Comparative genome analysis across a kingdom of eukaryotic organisms: Specialization and diversification in the Fungi

Michael J. Cornell, Intikhab Alam, Darren M. Soanes, et al.

Genome Res. published online November 5, 2007
Access the most recent version at doi:10.1101/gr.6531807

Resource

Comparative genome analysis across a kingdom of eukaryotic organisms: Specialization and diversification in the Fungi

Michael J. Cornell,[1,2] Intikhab Alam,[1] Darren M. Soanes,[3] Han Min Wong,[3] Cornelia Hedeler,[1] Norman W. Paton,[1] Magnus Rattray,[1] Simon J. Hubbard,[2] Nicholas J. Talbot,[3] and Stephen G. Oliver[2,4,5,6]

[1]*School of Computer Science, University of Manchester, Manchester M13 9PL, United Kingdom;* [2]*Faculty of Life Sciences, University of Manchester, Manchester M13 9PL, United Kingdom;* [3]*Department of Biosciences, University of Exeter EX4 4QD, United Kingdom;* [4]*Department of Biochemistry, University of Cambridge, Cambridge CB2 1 GA, United Kingdom*

The recent proliferation of genome sequencing in diverse fungal species has provided the first opportunity for comparative genome analysis across a eukaryotic kingdom. Here, we report a comparative study of 34 complete fungal genome sequences, representing a broad diversity of Ascomycete, Basidiomycete, and Zygomycete species. We have clustered all predicted protein-encoding gene sequences from these species to provide a means of investigating gene innovations, gene family expansions, protein family diversification, and the conservation of essential gene functions—empirically determined in *Saccharomyces cerevisiae*—among the fungi. The results are presented with reference to a phylogeny of the 34 fungal species, based on 29 universally conserved protein-encoding gene sequences. We contrast this phylogeny with one based on gene presence and absence and show that, while the two phylogenies are largely in agreement, there are differences in the positioning of some species. We have investigated levels of gene duplication and demonstrate that this varies greatly between fungal species, although there are instances of coduplication in distantly related fungi. We have also investigated the extent of orthology for protein families and demonstrate unexpectedly high levels of diversity among genes involved in lipid metabolism. These analyses have been collated in the *e–Fungi* data warehouse, providing an online resource for comparative genomic analysis of the fungi.

[Supplemental material is available online at www.genome.org.]

The Fungi represent a single eukaryotic kingdom, characterized by an osmotrophic growth habitat in which extracellular enzymes are secreted to break down complex substrates, the resulting simple sugars and amino acids being taken up by the growing fungus. Fungi exist in two distinct morphological growth forms, the unicellular yeasts (which grow by budding or simple fission) and the filamentous fungi (which produce polarized hyphal strands that aggregate to form a network called a mycelium). The osmotrophic growth habit of fungi is extremely effective for colonizing diverse habitats and has made the fungi the principal degraders of biomass in all terrestrial ecosystems (de Boer et al. 2005) and also important pathogens of both plants and animals.

The yeasts and filamentous fungi cover a huge evolutionary range. The Pezizomycotina (filamentous ascomycetes) and the Saccharomycotina (budding yeasts), for example, diverged from one another some 900–1000 million years ago (Mya) (Hedges et al. 2004), and the Saccharomycotina alone are more evolutionarily diverged than the Chordate phylum of the animal kingdom (Goffeau 2004).

[5]Present address: Department of Biochemistry, University of Cambridge, Sanger Building, 80 Tennis Court Road, Cambridge CB2 1 GA, UK.
[6]Corresponding author.
E-mail Steve.oliver@mole.bio.cam.ac.uk; fax 44-1223-766-002.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6531807. Freely available online through the *Genome Research* Open Access option.

It is also striking that among the 431 published studies describing completed genome sequences (October 2006), 361 describe bacterial genomes, 28 Archaeal genomes, and only 42 are genomes from eukaryotic organisms (for compendium, see http://www.tigr.org/tdb/ and the genomes online database http://www.genomesonline.org). Of the eukaryotic genome sequences currently available, more than half come from the kingdom Fungi, making them the only credible candidates for a kingdom-wide exploration of genomic relatedness. In addition to sampling this vast evolutionary diversity, there are complete genome sequences now available for very closely related fungal species—for instance, the *Saccharomyces* "sensu stricto" species group whose members are so closely related that they can mate to produce viable hybrids (Naumov et al. 1997, 2000; Marinoni et al. 1999). The diverse biology and morphology of fungi, coupled with the availability of substantial genomic sequence information, means that the Fungi are a unique group with which to study eukaryotic genome evolution, and the methods that are developed for the analyses, as well as the results obtained, are likely to help in comparative genome analyses of the other eukaryotic kingdoms. To this end, we have assembled genome sequences from filamentous fungi and yeast into a database, which we call *e-Fungi*. In this study, we report the results of a comprehensive comparative genomic analysis carried out using the *e-Fungi* resource.

The majority of the publicly available fully sequenced fun-

gal genomes come from Ascomycete species, but we have also investigated two sequenced Basidiomycete fungi, *Ustilago maydis* and *Phanerochaete chrysosporium*, plus the Zygomycete *Rhizopus oryzae* and the Microsporidian *Encephalitozoon cuniculi*. In addition, we have included two non-fungal species, the Oomycetes *Phytophthora sojae* and *Phytophthora ramorum*. Oomycetes are heterokonts that diverged from the Fungi very early in the evolution of eukaryotes. Nevertheless, in terms of their biological activity, they demonstrate convergent evolution with fungal plant pathogens. For example, oomycetes share a filamentous, osmotrophic growth habit and are capable of invasive growth and colonization of living plant tissue (Money et al. 2004).

We have used the features of *e-Fungi* to ask some fundamental questions about fungal evolutionary biology, including patterns of gene loss and gene duplication, extent of protein family conservation among the kingdom Fungi, and the distinctive genomic features that separate yeasts and filamentous species. Our study represents the most comprehensive analysis of the kingdom Fungi carried out to date and has highlighted some surprising levels of hitherto unrecognized functional and biochemical diversity among fungal species.

Results

Comparative genomics of the Fungi

Our first aim was to organize all available genomic data into a form that could be readily used to assess the evolutionary relatedness among fungal species and also to generate clusters of related sequences. The 36 genomes used in this analysis are listed in Table 1. There are clearly large differences in the size of fungal genomes and the number of predicted protein sequences. They range in size from 2.5 Mb, encoding 1996 predicted open reading frames (ORFs), for the microsporidian *E. cuniculi*, to >46 Mb, encoding some 17,467 ORFs, for *R. oryzae*. Among the Ascomycete species, the Pezizomycotina have larger genomes than the Saccharomycotina: in general, the Saccharomycotina genomes have sizes between 8 and 12 Mb, while the Pezizomycotina are between 29 and 39 Mb. The exception is *Yarrowia lipolytica*, which has a 20-Mb genome—nearly twice as large as that of the next largest in the Saccharomycotina. However, despite being larger, the *Y. lipolytica* genome does not appear to encode significantly more protein sequences than the other Saccharomycotina genomes. There are 6544 ORFs predicted for *Y. lipolytica* compared to 6317 in *Debaromyces hansenii*.

In order to generate clusters of proteins that are conserved among particular fungal species, translated amino acid sequences were first compared using BLASTP (Altschul et al. 1990). Sequences were then clustered on the basis of their similarity using an *E*-value cutoff of 10×10^{-5} (for more details, see Methods). This method generates clusters solely on the basis of similarity, and no distinction is made between orthologs and paralogs. However, using the information stored in the *e-Fungi* database, we are able to distinguish between orthologs and paralogs when running analyses. Orphan proteins, i.e., those not part of any cluster, are classified as such because they are not sufficiently similar to other proteins either in their own or other genomes. This could be for one of two reasons. First, orphans could arise due to errors in genome annotation. For example, the sequencing of members of the *Saccharomyces* "sensu stricto" complex by two different research groups (Cliften et al. 2003; Kellis et al. 2003) has resulted in very different numbers of predicted ORFs, despite similar degrees of sequence coverage being achieved. ORFs that are not real are clearly less likely to be conserved across species. Second, the occurrence of orphans could be due to a lack of sequence data from closely related fungal species. For example, there are several Saccharomycotina yeast species in this analysis and, as a consequence, >97% of the predicted protein products of *Eremothecium* (formerly *Ashbya*) *gossypii* genes are members of clusters. In comparison, *Schizosaccharomyces pombe* is the only Archaeascomycete yeast in the data set, and 20% of its ORFs encode proteins that are not included in any cluster.

Phylogenies of the fungi based on concatenated protein sequences and Dollo parsimony

A common approach in molecular phylogenetics is to use sequence comparisons of single genes or proteins from multiple species in order to construct phylogenies. However, the availability of complete genome sequences allows the construction of more accurate phylogenies by using multiple concatenated protein sequence alignments. We used a maximum likelihood approach in order to construct phylogenetic trees based on concatenated sequence alignments from multiple protein families. Maximum likelihood methods are thought to be least seriously influenced by saturation, and resulting long-branch artifacts, in comparison to parsimony and distance-based methods (Felsenstein 1978; Guindon and Gascuel 2003).

Protein families were selected for constructing the sequence similarity trees using protein clusters that spanned all the genomes but showed minimal duplication within genomes. The proteins obtained (see Supplemental Table 1) are generally involved in fundamental processes such as translation, transcription, and DNA replication and recombination. Because of the great divergence between species in this analysis, and the small size of the *E. cuniculi* genome, two sets of protein families were used to resolve the phylogeny. A well-resolved broader species tree based on concatenated sequences from 12 protein families, which included *Homo sapiens* and the model plant species *Arabidopsis thaliana*, clearly separates fungal, animal, and plant taxa with maximum bootstrap support, as shown in Figure 1A. The oomycete species are shown to have diverged prior to the divergence of fungi and animals. This is in agreement with previous phylogenetic analysis based on small subunit rRNA sequences (Baldauf et al. 2000). *E. cuniculi,* while appearing to be well resolved, has a very long branch, indicative of rapidly evolving genes. Again, this is consistent with previous phylogenetic analysis (Thomarat et al. 2004).

A second set of 29 protein families was used to generate a tree that resolved the deeper nodes, as shown in Figure 1B. Using this larger protein set, high levels of support have been achieved for the majority of branches. The tree shows that the divergence of the Zygomycetes followed by that of the Basidiomycetes mark the earliest bifurcations in the fungal tree, followed by the split of *S. pombe* (Taphrinomycotina) from the rest of the Ascomycetes. The Ascomycetes then divide into the Pezizomycotina and the Saccharomycotina. There is support for this tree from previous sequence-based phylogenetic studies. These results are also consistent with other recent phylogenies of the fungi (Fitzpatrick et al. 2006; Galagan et al. 2006; James et al. 2006), which confirm the well-supported aspects of this species tree.

Phylogenetic profiling, which uses the presence and absence of derived character states such as genes and gene families, is

Table 1. Overview of MCL clusters

Genome	Genome size (Mb)[a]	Total ORFs	ORFs found in clusters	Data source
Saccharomyces cerevisiae[b]	12.2	5869	5529	NCBI[c]
Saccharomyces paradoxus[d]	11.9	8955	6630	Saccharomyces Genome Database[e]
Saccharomyces mikatae[d]	11.5	9057	6495	Saccharomyces Genome Database[e]
Saccharomyces mikatae[f]	10.8	3100	2975	Saccharomyces Genome Database[e]
Saccharomyces kudriavzevii[f]	11.2	3768	3727	
Saccharomyces bayanus[d]	11.5	9424	6218	Saccharomyces Genome Database[e]
Saccharomyces bayanus[f]	11.9	4966	4943	Saccharomyces Genome Database[e]
Saccharomyces castellii[f]	11.4	4677	4549	Saccharomyces Genome Database[e]
Saccharomyces kluyveri[f]	11.0	2968	2938	Saccharomyces Genome Database[e]
Candida glabrata[b]	12.3	5180	4848	NCBI[c]
Kluyveromyces waltii	10.6	5230	5020	Kellis et al. 2004[g]
Eremothecium gossypii[b]	8.8	4718	4582	NCBI[c]
Kluyveromyces lactis[b]	10.7	5326	4989	NCBI[c]
Candida albicans[h]	27.6	14213	13647	NCBI[c]
Candida lusitaniae	12.1	5941	5243	Broad Institute[i]
Debaryomyces hansenii[b]	12.2	6310	5679	NCBI[c]
Yarrowia lipolytica[b]	20.6	6544	5536	NCBI[c]
Aspergillus nidulans	30.1	10701	9367	Broad Institute[i]
Aspergillus niger	37.2	11200	10072	DOE Joint Genome Institute[j]
Aspergillus fumigatus	29.4	9923	8841	CADRE Database[k]
Aspergillus oryzae	37.2	12062	10199	DOGAN Database[l]
Aspergillus terreus	29.3	10406	9253	Broad Institute[i]
Coccidioides immitis	28.9	10457	7949	Broad Institute[i]
Stagonospora nodorum	37.2	16597	10461	Broad Institute[i]
Botrytis cinerea	42.7	16448	10238	Broad Institute[i]
Sclerotinia sclerotiorum	38.3	14522	9947	Broad Institute[i]
Gibberella zeae	36.1	11640	10048	Broad Institute[i]
Trichoderma reesei	34.5	9783	8302	DOE Joint Genome Institute[j]
Magnaporthe grisea	38.8	12841	9816	Broad Institute[i]
Neurospora crassa	38.0	9826	7633	Broad Institute[i]
Chaetomium globosum	34.9	11124	9121	Broad Institute[i]
Schizosaccharomyces pombe[b]	12.5	4416	3212	NCBI[c]
Ustilago maydis	19.7	6522	5211	Broad Institute[i]
Phanerochaete chrysosporium	30.0	11776	6927	DOE Joint Genome Institute[j]
Rhizopus oryzae	46.1	17467	14092	Broad Institute[i]
Encephalitozoon cuniculi[b]	2.5	1996	1079	NCBI[c]
Phytophthora sojae	86.0	19071	15578	DOE Joint Genome Institute[j]
Phytophthora ramorum	66.7	15890	13848	DOE Joint Genome Institute[j]

The numbers of ORFs listed in Table 1 for *Y. lipolytica* and *D. hansenii* differ from those previously published (6703 and 6906) (see Table 2; Dujon et al. 2004). This appears to be due to the removal of pseudogenes from our analysis. When loading data into the *e-Fungi* database, we identified 151 *Y. lipolytica* and 579 *D. hansenii* CDS sequences that were annotated as pseudogenes.
[a]Genome sizes estimated from total size of all chromosomes or contigs.
[b]Genomes for which complete chromosome sequences are available.
[c]NCBI (http://www.ncbi.nlm.nih.gov/Genomes/).
[d]Genome sequenced by Kellis et al. 2003.
[e]Saccharomyces Genome Database (ftp://genome.ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes/).
[f]Genome sequenced by Cliften et al. 2003.
[g]Kellis et al. 2004, Supplemental data (http://www.nature.com/nature/journal/v428/n6983/extref/nature02424-s1.htm).
[h]Genome sequence for *C. albicans* is for the diploid form (Jones et al. 2004).
[i]Broad Institute (http://www.broad.mit.edu/annotation/fgi/).
[j]DOE Joint Genome Institute (http://www.jgi.doe.gov).
[k]CADRE Database (http://www.cadre-genomes.org.uk).
[l]DOGAN Database (http://www.bio.nite.go.jp/dogan/MicroTop?GENOME_ID=ao).

commonly used for tree construction (Fitz-Gibbon and House 1999; Wolf et al. 2002; Huson and Steel 2004). The Dollo parsimony method (Farris 1977), based on the model that derived character states that are lost cannot subsequently be regained, is applicable to such phylogenetic profiles for the reconstruction of the most parsimonious evolutionary paths. Huson and Steel (2004) showed this approach to perform favorably in comparison to distance-based methods for constructing gene content trees. Since the presence/absence data of character states can naturally be treated as binary characters, a matrix of 1/0 (presence/absence), constructed from 23,230 protein clusters present and absent across Ascomycetes, Basidiomycetes, and Zygomycetes, was subjected to the Dollo parsimony method. The resulting phylogenetic tree is shown in Figure 1C. In general, the Dollo parsimony tree supports the sequenced-based tree. There are differences in the positions of the members of the *Saccharomyces* "sensu stricto" complex, but these can be attributed to the differences in the reporting of ORFs by the different sequencing groups, which would clearly affect any analysis based on gene presence or absence. The two major differences between the trees

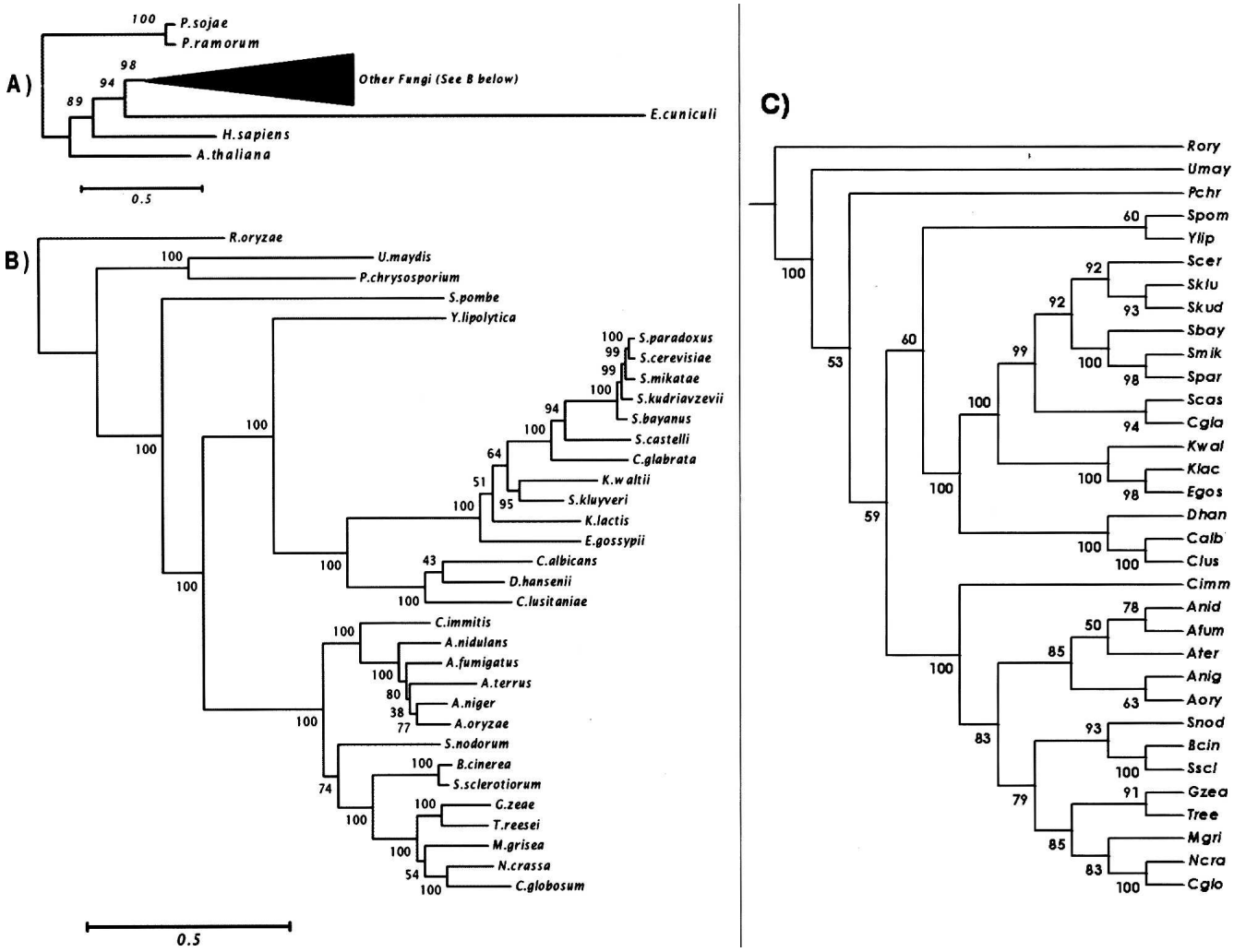

Figure 1. Species trees. (*A*) Broad species tree based on concatenated sequences from 30 universal protein clusters using maximum likelihood approach. (*B*) Basidiomycete and Ascomycete species tree based on 60 universal fungal proteins. (*C*) Fungal phylogenetic tree based on Dollo parsimony. Strongly supported branches agree with the sequence-based fungal species tree shown in *B*.

are, firstly, that the Dollo parsimony analysis places *S. pombe* next to *Y. lipolytica*, after the division of the Ascomycetes into the Pezizomycotina and the Saccharomycotina. Therefore, on the basis of gene presence/absence, *S. pombe* is closer to *Saccharomyces cerevisiae* than would be expected on the basis of sequence conservation. Secondly, the Dollo parsimony analysis does not position *Coccidioides immitis* next to the *Aspergillus* species, but places it as the most diverged member of the Pezizomycotina.

Gene duplication in the Fungi

Gene duplication is a major force in evolution, allowing the acquisition of new functions. Duplicated genes may be retained if they provide an advantage, either because of an increased gene dosage effect or because of increased functional diversification. Alternatively, the duplicated gene may be lost. In the evolution of the *S. cerevisiae* and *Candida globrata* genomes, following a whole-genome duplication (WGD), the majority of the duplicated genes have been lost. In Pezizomycotina species, a repeat-induced point mutation (RIP; a mechanism for the active removal of duplicated sequences) has been identified. The sequencing of the *Neurospora crassa* genome has shown that RIP activity has resulted in the presence of very few duplicated genes (Galagan and Selker 2004).

A previous study (Dujon et al. 2004) analyzed genome redundancy in five Saccharomycotina yeast genomes (*S. cerevisiae*, *C. glabrata*, *Kluyveromyces lactis*, *D. hansenii*, and *Y. lipolytica*). We have expanded this analysis to include 14 Pezizomycotina species, the Taphrinomycotina *S. pombe,* two Basidiomycetes, a Zygomycete, and a Microsporidian. The *Candida albicans* genome has been excluded from the analysis because the sequence used refers to the normal diploid form of the fungus and, therefore, we would expect a minimum of two copies for each protein-encoding sequence. Protein clusters were analyzed to determine whether they contained more than one protein from a given fungal species. The result of this analysis is shown in Supplemental Figure 1A. For Saccharomycotina genomes, our results are consistent with those of the previously published study (Dujon et al. 2004); although the level of duplication in the *D. hansenii* genome appears lower in our analysis. *S. cerevisiae* (438 duplication-containing clusters) has more than twice as many duplications as *K. lactis* or *Kluyveromyces waltii* (206 and 181 clusters,

respectively), which diverged prior to the WGD event. *C. glabrata* appears to have fewer duplication-containing clusters than *S. cerevisiae*. Only 325 were identified, indicating greater gene loss post-WGD. *Y. lipolytica* possesses the greatest number of highly duplicated clusters. Forty-seven clusters containing more than five proteins were identified, compared with only 10 for *S. cerevisiae*. Dujon et al. (2004) found that *K. lactis* had the lowest level of duplication of the yeast genomes analyzed, while our analysis of the *E. gossypii* genome indicates that it has even less, possibly a reflection of it possessing the smallest genome of the Saccharomycotina analyzed.

Gene duplication among the Pezizomycotina, in general, appears to be slightly higher than among the Saccharomycotina. The exceptions are *N. crassa* and *C. immitis* (372 and 374 clusters, respectively), which both possess fewer duplication-containing clusters than *S. cerevisiae*. Among the *Aspergillus* genomes, *Aspergillus niger* and *Aspergillus oryzae* (which are positioned next to each other on the phylogenetic tree) possess the most duplication-containing clusters. In the Basidiomycetes, duplication in the *P. chrysosporium* genome (885 clusters) appears much higher than in that of *U. maydis* (300 clusters). There is little duplication in the *E. cuniculi* genome, a likely reflection of its very small genome size. In contrast, the Zygomycete *R. oryzae* possesses by far the most duplication clusters (2481) of all the fungi analyzed, almost three times as many as the next highest, *P. chrysosporium*. In order to ensure that these results were not due to the presence of transposons, the analysis was repeated with protein clusters containing likely transposons removed. This did not greatly alter the results of the analysis (see Supplemental Fig. 1B); although some fungal genomes appear to contain large numbers of transposons, the sequence similarity of these elements means that they occur in only a small number of clusters.

As discussed above, the low number of duplications in the *N. crassa* genome is a result of RIP activity. Orthologs of the enzyme responsible, a DNA methyltransferase encoded by the *rid* gene, have been identified in other Pezizomycotina (Galagan and Selker 2004), and there is evidence of RIP activity in *Tan1* transposable element sequences in the genome of *A. oryzae* (Montiel et al. 2006). Our analysis shows that *N. crassa* has fewer duplication-containing clusters than all the Pezizomycotina except *C. immitis*, suggesting that the RIP process is more active in *N. crassa* than other Pezizomycotina species. This is in agreement with experimental evidence that shows that RIP is less active in *Magnaporthe grisea* than in *N. crassa* (Ikeda et al. 2002). To further investigate this, we repeated the analysis of Galagan and Selker (2004) and analyzed the percentage identity of best BlastP hits within each species. The results (shown in Supplemental Fig. 2) demonstrate that there is great variability within the Pezizomycotina in terms of the number of proteins with high percentage identity. For *M. grisea*, 911 proteins had BlastP hits with identity between 90% and 100%, while *Sclerotinia sclerotiorum* had 529 and *Chaetomium globosum* 419. In contrast, *Gibberella zeae* had 21, *A. oryzae* 34, *A. niger* 68, and *N. crassa* 61. Again, this result supports the suggestion that there are differing levels of RIP activity in Pezizomycotina species. Among the non-Pezizomycotina species analyzed, the most intriguing result is the large number of *R. oryzae* proteins with identity scores between 90% and 100%. A total of 4231 proteins were identified, almost 10 times as many as in *S. cerevisiae* (435). This result indicates that there has been a recent large-scale duplication of *R. oryzae* genes.

Gene duplication can lead to an organism possessing genes of novel function, potentially improving adaptation to its environment. It has therefore been proposed that duplication of different sets of genes may promote speciation (Lynch and Force 2000) and, conversely, that parallel duplication of the same sets of genes might allow distantly related species to adapt to the same niche (Hughes and Friedman 2003). An analysis of the *S. cerevisiae* and *S. pombe* genomes has shown that independent gene duplication of the same gene families has occurred to a far greater extent than expected by chance (Hughes and Friedman 2003).

We have investigated parallel duplication between pairs of fungal species by identifying instances of coduplication in protein clusters. The frequency with which these clusters occurred was compared to the expected frequency based on duplications occurring independently in each species using a χ^2 analysis. Some of the results of this analysis are shown in Table 2, and the remainder are in Supplemental Table 2. Our analysis does not distinguish between independently occurring duplications and those that occurred in ancestral species. Possibly as a result of this, we obtain significant chi-squared results for many pairs of fungi. However, we are able to identify instances where the χ^2 score is not in agreement with the phylogenetic tree. For example, for *S. cerevisiae*, the highest chi-squared score is obtained for *C. glabrata*, and the scores decrease in other Saccharomycotina species, with *Y. lipolytica* having a score similar to those of some of the Pezizomycotina. The score for *S. pombe* is higher than for *Y. lipolytica*, in agreement with the parallel duplication observed by Hughes and Friedman (2003). Both Basidiomycete species have low scores, but the score for *R. oryzae* is unexpectedly high, possibly indicating adaptation to a similar ecological niche. Both *S. cerevisiae* and *R. oryzae* are often found in environments associated with rotting fruit, although this is also true of many other fungal species (Wheeler et al. 2003; Tournas and Katsoudas 2005). Alternatively, it could be that this high level of coduplication reflects the fact that there has been a WGD event in the

Table 2. Chi-squared scores for coduplication in OrthoMCL clusters

Clusters	*S. cerevisiae*	*S. pombe*	*R. oryzae*
Scer	—	300.1	116.4
Cgla	1244.1	94.6	25.2
Kwal	566.4	100.1	14.0
Klac	476.7	94.8	17.8
Egos	528.2	89.1	21.6
Dhan	350.2	99.7	6.1
Clus	393.0	54.5	24.9
Ylip	88.2	15.1	1.0
Cimm	21.0	8.6	2.5
Anid	57.3	17.2	0.2
Afum	63.5	16.8	2.0
Ater	51.2	14.3	0.0
Anig	35.9	12.9	8.0
Aory	54.7	11.0	1.8
Snod	23.0	0.5	31.3
Bcin	84.9	22.8	0.1
Sscl	66.5	22.2	0.4
Gzea	30.9	6.5	17.4
Tree	71.1	17.2	0.9
Mgri	16.4	3.4	16.1
Ncra	56.9	16.0	0.8
Cglo	24.5	2.6	16.5
Spom	300.1	—	19.8
Pchr	12.9	2.8	29.1
Umay	9.5	22.9	2.8
Rory	116.4	19.8	—
Ecun	3.3	0	4.9

evolution of *S. cerevisiae* and, as discussed above, large-scale duplication in that of the *R. oryzae* genome. These two examples of parallel duplication seem specific to these pairs, the scores for *C. glabrata* and *S. pombe*, and for *C. glabrata* and *R. oryzae* are both much lower, despite *S. cerevisiae* and *C. glabrata* being closely related. It is interesting to note that there appears to be little parallel duplication between *S. pombe* and *R. oryzae*.

To investigate the genetic basis of the yeast and filamentous growth habits by fungi, we analyzed Pfam motifs associated with proteins from Saccharomycotina and Pezizomycotina species, to identify those motifs for which there is expansion in protein numbers for the Pezizomycotina compared to the Saccharomycotina and vice versa. Table 3 lists examples of the motifs identified. Many of the motifs expanded in the Pezizomycotina indicate increased metabolic flexibility compared to the Saccharomycotina. For example, there are expansions in protein families involved in transport into and out of the cell. Major facilitator superfamily transporters (PF07690), involved in the transport of small solutes, including sugar uptake and drug efflux (Maiden et al. 1987; Griffith et al. 1992), sugar transporter (PF00083) proteins, and ABC transporters (PF00005), which have been identified as multidrug resistance proteins (Jungwirth and Kuchler 2006), all show expansion. There is also an expansion in motifs associated with the utilization of different carbon sources, e.g., the alcohol dehydrogenase domains PF08240 and PF00106. Analysis of cytochrome P450 proteins (PF00067) shows that there has been a huge expansion in the number of P450 proteins in the Pezizomycotina compared to the Saccharomycotina. The numbers of P450 proteins identified in Pezizomycotina species are in broad agreement with a previous analysis (Deng et al. 2007). The P450 family is highly diverged, and P450 proteins have a range of roles, including the degradation of toxins and the production of secondary metabolites, and are important in fungal adaptation (Deng et al. 2007). As well as the expansion in the Pezizomycotina, our analysis indicates that *P. chrysosporium* and, to a lesser extent, *R. oryzae* also possess large numbers of P450s. The low number of cytochrome P450 proteins found in the Saccharomycotina is also shown by *S. pombe*, which possesses only two.

In addition to the expansion in Pfam motifs associated with responses to environmental stresses and resources, there is an expansion in the motifs associated with regulation of gene expression. Analysis of 84 Pfam motifs associated with DNA binding shows that, on average, Pezizomycotina species possess almost twice as many proteins containing these motifs as Saccharomycotina species. This expansion includes the most frequently identified Pfam motifs—PF00172, PF04082, PF00096, PF00098, and PF00170.

Protein clusters that have expanded in the Saccharomycotina compared to the Pezizomycotina are rarer, a likely reflection of the fact that the Saccharomycotina have fewer predicted ORFs. The four clusters that we have identified are all associated with cell wall structure and biosynthesis. There is a large expansion in the number of proteins containing PIR motifs (PF00399). These motifs directly link to the 1,3-beta-glucan and are important for cell wall anchoring (Kapteyn et al. 1999; Castillo et al. 2003). *S. cerevisiae* Pir2p (currently Hsp150p) has been shown to locate to the cell wall of *G. zeae*, suggesting that the incorporation mechanism of ASL-CWPs exists in both Saccharomycotina and Pezizomycotina (Narasimhan et al. 2003). Our analysis identified proteins containing PIR domains in several Pezizomycotina species. However, we failed to identify PIR proteins in any of *Aspergillus* species or in *C. immitis*, and they also appear to be absent from *C. globosum*. We were also unable to identify PIR motifs in any of the non-Ascomycete species. There are also expansions for three motifs associated with cell wall synthesis. These include glucan synthases (PF02364), which catalyze the formation of beta-1,3-glucan, and the PMT (PF02366) and MIR (PF02815) domains associated with mannosyltransferase activity. These differences are probably a consequence of the difference in the chemical composition of the cell wall between the Saccharomycotina and Pezizomycotina. The cell walls of the filamentous ascomycetes, such as *Neurospora* and *Aspergillus* species, contain far greater quantities of chitin, the major structural component of the hyphal wall, than those of yeasts, where chitin is confined to the bud scars (for review, see Bowman and Free 2006). In line with this, proteins with Pfam domains associated with chitin are much more frequently found in the Pezizomycotina (see Supplemental Fig. 3).

Protein family conservation and diversification among the Fungi

To investigate gene orthology between different species, we analyzed the co-occurrence of genomes in particular protein clusters. This provides an indication of whether a particular orthologous gene or gene family is conserved among particular fungal species and provides a powerful means of analyzing conservation in fun-

Table 3. Expansion of Pfam motifs within the Pezizomycotina and Saccharomycotina

Pfam motifs	Scer	Cgla	Kwal	Klac	Egos	Dhan	Clus	Ylip	Afum	Anid	Anig	Aory	Ater	Bcin	Cglo	Cimm	Gzea	Mgri	Ncra	Sscl	Snod	Tree
PF07690	45	34	63	56	31	88	61	123	220	267	379	415	303	135	168	127	251	176	111	181	299	166
PF00083	33	18	26	23	13	56	26	27	86	108	98	121	115	42	49	24	106	64	36	55	89	61
PF00005	31	25	28	23	20	31	34	32	52	49	71	92	58	47	55	49	59	47	35	57	61	57
PF00067	3	3	3	5	3	9	8	17	71	117	149	154	115	77	89	43	107	132	39	90	146	71
PF00106	14	13	17	18	16	35	22	28	113	157	196	170	148	87	98	61	149	121	61	92	157	123
PF08240	20	12	18	16	9	34	16	20	59	73	118	105	82	30	52	29	71	55	28	53	78	49
PF00172	52	39	47	58	40	88	69	50	188	239	278	202	184	75	132	95	260	111	90	106	155	188
PF00096	38	44	33	30	32	37	45	34	52	61	61	61	66	40	79	63	65	56	55	72	106	44
PF00399	10	11	6	5	3	2	2	8	0	0	0	0	0	1	0	0	1	1	1	1	2	1
PF02364	3	3	1	3	3	2	6	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PF02366	6	6	5	4	4	4	5	4	3	3	3	3	3	2	3	3	3	3	3	3	3	3
PF02815	7	7	5	5	4	4	5	4	3	3	3	3	3	2	3	3	3	3	3	4	3	3

Pfam motifs: PF07690 (MFS_1), PF00083 (Sugar_tr), PF00005 (ABC_tran), PF00067 (p450), PF00106 (adh_short), PF08240 (ADH_N), PF00172 (Zn_clus), PF00096 (zf-C2H2), PF00399 (PIR), PF02364 (Glucan_synthase), PF02366 (PMT), PF02815 (MIR).

gal species exhibiting distinct growth habits or lifestyles. Figure 2A shows the relative frequency of genome numbers in protein clusters for three species: *S. cerevisiae*, *C. glabrata*, and *S. pombe*. The plots are similar for *S. cerevisiae* and *C. glabrata*, reflecting the fact that they are positioned close together in our phylogenetic tree. There is an extra peak for *S. cerevisiae* at five genomes, reflecting its close relationship with the other members of the *Saccharomyces* "sensu stricto" clade, with which, indeed, it can interbreed. We expand on these relationships below.

The composition of protein clusters encoded by *S. cerevisiae* and conserved in 4–6, 9–11, and 13–15 species has been determined. For the clusters in these peaks, each genome was scored as present or absent, and then the percentage of clusters in which a given genome was present was calculated. The results are displayed in Figure 2B. The clusters containing 4–6 genomes are clearly composed of proteins from the most closely related members of the *Saccharomyces* "sensu stricto" complex—*Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, and *Saccharomyces kudriavzevii*. It is apparent that *S. kudriavzevii* has a lower score than the other "sensu stricto" genomes, reflecting the

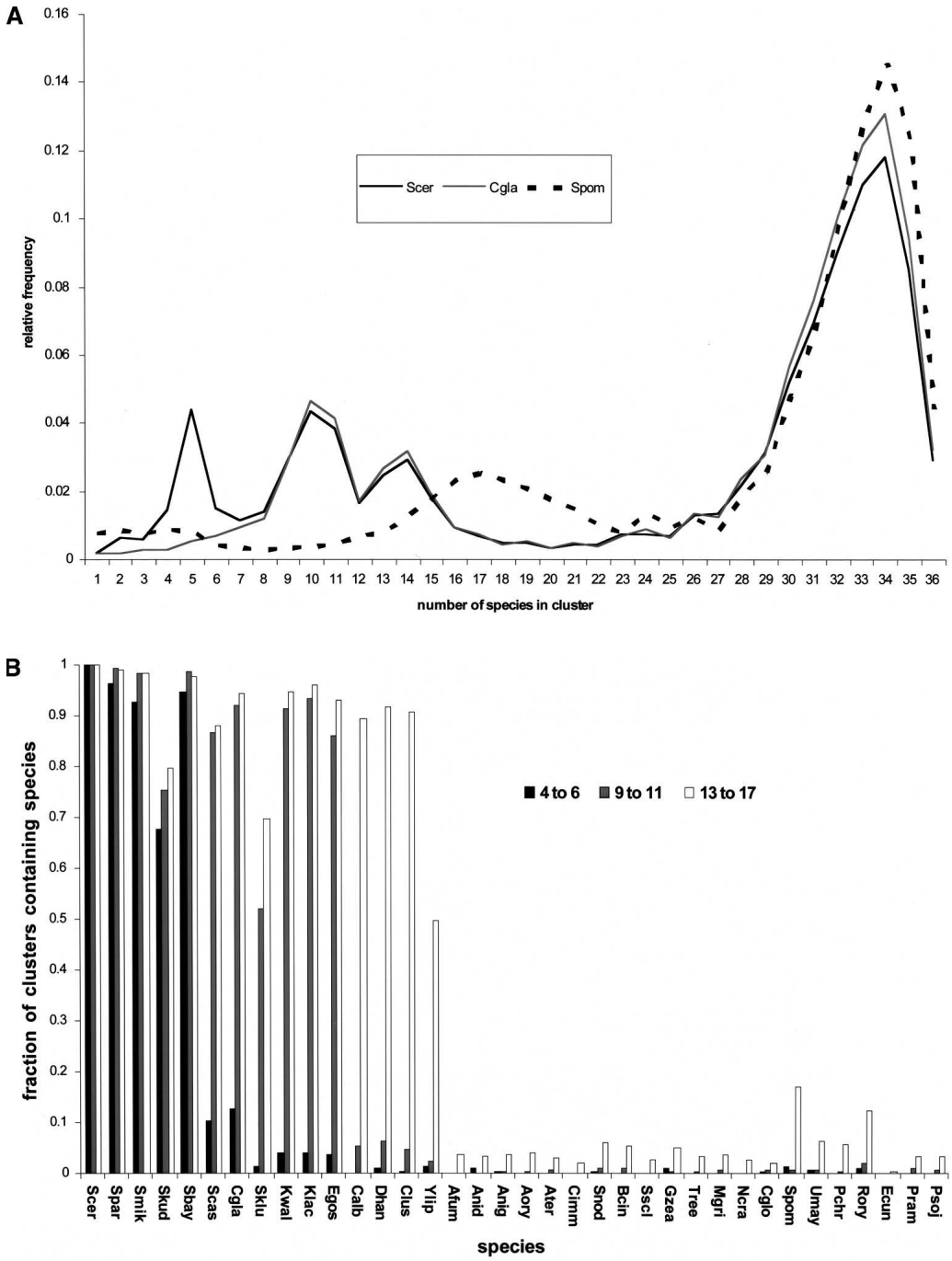

Figure 2. (Continued on next page)

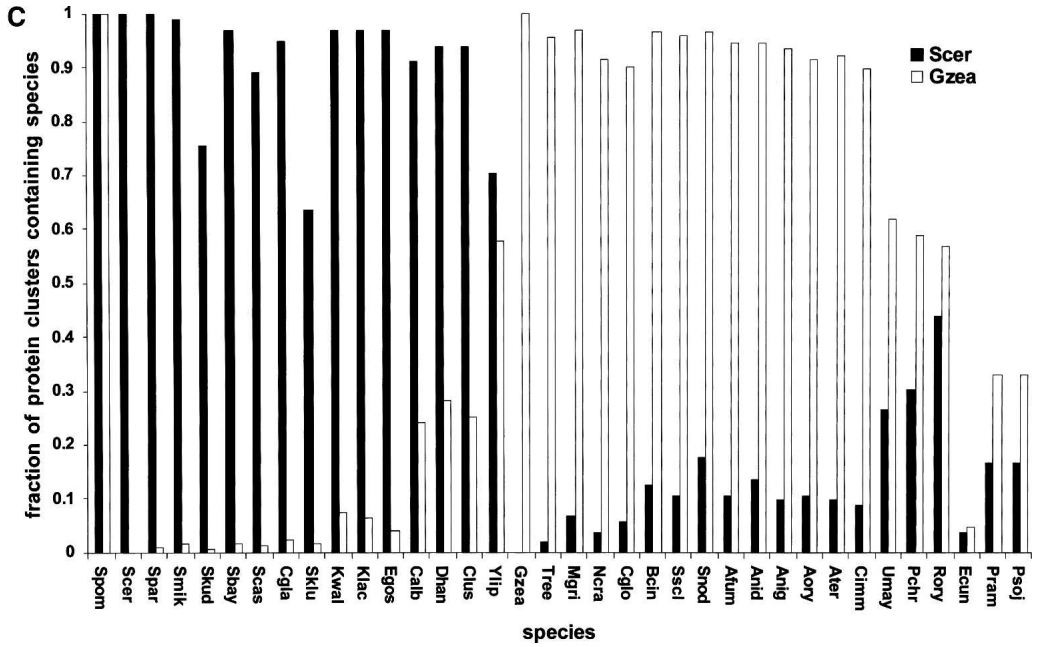

Figure 2. Protein family conservation for fungal species. (*A*) The graph shows the relative frequency of numbers of species represented in clusters containing *S. cerevisiae* (black), *C. glabrata* (gray), and *S. pombe* (dashed). (*B*,*C*) The graphs show the fraction of protein clusters containing a representative from a fungal genome. Protein clusters are selected on the basis of the following criteria: (*B*) Protein clusters containing *S. cerevisiae* and a total of four to six genomes (black), nine to 11 genomes (gray), and 13 to17 genomes (white). (*C*) Protein clusters containing *S. pombe* and a total of 13 to 24 genomes including *S. cerevisiae* (black) and *G. zeae* (white).

lower number of ORFs reported for this species (Cliften et al. 2003). This finding is consistent with our phylogenetic analysis, which showed these five species to be very closely related. Two other *Saccharomyces* species, *Saccharomyces castellii* and *Saccharomyces kluyveri*, are members of the *Saccharomyces* "sensu lato" group, which (unlike the "sensu stricto" clade) are unable to mate with *S. cerevisiae* (Marinoni et al. 1999). Our phylogenetic tree confirms that they are more distantly related to *S. cerevisiae*, and the protein products of their genes are found in far fewer of these clusters (~11% and 1.4%, respectively). It appears, therefore, that the peak at 4–6 genomes represents a group of proteins specific to members of the *Saccharomyces* "sensu stricto" complex, which are largely absent from other budding yeasts.

We analyzed a set of 977 clusters containing essential *S. cerevisiae* proteins. The distribution of cluster sizes obtained using this set is different from that obtained using all *S. cerevisiae* proteins, the peak at 4–6 proteins is absent (see Supplemental Fig. 4). Only three essential genes that are peculiar to the *Saccharomyces* "sensu stricto" clade were identified: *KAR1* (YNL188W), *SPC29* (YPL124W), and *TEN1* (YLR010C). Kar1p is an essential protein involved in nuclear fusion during mating and in spindle pole body duplication during mitosis (Pereira et al. 1999). Spc29p is also a component of the spindle pole body, linking the central plaque component Spc42p to the inner plaque component Spc110p (Elliott et al. 1999). Ten1p is involved in the regulation of telomere length as part of the nuclear telomere cap complex (Grandin et al. 2001).

Analysis of protein clusters containing 9–11 genomes shows that proteins from *S. castellii* and *S. kluyveri* now occur in the majority of the clusters, along with those from *C. glabrata*, *E. gossypii*, and *K. lactis*. The fraction of clusters in which a genome is found reduces sharply at *C. albicans* and *D. hansenii*. Again, this is consistent with the phylogenetic relationships. Analysis of clusters containing 13–15 genomes shows that *C. albicans*, *D. hansenii*, and *Y. lipolytica* now occur in the majority of the clusters, and the fraction now reduces sharply at the Pezizomycotina. It is interesting to note that despite *S. pombe*, *P. chrysosporium*, *U. maydis*, and *R. oryzae* being more diverged from *S. cerevisiae* than these filamentous fungi, they occur in a greater fraction of these protein clusters. This result is in agreement with the positioning *S. pombe* closer than expected to *S. cerevisiae* in the Dollo parsimony tree and the higher than expected coduplication results for *S. cerevisiae*, *S. pombe*, and *R. oryzae*. It suggests that targeted gene loss may have occurred, indicating that some non-Ascomycete species have retained some sequences found in the Saccharomycotina that have been lost in the Pezizomycotina.

Gene Ontology annotation of proteins was used to identify GO terms that are over-represented in clusters associated with Saccharomycotina (those containing 1–15 species) compared to those associated with all fungal species (containing 30–36 species). There is an inherent bias in this analysis, since proteins that do not possess orthologs beyond the Saccharomycotina are less likely to have GO annotations. For Saccharomycotina-specific clusters, GO terms associated with the fungal cell wall (GO:0009277), with Saccharomycotina-specific transcription factors (GO:0003704, GO:0016455), and components of the spindle pole body (GO:0005816) appear to be over-represented.

As Figure 2A shows, the frequency distribution of protein clusters obtained for *S. pombe* is quite different from that obtained for the budding yeasts. There is still the broad peak on the righthand side, representing proteins that are common to many or all fungal genomes, but there are no sharp peaks representing branches in the phylogenetic tree. Instead, there is a broad peak between six and 15 genomes. Further analysis shows that this is actually several overlapping peaks. Figure 2C shows the species represented in clusters containing either *S. pombe* and *S. cerevisiae*

proteins, or proteins from *S. pombe* and the filamentous fungus *G. zeae*. The results show clearly that these clusters can be divided into subgroups of budding-yeast-specific proteins and filamentous-fungi-type proteins. This implies that, although *S. pombe* split from the rest of the fungi prior to the divergence of the budding yeasts and filamentous fungi, it has retained a set of budding-yeast-type proteins (which were subsequently lost from the filamentous fungi) and a set of filamentous-fungi-type proteins (which have been lost from the budding yeasts). The GO annotations associated with proteins in these clusters were also analyzed. There are 102 clusters specific to *S. pombe* and the Saccharomycotina compared to 464 specific to *S. pombe* and the Pezizomycotina. Among the *S. pombe* and Saccharomycotina clusters, we identified four containing nuclear pore proteins (GO:0006913). There were no *S. pombe*/Pezizomycotina clusters with this annotation. Among the *S. pombe*/Pezizomycotina clusters, we identified 16 clusters associated with cellular transport (GO:0006810), 16 with DNA repair (GO:0006281), nine with meiosis, and seven associated with progression through S phase in the mitotic cell cycle (GO:0000084). Analysis of clusters containing *Y. lipolytica* proteins shows that it also contains a set of Pezizomycotina-associated proteins that have been lost from other Saccharomycotina species (see Supplemental Fig. 5).

Analysis of fatty acid β-oxidation pathways in the Fungi

To investigate the diversity and conservation of key metabolic pathways among the Fungi, protein cluster data were combined with metabolic pathway data from the LIGAND database (Goto et al. 2002). Protein clusters that contained components of metabolic pathways were then analyzed to determine which of the genomes were represented. These were subsequently mapped onto known metabolic pathways to investigate whether multiple pathway components were missing from any species.

Fatty acid β-oxidation is a process that degrades fatty acids into acetyl-CoA, which is then fed into the TCA cycle to produce ATP or (in plants and fungi) can be used to synthesize carbohydrates via the glyoxylate cycle and gluconeogenesis. Fatty acids are activated by the addition of coenzyme A, in a reaction catalyzed by fatty acyl-CoA ligases that is often concomitant with transport. Subsequent steps bring about the complete oxidation of the β-carbon and the release of acetyl-CoA. The resulting acyl-CoA, shortened by two carbon units, can undergo additional β-oxidation cycles. In mammals, β-oxidation occurs in both the mitochondria and peroxisomes (Kunau et al. 1995). One of the ways these pathways differ is in the enzyme that catalyzes the initial reaction in each round of acyl-CoA oxidation. In the mitochondrial pathway, this is catalyzed by acyl-CoA dehydrogenase (EC 1.3.99.3), with electrons produced from this reaction being passed into the electron transport chain to produce ATP. In the peroxisomal pathway, this initial reaction is catalyzed by acyl-CoA oxidase (EC 1.3.3.6), with electrons being passed directly to molecular oxygen producing hydrogen peroxide, which is detoxified by catalase. In yeasts, such as *S. cerevisiae*, β-oxidation occurs solely in peroxisomes (Hiltunen et al. 2003). A genetic study in *A. nidulans* suggested that, in this filamentous Ascomycete, β-oxidation occurs in both peroxisomes and mitochondria, the latter being specific for short-chain fatty acids (Maggio-Hall and Keller 2004). A third type of β-oxidation pathway has been discovered in another filamentous Ascomycete, *N. crassa*. This pathway occurs in the glyoxysome (a microbody that is distinct from peroxisomes, and which contains the enzymes of the glyoxylate cycle). In this pathway, the initial reaction in each round of β-oxidation is catalyzed by acyl-CoA dehydrogenase (Thieringer and Kunau 1991). Recent studies suggest that *N. crassa* may lack peroxisomes (Schliebs et al. 2006).

By surveying the available fungal genomes for the presence or absence of acyl-CoA dehydrogenases and acyl-CoA oxidases, it can be established whether each organism has peroxisomal or non-peroxisomal pathways for the β-oxidation of fatty acids (Fig. 3). The results suggest that all members of the class Saccharomycetes, except *Y. lipolytica*, possess only the peroxisomal β-oxidation pathway. The *S. cerevisiae* genome contains only one acyl-CoA oxidase gene, which encodes an enzyme that can accept fatty acid chains of differing lengths (Hiltunen et al. 2003). This is also the case with *C. glabrata*, *K. waltii*, *K. lactis*, *E. gossypii*, and other *Saccharomyces* species. Like *N. crassa*, two other species of filamentous ascomycete (*M. grisea* and *Trichoderma reesei*) lack a predicted peroxisomal β-oxidation pathway. All other species of filamentous ascomycetes, the dimorphic species *Y. lipolytica*, as well as the Basidiomycetes, Zygomycetes, and Oomycetes have both peroxisomal and non-peroxisomal β-oxidation pathways.

It is striking, however, that the fission yeast *S. pombe* and the obligate intracellular parasite *E. cuniculi* do not seem to possess either β-oxidation pathway. This observation has been confirmed by the lack of a homolog of the *S. cerevisiae FOX2* gene (which encodes a multifunctional β-oxidation protein) in the genomes of either of these two fungi. The lack of β-oxidation has been noted before in the case of *E. cuniculi* and is not surprising, since this unicellular eukaryote does not possess mitochondria or peroxisomes (Katinka et al. 2001), but the observation that the well-studied fission yeast *S. pombe* also lacks this pathway is a novel finding that has emerged from our comparative genomic analysis. In order to search for further evidence of *S. pombe* β-oxidation, we identified 72 protein clusters containing Pfam motifs associated with acyl-CoA dehydrogenases (PF00441, PF08028, PF02770, and PF02771). None of these clusters contains a *S. pombe* protein.

The genome of the dimorphic yeast *Y. lipolytica* contains six acyl-CoA oxidase-encoding genes. Five have been cloned and their products shown to have different substrate chain-length preferences (Wang et al. 1999). This fact is not unreasonable in light of the variety of hydrocarbons and fatty acids this yeast can use as a source of carbon. The two species of *Phytophthora* also contain large numbers of genes encoding acyl-CoA oxidases; these are likely to have different substrate specificities. Likewise the genomes of the filamentous Ascomycetes, Basidiomycetes, and Zygomycetes, as well as the yeasts *C. albicans*, *Candida lusitaniae*, and *D. hansenii*, all possess multiple genes for such enzymes.

The fact that mammals and many fungi possess both peroxisomal and non-peroxisomal β-oxidation pathways suggests that, during the evolution of some species of fungi, one or both of these pathways have been lost. Relating this to the fungal phylogeny generated in this study (Fig. 1), the evidence suggests that non-peroxisomal β-oxidation was lost after the divergence of the ancestor of *Y. lipolytica* from the rest of the Saccharomycetes. Strictly peroxisomal β-oxidation seems to have been lost after the divergence of the ancestor of *N. crassa* and *M. grisea* from the ancestor of *G. zeae*, as well as in the lineage of *T. reesei*, and may have been replaced with the glyoxysomal derivative of this metabolic pathway. Interestingly, recent evidence has shown that *M. grisea* still requires the peroxisomal biogenesis machinery to carry out fatty acid β-oxidation; therefore, the compartments

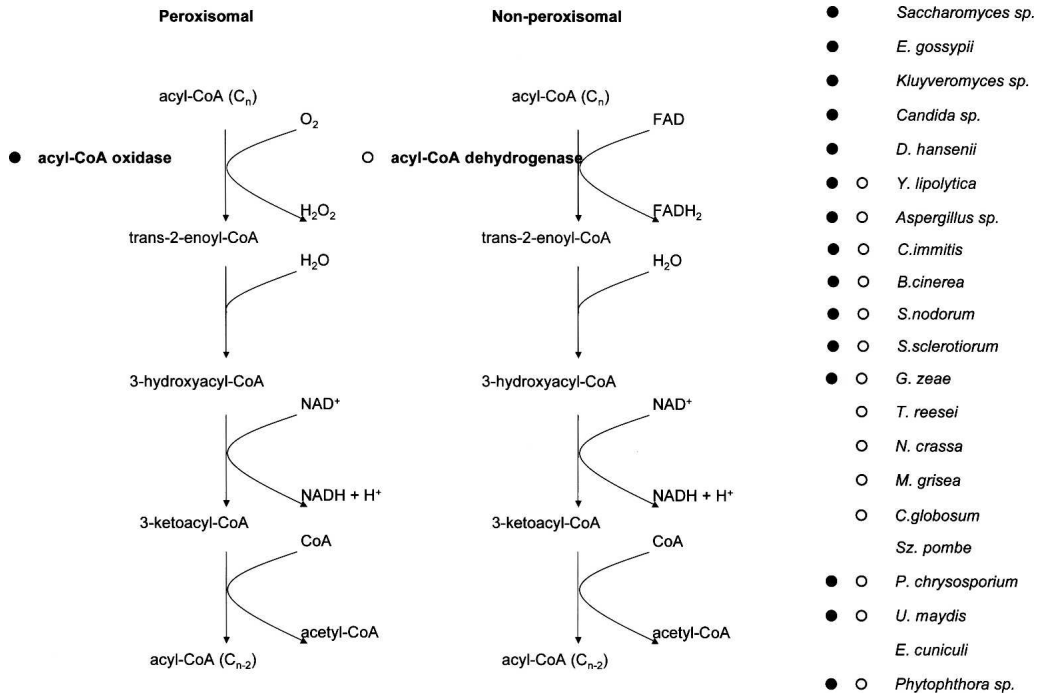

Figure 3. The occurrence of peroxisomal/non-peroxisomal pathways for beta-oxidation of fatty acids in fungal and oomycete species. The first step in peroxisomal beta-oxidation is catalyzed by acyl-CoA oxidase. The first step in non-peroxisomal beta-oxidation is catalyzed by acyl-CoA dehydrogenase. The occurrence of each pathway in fungal and oomycete species was defined based on the identification of acyl-CoA oxidase and acyl-CoA dehydrogenase encoding enzymes in the genome of each species. Acyl-CoA oxidases were identified by homology with Pox1 from *S. cerevisiae* and acyl-CoA dehydrogenases by homology with an acyl-CoA dehydrogenase from *N. crassa* (NCBI accession no. CAD70984.1).

are highly conserved (Bhambra et al. 2006; Ramos-Pamplona and Naqvi 2006). Indeed, evidence from *N. crassa* shows that orthologs of peroxisomal biogenesis proteins are necessary for formation of glyoxysomes (Managadze et al. 2007).

Discussion

Comparative genomics is a powerful technique for identifying orthologous genes in different species (Tatusov et al. 1997), for inferring the presence of metabolic pathways in organisms in which they have not been discovered previously (Giles et al. 2003), and for defining transcription-factor-binding sites and other regulatory motifs within DNA sequences (Cliften et al. 2001; Kellis et al. 2003). The availability of complete genome sequences for an increasingly large number of organisms has generated a need for techniques that enable the analysis of gene action and interaction and that are as comprehensive as the genome sequencing efforts that preceded them. This has led to the development of the different levels of functional genomic analysis—the transcriptome, proteome, and metabolome (Oliver 2000), as well as high-throughput approaches for determining phenotypes (Oliver 1996, 1997; Winzeler et al. 1999; Giaever et al. 2002). While the full range of functional genomic technologies is currently being applied to a few model organisms, of which the fungus *S. cerevisiae* is in the vanguard, this situation is changing rapidly. For instance, DNA microarrays are starting to be fabricated for many less well-studied organisms (e.g., *A. nidulans*; Sims et al. 2005). In this situation, there is a need to carry out comparative genomic analyses that exploit not just the genome sequences, but also the burgeoning amount of functional genomic data that are becoming available for a large number of species.

We have chosen to construct a bioinformatics platform for carrying out comparative functional genomics with the Fungi for several reasons. The Fungi are a group of organisms of great scientific, medical, agricultural, and industrial importance. There is a wealth of genome sequence and EST data for the Fungi, including for important model organisms (e.g., *S. cerevisiae*, *S. pombe*, and *N. crassa*), human (e.g., *C. albicans*, *A. fumigatus*), and plant (e.g., *M. grisea, U. maydis*) pathogens, as well as major process organisms in industry (e.g., *A. niger*), and others that play major ecological roles in the recycling of plant materials (e.g., *P. chrysosporium*). Moreover, techniques for the genetic manipulation of many of these fungi have been developed and are rapidly improving in both efficiency and scope. Finally, it is only in the Fungi that it is possible, at the moment, to perform comparative genomic analyses over an entire eukaryotic kingdom. The fungal kingdom shows great diversity, and thus comparisons at both short and long evolutionary distances may be made. This holds the prospect that the results of our analyses, and the tools that we develop to perform them, will have both general biological interest and utility as further eukaryotic genome sequences are generated.

The *e-Fungi* data warehouse is an object-oriented data warehouse based on the Genome Information Management System (GIMS) architecture (Cornell et al. 2003) that we developed previously to enable integrative analyses of genome sequence and functional genomic data for *Saccharomyces cerevisiae*. It is available online at http://www.e-fungi.org.uk/database.html and provides >90 detailed querying forms (so-called "canned" queries) to enable users to perform their own investigations. This multispe-

cies warehouse is a work in progress, and is currently (July 2007) populated with sequence and functional data for 34 fungal species and two oomycetes.

The Fungi are not only an evolutionarily diverse group, but also their genomes span a size range of some 18-fold from the smallest (that of the microsporidian *E. cuniculi*) to the largest (the Zygomycete, *R. oryzae*). Our whole-genome-based phylogenetic analysis demonstrates that the Oomycetes are only very distant relatives of the fungi, as previously predicted from SSU rDNA analysis. Oomycetes diverged from the most recent common ancestor of both groups much earlier than the split between the fungi and the plants. Moreover, it is interesting to note that, despite mycology being the traditional province of departments of botany, animals are more closely related to the true fungi than plants.

For all this, the influence of overall size on the construction of whole-genome phylogenies should not be overlooked. Differences in total gene numbers complicate the construction of whole-genome phylogenies. *E. cuniculi* could only be placed in the coarse-grained tree, based on 12 protein families (Fig. 1A), since we could not find sufficient representatives in this genome of the 29 families used in the higher-resolution tree (Fig. 1B). Yet other problems associated with total gene numbers arise when the gene presence/absence criterion of the Dollo parsimony approach is used. Here, there were difficulties in the resolution of the *Saccharomyces* "sensu stricto" clade, which were the consequence of the major differences in gene annotation between two different sequencing centers, despite their very similar sequence coverage. This emphasizes that the genomes chosen as the model for the construction of microarrays for sequence comparisons using the CGH approach (Edwards-Ingram et al. 2004) must be comprehensively sequenced, and accurately and uniformly annotated.

The relative sizes of genomes reflects, at least in part, differences in the levels of gene duplication between species—a contention that is confirmed by the fact that *E. gossypii* has both the smallest genome and the lowest level of gene duplication of any of the Saccharomycotina. In general, among the Ascomycetes, the filamentous Pezizomycotina have greater levels of gene duplication than do the yeast-like Saccharomycotina. However, it appears that the levels of duplication in the genome of *R. oryzae* far exceed those in any of the fungal species in this study. As well as differences in the amounts of gene duplication there appear to be differences in the loss of duplicates. Analysis of sequence similarities within Pezizomycotina species suggests differences in the amount of RIP activity, and there appear to be differences in gene loss following the WGD in the genomes of *S. cerevisiae* and *C. glabrata*. We have also demonstrated that there is evidence of coevolution, as measured by coduplication for *S. cerevisiae* and *S. pombe* and, unexpectedly, for *S. cerevisiae* and *R. oryzae*.

Protein clusters that are expanded in the Pezizomycotina, in comparison to the Saccharomycotina, include various families of transport proteins, proteins involved in the utilization of different carbon sources, and transcription factors. All of this reflects the greater metabolic versatility and broader substrate range of the filamentous Ascomycetes compared to the yeasts. An exception to this general rule is the dimorphic Saccharomycotina *Y. lipolytica,* which shares a number of gene families with the Pezizomycotina that have been lost from other members of the Saccharomycotina.

Protein clusters that are expanded in the Saccharomycotina, as compared to the Pezizomycotina, are (predictably) rare, and the examples that we found are involved in cell wall structure. This may reflect the very different chemical composition of the cell walls of yeasts and fungal hyphae. This point is emphasized by the fact that the genomes of both filamentous Ascomycetes and Basidiomycetes have expanded families of chitosanases—chitin is the principal polysaccharide of hyphal cell walls, and chitosan is found at discrete stages of the development of most fungi, but both are only minor components of yeast cell walls.

The representation of the different fungal species among the various protein clusters that we have defined is instructive. The protein products of essential genes of *S. cerevisiae* (Giaever et al. 2002), for instance, belong to clusters that contain representatives from species across the whole range of the fungal kingdom. Very few essential genes are peculiar to the *Saccharomyces* "sensu stricto" clade. However, the few examples of proteins with essential functions that are found in clusters whose members are exclusive to the *Saccharomyces* "sensu stricto" are all involved in chromosome segregation and nuclear fusion. This clade-specific set of proteins may therefore contribute to the reproductive isolation of the *Saccharomyces* "sensu stricto" from the other members of the genus (the *Saccharomyces* "sensu lato") with whom they cannot mate. It will be interesting to see if a similar clade-specific gene set is revealed when genome sequences for interbreeding species within the *Saccharomyces* "sensu lato" (Marinoni et al. 1999) become available.

The study has also highlighted the metabolic diversity of fungi reflected in the gene inventories associated with particular metabolic processes such as fatty acid β-oxidation, which appears to be absent as a process in the Taphrinomycotina *S. pombe*, and compartmentalized distinctly in the budding yeasts and filamentous fungi. This demonstrates that comparative genomic analysis can reveal new fundamental characteristics of even well-studied fungal species such as *S. pombe*.

In summary, the *e-Fungi* data warehouse has been designed to provide an e-science framework for analysis of the diverse functional genomic information currently being generated for the Fungi. Using the transcriptional profiling, proteomic, and deletion-mutant data that have been extensively compiled for *S. cerevisiae* as a reference set, we aim to build a resource into which newly acquired data from filamentous fungi, including pathogenic species and industrially relevant fungi, can be integrated. The ability to carry out extensive, multifaceted querying is a key feature of the data warehouse, which is being implemented using e-science protocols that harness data stored remotely in distinct databases, allowing them to be interrogated at a single site. This study has provided the first evidence of the utility of such a resource for investigating fundamental features of the fungi and exploring their genetic and genomic relatedness.

Methods

Orthology assignments

Predicted protein sequences for 36 genomes were downloaded from respective sequencing project repositories (see Table 1). All-against-all BLAST searches (Altschul et al. 1990) were performed to reveal the similarities among 348,995 predicted protein sequences, identifying 47,342,483 similarities below the maximum *E*-value 1×10^{-5}. OrthoMCL clustering (Li et al. 2003) was applied with default parameters, generating 30,084 clusters, including 5406 containing only paralogs.

Formulating data sets for fungal species tree

Additional homology searches were performed using CHASE (Alam et al. 2004) to select universal (i.e., present in all species) protein clusters. Two data sets were formulated: (1) To get a better rooting of the fungal phylogenetic tree, *H. sapiens* and *A. thaliana* sequences were collected for 12 universal fungal protein clusters from the HomoloGene orthologs databases of the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/HomoloGene) using *S. cerevisiae* as an intermediate. (2) To resolve deep branches of the tree, an additional 17 protein clusters were selected that had the lowest number of duplicates.

Phylogenetic analysis procedure to generate fungal species trees

Multiple sequence alignments were generated for universal protein families using ClustalW (Thompson et al. 1994). Protein sequences were concatenated, and columns with >50% gaps were excluded from the concatenated alignment. The alignments of concatenated sequences were used for phylogenetic tree construction using PhyML (Guindon and Gascuel 2003), which is a fast heuristic algorithm for estimating large phylogenies by maximum likelihood. We used 100 bootstrap replicates to assess confidence in the inferred relationships and assign bootstrap support levels to each clade in the maximum likelihood tree. We used a JTT model of amino acid evolution (Jones et al. 1992) with a GTR+Γ+I model of evolution with four discrete-Gamma rate categories (Yang 1994). The resulting trees were visualized using MEGA (Kumar et al. 2004). We initially used PhyML to evaluate bootstrap support, but it was observed that both maximum likelihood and Bayesian methods were prone to local maxima problems. We therefore selected the top trees obtained by PhyML in all bootstrap samples and exhaustively computed each tree's likelihood for each bootstrap sample, in order to avoid problems with PhyML not finding the maximum likelihood tree in each case.

Dollo parsimony analyses

The presence and absence of 18,656 protein clusters containing proteins from Ascomycete, Basidiomycete, and Zygomycete were converted into a 1/0 matrix for Dollo parsimony analysis. The seqboot program from the Phylip package (Felsenstein 1989) was used to generate 100 Bootstrap data sets based on phyletic profiles of 23,230 protein clusters. This data set was subjected to Phylip's dollop program. Phylip's consense program was used to compute the consensus tree.

Pfam domain assignments

Protein sequences were scanned using Pfam database release 18 (Finn et al. 2006) with hmmpfam (Eddy 1998) and an *E*-value cutoff of 1×10^{-1}.

Identification of Pfam domains expanded in the Pezizomycotina and Saccharomycotina

Pfam domains were identified using chi-squared analysis, assuming that domains would be equally distributed between Saccharomycotina and Pezizomycotina. For domains with significant χ^2 scores, we checked to ensure that the expansion was not limited to a single species.

Identification of transposon sequences in fungal species

Because most of the genomes analyzed in this study have not been fully annotated, we identified transposon sequences by the presence of their characteristic Pfam motifs. A search of the Pfam database (http://pfam.sanger.ac.uk) allowed us to identify 48 such domains (see Supplemental Table 3).

Identification of over-represented GO terms

The GO terms associated with proteins clusters were counted, with each term being counted only once per cluster to remove any bias toward clusters containing large numbers of paralogs. For sets of clusters, the GO terms were compared using chi-squared analysis, and (for the sake of brevity) only the most significant clusters are discussed in this paper.

Acknowledgments

We thank the Broad Institute and the DOE Joint Genome Institute for releasing data ahead of publication. We acknowledge the financial support of the Biotechnology and Biological Sciences Research Council (BBSRC). The development of the *e-Fungi* database has been funded by the BBSRC Bioinformatics and e-Science programme II.

References

Alam, I., Fuellen, G., Rehmsmeier, M., and Dress, A. 2004. Comparative homology agreement search: An effective combination of homology-search methods. *Proc. Natl. Acad. Sci.* **101:** 13814–13819.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290:** 972–977.

Bhambra, G.K., Wang, Z.-Y., Soanes, D.M., Wakley, G.E., and Talbot, N.J. 2006. Peroxisomal carnitine acetyl transferase is required for elaboration of penetration hyphae during plant infection by *Magnaporthe grisea*. *Mol. Microbiol.* **61:** 46–60.

Bowman, S.M. and Free, S.J. 2006. The structure and synthesis of the fungal cell wall. *Bioessays* **28:** 799–808.

Castillo, L., Martinez, A.I., Garcera, A., Elorza, M.V., Valentin, E., and Sentandreu, R. 2003. Functional analysis of the cysteine residues and the repetitive sequence of *Saccharomyces cerevisiae* Pir4/Cis3: The repetitive sequence is needed for binding to the cell wall beta-1,3-glucan. *Yeast* **20:** 973–983.

Cliften, P.F., Hillier, L.W., Fulton, L., Graves, T., Miner, T., Gish, W.R., Waterston, R.H., and Johnston, M. 2001. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11:** 1175–1186.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.

Cornell, M., Paton, N.W., Hedeler, C., Kirby, P., Delneri, D., Hayes, A., and Oliver, S.G. 2003. GIMS: An integrated data storage and analysis environment for genomic and functional data. *Yeast* **20:** 1291–1306.

de Boer, W., Folman, W.B., Summerbell, R.C., and Boddy, L. 2005. Living in a fungal world: The impact of fungi on soil bacterial niche development. *FEMS Microbiol. Rev.* **29:** 795–811.

Deng, J., Carbone, I., and Dean, R.A. 2007. The evolutionary history of cytochrome P450 genes in four filamentous Ascomycetes. *BMC Evol. Biol.* **7:** 30. doi: 10.1186/1471-2148-7-30.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E., et al. 2004. Genome evolution in yeasts. *Nature* **430:** 35–44.

Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755–763.

Edwards-Ingram, L.C., Gent, M.E., Hoyle, D.C., Hayes, A., Stateva, L.I., and Oliver, S.G. 2004. Comparative genomic hybridization provides new insights into the molecular taxonomy of the *Saccharomyces* "sensu stricto" complex. *Genome Res.* **14:** 1043–1051.

Elliott, S., Knop, M., Schlenstedt, G., and Schiebel, E. 1999. Spc29p is a component of the Spc110p subcomplex and is essential for spindle pole body duplication. *Proc. Natl. Acad. Sci.* **96:** 6205–6210.

Farris, J.S. 1977. Phylogenetic analysis under Dollo's Law. *Syst. Zool.* **26:** 77–88.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27:** 401–410.

Felsenstein, J., 1989. PHYLIP—Phylogeny inference package (version 3.2) *Cladistics* **5:** 164–166.
Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., et al. 2006. Pfam: Clans, web tools and services. *Nucleic Acids Res.* **34:** D247–D251. doi: 10.1093/nar/gkj149.
Fitz-Gibbon, S.T. and House, C.H. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* **27:** 4218–4222.
Fitzpatrick, D.A., Logue, M.E., Stajich, J.E., and Butler, G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol. Biol.* **6:** 99. doi: 10.1186/1471-2148-6-99.
Galagan, J.E. and Selker, E.U. 2004. RIP: The evolutionary cost of genome defense. *Trends Genet.* **20:** 417–423.
Galagan, J.E., Henn, M.R., Ma, L.J., Cuomo, C.A., and Birren, B. 2006. Genomics of the fungal kingdom: Insights into eukaryotic biology. *Genome Res.* **15:** 1620–1631.
Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418:** 387–391.
Giles, P.F., Soanes, D.M., and Talbot, N.J. 2003. A relational database for the discovery of genes encoding amino acid biosynthetic enzymes in pathogenic fungi. *Comp. Funct. Genomics* **4:** 4–15.
Goffeau, A. 2004. Evolutionary genomics: Seeing double. *Nature* **430:** 25–26.
Goto, S., Okuno, Y., Hattori, M., Nishioka, T., and Kanehisa, M. 2002. LIGAND: Database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30:** 402–404.
Grandin, N., Damon, C., and Charbonneau, M. 2001. Ten1 functions in telomere end protection and length regulation in association with Stn1 and Cdc13. *EMBO J.* **20:** 1173–1183.
Griffith, J.K., Baker, M.E., Rouch, D.A., Page, M.G.P., Skurray, R.A., Paulsen, I.T., Chater, K.F., Baldwin, S.A., and Henderson, P.J.F. 1992. Membrane transport proteins: Implications of sequence comparisons. *Curr. Opin. Cell Biol.* **4:** 684–695.
Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52:** 696–704.
Hedges, S.B., Blair, J.E., Venturi, M.L., and Shoe, J.L. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol. Biol.* **4:** 2. doi: 10.1186/1471-2148-4-2.
Hiltunen, J.K., Mursula, A.M., Rottensteiner, H., Wierenga, R.K., Kastaniotis, A.J., and Gurvitz, A. 2003. The biochemistry of peroxisomal beta-oxidation in the yeast *Saccharomyces cerevisiae*. *FEMS Microbiol. Rev.* **27:** 35–64.
Hughes, A.L. and Friedman, R. 2003. Parallel evolution by gene duplication in the genomes of two unicellular fungi. *Genome Res.* **13:** 1259–1264.
Huson, D.H. and Steel, M. 2004. Phylogenetic trees based on gene content. *Bioinformatics* **20:** 2044–2049.
Ikeda, K., Nakayashiki, H., Kataoka, T., Tamba, H., Hashimoto, Y., Tosa, Y., and Mayama, S. 2002. Repeat-induced point mutation (RIP) in *Magnaporthe grisea*: implications for its sexual cycle in the natural field context. *Mol. Microbiol.* **45:** 1355–1364.
James, T.Y., Kaulf, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J., Celio, G., Guedin, C., Fraker, E., and Miadlikowska, J. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443:** 818–822.
Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8:** 275–282.
Jones, T., Federspiel, N.A., Chibana, H., Dungan, J., Kalman, S., Magee, B.B., Newport, G., Thorstenson, Y.R., Agabian, N., Magee, P.T., et al. 2004. The diploid genome sequence of *Candida albicans*. *Proc. Natl. Acad. Sci.* **101:** 7329–7334.
Jungwirth, H. and Kuchler, K. 2006. Yeast ABC transporters—A tale of sex, stress, drugs and aging. *FEBS Lett.* **580:** 1131–1138.
Kapteyn, J.C., Van Egmond, P., Sievi, E., Van Den Ende, H., Makarow, M., and Klis, F.M. 1999. The contribution of the *O*-glycosylated protein Pir2p/Hsp150 to the construction of the yeast cell wall in wild-type cells and beta 1,6-glucan-deficient mutants. *Mol. Microbiol.* **31:** 1835–1844.
Katinka, M.D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., et al. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414:** 450–453.
Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.
Kellis, M., Birren, B.W., and Lander, E.S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428:** 617–624.
Kumar, S., Tamura, K., and Nei, M. 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5:** 150–163.
Kunau, W.H., Dommes, V., and Schulz, H. 1995. Beta-oxidation of fatty acids in mitochondria, peroxisomes, and bacteria: A century of continued progress. *Prog. Lipid Res.* **34:** 267–342.
Li, L., Stoeckert, C.J., and Roos, D.S. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13:** 2178–2189.
Lynch, M. and Force, A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154:** 459–473.
Maggio-Hall, L.A. and Keller, N.P. 2004. Mitochondrial beta-oxidation in *Aspergillus nidulans*. *Mol. Microbiol.* **54:** 1173–1185.
Maiden, M.C., Davis, E.O., Baldwin, S.A., Moore, D.C., and Henderson, P.J. 1987. Mammalian and bacterial sugar transport proteins are homologous. *Nature* **325:** 641–643.
Managadze, D., Würtz, C., Sichting, M., Niehaus, G., Veenhuis, M., and Rottensteiner, H. 2007. The peroxin PEX14 of *Neurospora crassa* is essential for the biogenesis of both glyoxysomes and Woronin bodies. *Traffic* **8:** 687–701.
Marinoni, G., Manuel, M., Petersen, R.F., Hvidtfeldt, J., Sulo, P., and Piskur, J. 1999. Horizontal transfer of genetic material among *Saccharomyces* yeasts. *J. Bacteriol.* **181:** 6488–6496.
Money, N.P., Davis, C.M., and Ravishankar, J.P. 2004. Biomechanical evidence for convergent evolution of the invasive growth process among fungi and oomycete water molds. *Fungal Genet. Biol.* **41:** 872–876.
Montiel, M.D., Lee, H.A., and Archer, D.B. 2006. Evidence of RIP (repeat-induced point mutation) in transposase sequences of *Aspergillus oryzae*. *Fungal Genet. Biol.* **43:** 439–445.
Narasimhan, M.L., Lee, H., Damsz, B., Singh, N.K., Ibeas, J.I., Matsumoto, T.K., Woloshuk, C.P., and Bressan, R.A. 2003. Overexpression of a cell wall glycoprotein in *Fusarium oxysporum* increases virulence and resistance to a plant PR-5 protein. *Plant J.* **36:** 390–400.
Naumov, G.I., Naumova, E.S., and Sniegowski, P.D. 1997. Differentiation of European and Far East Asian populations of *Saccharomyces paradoxus* by allozyme analysis. *Int. J. Syst. Bacteriol.* **47:** 341–344.
Naumov, G.I., Naumova, E.S., Masneuf, I., Aigle, M., Kondratieva, V.I., and Dubourdieu, D. 2000. Natural polyploidization of some cultured yeast *Saccharomyces sensu stricto*: Auto- and allotetraploidy. *Syst. Appl. Microbiol.* **23:** 442–449.
Oliver, S.G. 1996. From DNA sequence to biological function. *Nature* **379:** 597–600.
Oliver, S.G. 1997. From gene to screen with yeast. *Curr. Opin. Genet. Dev.* **7:** 405–409.
Oliver, S. 2000. Guilt-by-association goes global. *Nature* **403:** 601–603.
Pereira, G., Grueneberg, U., Knop, M., and Schiebel, E. 1999. Interaction of the yeast gamma-tubulin complex-binding protein Spc72p with Kar1p is essential for microtubule function during karyogamy. *EMBO J.* **18:** 4180–4195.
Ramos-Pamplona, M. and Naqvi, N.I. 2006. Host invasion during rice-blast disease requires carnitine-dependent transport of peroxisomal acetyl-CoA. *Mol. Microbiol.* **61:** 61–75.
Schliebs, W., Würtz, C., Kunau, W.H., Veenhuis, M., and Rottensteiner, H. 2006. A eukaryote without catalase-containing microbodies: *Neurospora crassa* exhibits a unique cellular distribution of its four catalases. *Eukaryot. Cell* **5:** 1490–1502.
Sims, A.H., Gent, M.E., Lanthaler, K., Dunn-Coleman, N.S., Oliver, S.G., and Robson, G.D. 2005. Transcriptome analysis of recombinant protein secretion by *Aspergillus nidulans* and the unfolded protein response in vivo. *Appl. Environ. Microbiol.* **71:** 2737–2747.
Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.
Thieringer, R. and Kunau, W.H. 1991. The beta-oxidation system in catalase-free microbodies of the filamentous fungus *Neurospora crassa*. Purification of a multifunctional protein possessing 2-enoyl-CoA hydratase, L-3-hydroxyacyl-CoA dehydrogenase, and 3-hydroxyacyl-CoA epimerase activities. *J. Biol. Chem.* **266:** 13110–13117.
Thomarat, F., Vivares, C.P., and Gouy, M. 2004. Phylogenetic analysis of the complete genome sequence of *Encephalitozoon cuniculi* supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes. *J. Mol. Evol.* **59:** 780–791.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap

penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Tournas, V.H. and Katsoudas, E. 2005. Mould and yeast flora in fresh berries, grapes and citrus fruits. *Int. J. Food Microbiol.* **105:** 11–17.
Wang, H., LeDall, M.T., Wache, Y., Laroche, C., Belin, J.M., and Nicaud, J.M. 1999. Cloning, sequencing, and characterization of five genes coding for acyl-CoA oxidase isozymes in the yeast *Yarrowia lipolytica*. *Cell Biochem. Biophys.* **31:** 165–174.
Wheeler, R.T., Kupiec, M., Magnelli, P., Abeijon, C., and Fink, G.R. 2003. A *Saccharomyces cerevisiae* mutant with increased virulence. *Proc. Natl. Acad. Sci.* **100:** 2766–2770.
Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285:** 901–906.
Wolf, Y.I., Rogozin, I.B., Grishin, N.V., and Koonin, E.V. 2002. Genome trees and the tree of life. *Trends Genet.* **18:** 472–479.
Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39:** 306–314.

Received March 22, 2007; accepted in revised form September 17, 2007.