003) examined 100 deletion mutants in *S. pombe* and concluded that 17.5% of the genes were essential. Giaever et al. (2002) created deletion mutants for 5916 *S. cerevisiae* genes and found that 18.7% were essential for growth on rich glucose

Table 3. The purposed functions from the *Saccharomyces* Genome Database (SGD) of 17 putative yeast genes that have homologs in at least 11 of 13 other fungal genomes tested, but without matches in 6 nonfungal genomes or in GenBank NR, EST, GSS, or HTGS

Gene	Biological process	Molecular function	Absent[d]	e-value[e]	Copy number[f]
YBL095W	Unknown	Unknown	Pc, P_P	1E-17	Single
YDL127W	Cell cycle	Cyclin-dependent protein kinase	Tr	1E-24	Multiple
YGR213C[a,b]	Unknown	Unknown	Cc, Pp	1E-25	Multiple
YGR215W	Protein biosynthesis	Ribosome structural constituent	Cn, Pp	1E-8	Single
YHL047C	Iron ion homeostasis	Siderochrome iron transporter	Pp	1E-58	Multiple
YHR194W	Mitochondrion organization/biogenesis	Unknown		1E-93	Multiple
YIL085C	Cell wall organization and biogenesis	Mannosyltransferase		1E-58	Multiple
YIL099W	Sporulation (sensu *Saccharomyces*)	Glucan 1,4-α glucosidase	Pp	1E-35	Single
YIR007W[a]	Unknown	Unknown		1E-115	Single
YJL087C[a,c]	tRNA splicing	tRNA ligase (ATP)		1E-95	Single
YJL127C	Chromatin modeling	Histone acetyltransferase	Cn, Pc	1E-38	Single
YLR203C	Protein biosynthesis	Unknown		1E-57	Single
YOR099W[a,b]	O-linked glycosylation	α-1,2-Mannosyltransferase		1E-96	Multiple
YOR132W	Endosome to Golgi transport	Unknown	Pp	1E-45	Single
YOR359W[a]	Protein-vacuolar targeting	Intracellular transporter	Cn, Pc	1E-10	Single
YOR365C	Unknown	Unknown		1E-61	Multiple
YPL105C[a]	Unknown	Unknown	Pp	1E-8	Multiple

[a]These yeast genes are mentioned in U.S. patents dealing with antifungal drug discovery.
[b]The yeast mutant in which this gene is suppressed shows a severe growth defect according to SGD.
[c]The yeast mutant in which this gene is suppressed is not viable according to SGD.
[d]This yeast gene did not have a homolog in these genomes: Cc—*Coprinus cinereus*; Cn—*Cryptococcus neoformans*; Pc—*Phanerochaete chrysosporium*; Pp—*Phakopsora pachyrhizi*; Tr—*Trichoderma reesei*.
[e]Mean e-value from matches between the yeast gene and the fungal homologs calculated as the antilog of the average log e-value.
[f]Copy number was assessed by comparing each yeast gene to the yeast genome using TBLASTN. Matches with e-values less than 1E-5 were considered homologs. Two of the genes, YIL085C and YOR099W, show homology to each other at an e-value of 1E-67.

medium. These studies show an average of 18% for essential genes in these yeast genomes.

Prokaryotic Homologs

The 3340 common fungal sequences were also matched against the proteobacteria *Agrobacterium tumefasciens and Xylella fastidiosa*, which resulted in 1090 homologs (Fig. 3). Of this number, 1074 had a homolog in at least one plant or one animal genome of those tested above, demonstrating that this proteobacterial set of just over 1000 genes was conserved among eukaryotes tested. Similarly, Mannhaupt et al. (2003) found that of the homologs of *N. crassa* present in other eukaryotes (EMBL database), 33% could also be found in prokaryotes (e-value ≤ 1E-8).

Among the 3340 common fungal sequences, the intersection between the 3080 plant homologs and the 2907 animal homologs encompassed most of the 1090 proteobacterial homologs, with a total of 938 (Fig. 4). The purported functions of the 938 yeast genes with homologs in plants, animals, and bacteria were obtained from the *Saccharomyces* Genome Database and were divided among the following: metabolic functions (∼400), RNA/DNA processing (∼200), and signaling (∼150), with approximately 150 of unknown function. Among these 938 core genes, 197 have been found to be essential (*Saccharomyces* Genome Database).

Number of Core Fungal Genes

The number of common fungal genes with at least one homolog among any of the six nonfungal species totaled 3280 (Fig. 4), and of the common fungal set of 3340 genes, 60 remained without a nonfungal match. These 60 yeast genes were matched against various GenBank databases (NR, EST, GSS, HTGS) using TBLASTN, and 43 matched a nonfungal organism at an e-value ≤ 1E-5. Among these 43, many had top nonfungal matches (17 sequences) with the large vertebrate genomes (human, mouse, rat).

To account for the possibility that the matches with the nonfungal genomes might be due to fungal contamination of the nonfungal genomes, a second nonfungal match of similar expect value as the top match was required to classify the gene as not exclusively fungal. For example, if the top nonfungal match was with a mouse sequence, then one would expect that a similarly strong match would be found with a human or rat sequence from GenBank. In six cases, matches with nonfungal taxa turned out to be possible fungal contaminations in the sequence material. For example, in matches with the GenBank EST database, there were 3 yeast genes among the 60 that had matches with plants (wheat or bean), but all the other matches were with fungal sequences; these nonfungal matching sequences were downloaded from GenBank and then matched against GenBank NR, and they turned out to have no matches with any

Table 4. Possible functions of five core fungal genes without homologs in other organisms and without an annotated function in the *Saccharomyces* Genome Database

Gene	Conserved segment	GenBank NR match and e-value
YIR007W	129 aa–171 aa	Permease (12), transmembrane protein (12), immunoglobulin (12)
	190 aa–229 aa	Membrane protein (0.1), membrane protein (.55)
	264 aa–298 aa	ATPase (7.6), transposase (10)
	440 aa–473 aa	No nonhypothetical matches
	476 aa–529 aa	Chemotaxis protein (2.3), amyloid precursor (4.1), helicase (5.5)
	530 aa–563 aa	Fumarate reductase (4.2, 7.6), membrane receptor (7.6)
	564 aa–589 aa	Viral polyprotein (1.3), threonyl-tRNA synthetase (1.3)
YHR194M	72 aa–200 aa	Membrane protein (1E-36), protein kinase (6.2)
	226 aa–362 aa	Membrane protein (3E-35), adhesin protein (0.43)
	371 aa–414 aa	Membrane protein (1E-10)
	460 aa–569 aa	Membrane protein (5E-26)
YOR365C	186 aa–303 aa	Permease (0.11), spermidine synthase (0.11), Ih channel protein (0.56)
	358 aa–407 aa	No nonhypothetical matches
	420 aa–485 aa	Unknown cytoplasmic protein (9E-10), probable membrane protein (4E-5)
	494 aa–514aa	Unknown cytoplasmic protein (0.66)
	562 aa–591 aa	Unknown cytoplasmic protein (0.039)
YGR213C	9 aa–29 aa	No nonhypothetical matches
	74 aa–192 aa	Putative membrane protein (4E-17), putative transporter/flippase (7E-7)
	221 aa–263 aa	Putative membrane protein (0.006), putative transporter/flippase (0.067)
YBL095W	96 aa–219 aa	Thioesterase (0.089, 1.3), carboxyl-terminal modulator protein (1.7)

Note. Sequence alignments were made for each yeast gene and its fungal homologs, and the conserved segments were compared to the GenBank NR Protein database with an e-value up to 100.

other plant sequences, and hence the original query sequences were classified as exclusively fungal sequences.

In total, there were 17 core fungal genes out of 3340 that had no matches with any nonfungal organism based on comparisons with several GenBank databases and the additional criteria stated above (Table 3). Among these 17 genes, 7 had homologs in all of the 13 other fungal genomes, 5 had homologs in 12 of the 13 other fungal genomes, and 5 had homologs in 11 of the 13 other genomes (Table 3). Of these 15 cases where the yeast genes lacked a homolog, 14 were with basidiomycetous species (7 for *Phakopsora pachyrhizi*, 3 each for *Cryptococcus neoformans* and *Phanerochaete chrysosporium*, and 1 for *Coprinus cinereus*). Only one case was with an ascomycetous species, *Trichoderma reesei*. The inability to find a homolog in the basidiomycetes for certain yeast genes might be because there is no homolog, and not just because the genomes are incompletely sequenced.

Characteristics of the Core Fungal Genes

The purported functions of these 17 core fungal genes were obtained from the *Saccharomyces* Genome Database and are listed in Table 3. In summary, five have unknown functions, two are involved in protein biosynthesis, two are involved in transport, two have miscellaneous functions, and one is involved in sporulation. Deletion mutants have been made for all of these genes (*Saccharomyces* Genome Database), and deletion of YJL087C gave a nonviable phenotype. For YGR213C and YOR099W, there were apparent severe growth defects after several generations. Supplemental details on the 17 core fungal genes, such as BLAST results, can be found at www.uoguelph.ca/~thsiang/pubs/supplement/jmolevol04.

Among the genes exclusive to fungi, there were some with annotated functions that would seem to be necessary in other organisms. For example, YJL087C has a RNA ligase function (Table 3), but RNA ligases are found in other organisms since they are involved in repairing or restructuring RNA sequences (Belfort and Winer 1997). An explanation for this can be found in Sawaya et al. (2003), who reported that homologs of YJL087C are present in several genera of fungi such as *Candida, Schizosaccharomyces*, and *Aspergillus*, but are absent from archaea and nonfungal eukarya, perhaps because these other taxa use a different end-joining mechanism for tRNA splicing.

Since these 17 genes were found only in fungi, they may be prime targets for antifungal drug discovery. Indeed, 7 of these 17 yeast genes are listed in U.S. patents dealing with antifungal drug discovery (Table 3). According to Brown and Warren (1998), there are five important characteristics for a target site of antimicrobial activity: occurrence over a broad spectrum of pathogen species; lack of a homolog in the host species; occurrence as unique copy in pathogens,

since duplications may facilitate the evolution of resistance; be essential for the persistence of infection; and be amenable to high throughput production. All the core fungal genes satisfy the first and second requirements. Nine of the 17 exist as single copies or lack homologs in the yeast genome (Table 3), and some of these with highly conserved sequences among these fungi (mean e-value among fungal matches $\leq$ 1E-35; Table 3) should be further assessed for suitability as target sites of antimicrobial activity, notably YGR215W, YIL099W, YJL127C, and YOR132W.

The chromosomal locations of these 17 core fungal genes (indicated by the second letter in each name; Table 3) were scattered across the yeast genome, and the characteristics of these 17 exclusively fungal genes did not seem to differ greatly from other yeast genes. The GC content of over 13,000 predicted coding sequences from the yeast genome is 39.7% (www.kazusa.or.jp/codon), while the average GC content of the 17 exclusively fungal genes is 40.3%. There are four sequences which showed a departure from this average by greater than 3%: YOR365C (35.3%), YJL087C (36.0%), YHR194W (44.4%), and YIL099W (47.0%). Combined codon usage for these 17 genes is very similar to that for the coding sequences except for the following triplets and the encoded amino acid (with the difference in frequency per thousand): GAA = Glu (6.6), GGU = Gly (4.9), UUA = Leu (4.0), UUC = Phe (−4.2), UAU = Tyr (−4.9), UUU = Phe (−5.0), and UGG = Trp (−5.4). Negative numbers denote a greater frequency in the combined coding sequences of the 17 genes. The rest of the frequency differences were all less than 4.0 and greater than −4.0.

Among the 17 core fungal genes without homologs in other organisms, 5 were of unknown function. All five genes were further explored by aligning each with its fungal homologs and testing the most conserved regions for matches in the GenBank NR protein database. Four of the five genes were found to have conserved portions that showed matches with membrane or membrane-related proteins (Table 4). Studies have found that many of the predicted proteins of unknown function in the yeast proteome are membrane-related (Diehn et al. 2000). Supplemental web data to the article by Diehn et al. (2000) states that "among unknown yeast genes, 12% contain signal peptides and 31% contain transmembrane domains" (genome-www.stanford.edu/mbp/supplement.shtml#unknown). Further study is required to determine the exact function of these genes.

Evolutionary Age of Common Fungal Genes

Among 6355 predicted or known yeast genes, 3340 had homologs in at least 11 other fungal genomes, and these genes should be older than the divergence of ascomycetes and basidiomycetes. Among these 3340, 2723 could also be found in an animal and a plant genome; these genes can be regarded as an eukaryotic core for fungal genes and should be older than the divergence of the major eukaryotic kingdoms. Within this eukaryotic core, 938 of these genes also had homologs in one or two proteobacterial genomes, which can be considered to represent a prokaryotic (proteobacterial) core for fungal genes, dating from the divergence of eukaryotes from prokaryotes. Among the 3340 common fungal genes, 17 were found in no other organisms except fungi. These are genes which could have arisen after fungi diverged from other eukaryotes and before the divergence of ascomycota from basidiomycota. For only 1 of these 17 genes did a deletion mutation result in a nonviable phenotype according to the *Saccharomyces* Genome Database. While the other 16 genes seemed to have been evolutionarily conserved among fungi but not other organisms, they are not considered to be essential genes. More intensive study is required of the core fungal genes and other common fungal genes to discover which genes make fungi different from other organisms.

Acknowledgments. This research was initiated while T. Hsiang was on Research Leave at Simon Fraser University working with D.L. Baillie. The authors are very grateful to Z. Punja for providing office space and arranging access to the SFU library and Internet for T. Hsiang, as well as the fruitful discussions with him regarding fungal genomics. We are also very grateful to P.H. Goodwin and J.T. Trevors for their helpful comments on the manuscript.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3789–3402

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant, *Arabidopsis thaliana*. Nature 408:796–815

Baldauf SL (2003) The deep roots of eukaryotes. Science 300:1703–1706

Belfort M, Weiner A (1997) Another bridge between kingdoms: tRNA splicing in Archaea and Eukaryotes. Cell 89:1003–1006

Berbee ML, Taylor JW (2001) Systematics and evolution. In: Mclaughlin DJ, Mclaughlin EG, Lemke PA (eds) The Mycota VIIB. Springer, Berlin, pp 229–245

Boffell D, Nobrega MA, Rubin EM (2004) Comparative genomics at the vertebrate extremes. Nat Rev Genet 5:456–465

Brown JR, Warren PV (1998) Antibiotic discovery: Is it all in the genes? Drug Discovery Today 3:564–566

Chervitz SA, Aravind L, Sherlock G, et al. (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science 282:2022–2028

Decottignies A, Sanchez-Perez I, Nurse P (2003) *Schizosaccharomyces pombe* essential genes: a pilot study. Genome Res 13:399–406

Diehn M, Eisen MB, Botstein D, Brown PO (2000) Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. Nat Genet 25:58–62

Dutilh BE, Huynen MA, Bruno WJ, Snel B (2004) The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J Mol Evol 58:527–539

Galagan JE, Calvo SE, Borkovich KA, et al. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422:859–868

Giaever G, Chu AM, Ni L, et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418:387–91

Goffeau A, Barrell BG, Bussey H, et al. (1996) Life with 6000 genes. Science 274:546–567

Harris TW, Lee R, Schwarz E, et al. (2003) WormBase: a cross-species database for comparative genomics. Nucleic Acids Res 31:133–137

Heckman DS, Geiser DM, Eidell BR, Stauffer RL, Kardos NL, Hedges SB (2001) Molecular evidence for the early colonization of land by fungi and plants. Science 293:1129–1133

Hedges SB, Kumar S (2003) Genomic clocks and evolutionary time scales. Trends Genet 19:200–206

Hsiang T, Baillie DL (2004) Recent progress, developments and issues in comparative fungal genomics. Can J Plant Pathol 26:19–30

Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT, Davis RW, Scherer S (2004) The diploid genome sequence of *Candida albicans*. Proc Natl Acad Sci USA 101:7329–7334

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254

Koonin EV, Federov ND, Jackson JD, et al. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5:R7

Mannhaupt G, Montrone C, Haase D, Mewes HW, Aign V, Hoheisel JD, Fartmann B, Nyakatura G, Kempken F, Maier J, Schulte U (2003) What's in the genome of a filamentous fungus? Analysis of the *Neurospora* genome sequence. Nucleic Acids Res 31:1944–1954

Mardis E, McPherson J, Martienssen R, Wilson RK, McCombie WR (2002) What is finished and why does it matter? Genome Res 12:669–671

Martinez D, Larrondo LF, Putnam N, Gelpke MD, Huang K, Chapman J, Helfenbein KG, Ramaiya P, Detter JC, Larimer F, Coutinho PM, Henrissat B, Berka R, Cullen D, Rokhsar D (2004) Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. Nat Biotechnol 22:695–700

Mata J, Bahler J (2003) Correlations between gene expression and gene conservation in fission yeast. Genome Res 13:2686–2690

Mewes HW, Albertmann K, Bahr M, Frishman D, Gkeissner A, Hand J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F, Zollner A (1997) Overview of the yeast genome. Nature 387:7–65

Philippe H, Germot A, Moreira D (2000) The new phylogeny of eukaryotes. Curr Opin Genet Dev 10:596–601

Piskur J, Langkjaer RB (2004) Yeast genome sequencing: the power of comparative genomics. Mol Microbiol 53:381–389

Rubin GM, Yandell MK, Wortman JR, et al. (2000) Comparative genomics of eukaryotes. Science 287:2204–2215

Sawaya R, Schwer B, Shuman S (2003) Genetic and biochemical analysis of the functional domains of yeast tRNA ligase. J Biol Chem 278:43928–43938

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Tzung KW, Williams RM, Scherer S, Federspiel N, Jones T, Hansen N, Bivolarevic V, Huizar L, Komp C, Surzycki R, Tamse R, Davis RW, Agabian N (2001) Genomic evidence for a complete sexual cycle in *Candida albicans*. Proc Natl Acad Sci USA 98:3249–3253

Winzeler E, Shoemaker DD, Astromoff A, et al. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science 285:901–906

Wood DW, Setubal JC, Kaul R, et al. (2001) The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science 294:2317–2323

Yu J, Hu S, Wang J, et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L spp. *indica*). Science 296:79–92