Assessing the Performance of Single-Copy Genes for Recovering Robust Phylogenies

G. Aguileta, S. Marthey, H. Chiapello, M.-H. Lebrun, F. Rodolphe, E. Fournier, A. Gendrault-Jacquemard, Tatiana Giraud

Syst. Biol. 57(4):613–627, 2008

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802306527

Assessing the Performance of Single-Copy Genes for Recovering Robust Phylogenies

G. AGUILETA,[1]* S. MARTHEY,[2] H. CHIAPELLO,[2] M.-H. LEBRUN,[4] F. RODOLPHE,[2] E. FOURNIER,[3]
A. GENDRAULT-JACQUEMARD,[2] AND T. GIRAUD[1]

[1]*Laboratoire Ecologie, Systématique et Evolution, Université Paris-Sud, Orsay, UMR8079, Orsay Cedex, F-91405, France; and CNRS, UMR 8079, Orsay Cedex, F-91405, France; E-mail: gabriela.aguileta@u-psud.fr (G.A.)*
[2]*Mathématique, Informatique et Génome–INRA Bâtiment 233, Domaine de Vilvert, 78350 Jouy en Josas Cedex, France*
[3]*BIOGER, UMR 1290 INRA AgroParisTech–INRA, Route de St Cyr, 78026 Versailles, France*
[4]*Microbiologie, Adaptation, Pathogénie, UMR 5240 CNRS-UCB-INSA-BCS, Bayer Cropscience, 14, rue Pierre Baizet, 69009 Lyon, France*

Abstract.— Phylogenies involving nonmodel species are based on a few genes, mostly chosen following historical or practical criteria. Because gene trees are sometimes incongruent with species trees, the resulting phylogenies may not accurately reflect the evolutionary relationships among species. The increase in availability of genome sequences now provides large numbers of genes that could be used for building phylogenies. However, for practical reasons only a few genes can be sequenced for a wide range of species. Here we asked whether we can identify a few genes, among the single-copy genes common to most fungal genomes, that are sufficient for recovering accurate and well-supported phylogenies. Fungi represent a model group for phylogenomics because many complete fungal genomes are available. An automated procedure was developed to extract single-copy orthologous genes from complete fungal genomes using a Markov Clustering Algorithm (Tribe-MCL). Using 21 complete, publicly available fungal genomes with reliable protein predictions, 246 single-copy orthologous gene clusters were identified. We inferred the maximum likelihood trees using the individual orthologous sequences and constructed a reference tree from concatenated protein alignments. The topologies of the individual gene trees were compared to that of the reference tree using three different methods. The performance of individual genes in recovering the reference tree was highly variable. Gene size and the number of variable sites were highly correlated and significantly affected the performance of the genes, but the average substitution rate did not. Two genes recovered exactly the same topology as the reference tree, and when concatenated provided high bootstrap values. The genes typically used for fungal phylogenies did not perform well, which suggests that current fungal phylogenies based on these genes may not accurately reflect the evolutionary relationships among species. Analyses on subsets of species showed that the phylogenetic performance did not seem to depend strongly on the sample. We expect that the best-performing genes identified here will be very useful for phylogenetic studies of fungi, at least at a large taxonomic scale. Furthermore, we compare the method developed here for finding genes for building robust phylogenies with previous ones and we advocate that our method could be applied to other groups of organisms when more complete genomes are available. [Ascomycota; Basidiomycota; fungi; FUNYBASE; incongruence; multigene phylogenies; topological score; tree of life; Web site; phylogenetic informativeness.]

To date, genes used to build phylogenies have mostly been chosen based on historical or practical criteria. Phylogenies have been constructed using the same genes as in previous studies or using genes that were available in the focal sets of species. Using genes whose histories include duplications, horizontal transfer, lineage sorting, or selection-based biases may result in gene/species tree discrepancies. The strength of the inference of species trees increases when multiple independent loci converge on a single answer. However, because some genes have higher phylogenetic inference power than others (Townsend, 2007), the choice of a few genes with high phylogenetic informativeness can allow the construction of robust phylogenies and minimize the amount of data needed to be sequenced.

Fungi constitute good models for phylogenomics because, to date, they are the only eukaryotic clade in which more than 30 complete genomes are available (Galagan et al., 2005). Furthermore, Fungi constitute one of the main clades of eukaryotic diversity. Roughly 80,000 species have been described but the actual number has been estimated at approximately 1.5 million (Hawksworth, 1991). Fungi play pivotal ecological roles in virtually all ecosystems, through their saprophytic, pathogenic, mutualistic, and symbiotic species. The economic impact of fungi is large, involving human and crop pathogens. Fungi are also used in food processing, biotechnology, and as sources of lifesaving antibiotics. The two major groups that have been traditionally recognized among the true fungi are the Ascomycota, including the yeasts and filamentous fungi, with several important model species (e.g., *Saccharomyces cerevisiae*, *Neurospora crassa*) and the Basidiomycota, including the conspicuous mushrooms, the rusts, and the smuts. Ascomycota and Basidiomycota have been called the Dicaryomycota (Schaffer, 1975) and have been resolved as sister taxa (e.g., Lutzoni et al., 2004; James et al., 2006). The other major groups of fungi, basal to the Dicaryomycota, include the Glomeromycota, the Zygomycota, and the Chytridimycota (James et al., 2006).

The recent availability of numerous fungal genome sequences provides large numbers of genes that could be used for building robust phylogenies. Complete fungal genomes have been successfully used to build organismal phylogenies (Rokas et al., 2003; Fitzpatrick et al., 2006; Robbertse et al., 2006). However, if we are to reconstruct phylogenetic relationships among fungal species whose complete genome is not sequenced, as is generally the case, only a limited number of DNA fragments can practically be sequenced. It would therefore be useful to determine how much data are sufficient to recover a well-resolved and correct species tree, and whether some genes better reflect whole-genome relatedness. Several factors may indeed influence the

performance of a particular gene in recovering a robust and valid phylogeny, such as its length, its rate and mode of evolution, and its demographic and selective histories. Identifying a few genes sufficient to build robust phylogenies will economize on costs and improve accuracy. It will furthermore allow homogenization of data sets, which will be useful for building the tree of life using independent studies, which have so far little overlap in their data partitions (Lutzoni et al., 2004). Identifying genes with high phylogenetic performance may also be of great interest for species bar coding (i.e., species identification based on a few DNA sequences).

The aim of this study was therefore to develop a method for finding genes with a high phylogenetic inference power. We used fungi as models because of the availability of numerous genome sequences. Our specific aims were to assess which genes recovered the optimal fungal phylogenies among the single-copy orthologs shared by most fungal genomes. Furthermore, we wished to determine how many genes are required to obtain an accurate and well-supported phylogeny. Similar approaches have been developed previously, applied to yeasts (Rokas et al., 2003) or bacteria (Konstantinidis et al., 2006); other studies have also looked for the best-performing genes, among a few available in vertebrates, but not at a whole-genome scale (Graybeal, 1994; Cummings et al., 1995; Zardoya and Meyer, 1996; Miya and Nishida, 2000; Springer et al., 2001; Mueller, 2006). These studies have found a high variability in the performance of single genes and suggested that a few genes, if appropriately chosen, may be sufficient to recover a reliable species tree. Taking advantage of the complete sequenced fungal genomes, a few previous studies have built phylogenies (Fitzpatrick et al., 2006; Kuramae et al., 2006; Robbertse et al., 2006), and one recently investigated how many and which genes were sufficient to resolve the fungal tree of life (Kuramae et al. 2007).

Here we developed an automated procedure for extracting only single-copy orthologous genes from complete fungal genomes, as the presence of paralogs may hinder correct phylogenetic reconstruction (Koonin, 2005). Maximum likelihood trees based on individual orthologous clusters and concatenated alignments were constructed. The performance of each gene was tested by assessing the congruence of every single-gene phylogeny with that of the reference tree, using three different metrics. We furthermore tested for parameters that could predict the performance of single genes in yielding a good phylogeny, such as gene size, number and proportion of variable sites, and putative function. We then compared our results and methods to those of previous studies with similar aims.

MATERIALS AND METHODS

Ortholog Search

Predicted proteins were extracted from 30 fungal genomes (Table 1). We decided to use exclusively protein sequences because Basidiomycota and Ascomycota have nucleotide sequences that are too divergent to be aligned with confidence. From predicted proteins, single-copy orthologs were identified following the method described in Dujon et al. (2004). First, an all versus all BLASTP search was performed using the NCBI BLAST2 software (Altschul et al., 1997) with the BLOSUM62 matrix and affine gap penalties of 11 (gap) + 2 (ext). Pairwise alignments were considered nonspurious after HSP tiling if they met three criteria: (i) coverage of at least 70% of the query sequence, (ii) identity of at least 30%, and (iii) E-value cutoff of 6e−6. HSP tiling was then performed using the "tile_hsp" function of the BioPerl module Bio::Search::BlastUtils (documentation at http://search.cpan.org/~birney/bioperl1.2.3/Bio/Search/BlastUtils.pm). Default values were taken for all parameters. For each hit, both the query and the subject sequences have been tiled independently and only if they came from nonoverlapping regions. If tiling is operated, the function computes the following data across all tiled HSPs: total alignment length, total identical residues, and total conserved residues. If no tiling is operated, only the values of the best HSP are kept. Note that for all tiled and nontiled hits, only the best HSP e-value is considered for MCL clustering. Therefore, HSP tiling can only increase the number of pairwise alignments that meet the criterion of at least 70% of coverage on query sequence and at least 30% of identity. HSP tiling has no impact on the clustering MCL procedure itself.

The BLAST results were then analyzed with the program Tribe-MCL (06–058 release) obtained from the Web site http://www.micans.org/mcl/ (Enright et al., 2002). The program Tribe-MCL uses Markov clustering (MCL) by creating a similarity matrix from BLAST e-values and then clustering proteins into related groups. The main parameter that influences the size of a cluster in Tribe-MCL is the inflation value, which can be adjusted from 1.1 (fewer clusters are formed but with more proteins in each) to 5.0 (more but smaller clusters are formed and proteins with high similarity remain clustered together). MCL avoids clustering fragmented proteins or domains coming from different complex multidomain proteins, both of which are very common in eukaryotic genomes, and it has been shown to outperform other cluster algorithms (Costa et al., 2005; Brohee and van Helden, 2006). In order to be conservative and to retrieve only the true orthologs present in all fungal genomes, as stringently as possible, we used the conservative inflation value of $I = 4$ (see previous analyses on the impact of the inflation parameter; Enrigth et al., 2002; Dujon et al., 2004; Robbertse et al., 2006). We also filtered clusters that contain exactly one protein per fungal genome (single-copy clusters).

The number of recovered single-copy clusters varies considerably depending on the data set used, as the inclusion of more divergent genomes will result in a lower number of orthologs shared by all. Errors in protein prediction (e.g., proteins deduced from genes identified using automatic annotation pipelines that are split, fused or with unresolved introns; Gilks et al., 2005) can also drastically reduce the number of detected orthologs common to all genomes, because some orthologs will then be

TABLE 1. The 30 complete genomes used for the search of single-copy orthologs.

Species	Source	Number of proteins	Release	Online database	Taxonomy
Ashbya gossypii [b]	AGD	4726	"2.1"	http://agd.unibas.ch/Ashbya_gossypii/	Hemi-Ascomycete
Aspergillus fumigatus[b]	NCBI	9923	07/02/2006	ftp://ftp.ncbi.nih.gov/genomes/Fungi/	Ascomycete
Aspergillus nidulans[b]	BROAD	9541	27/10/2003	http://www.broad.mit.edu/annotation/fungi/fgi/	Ascomycete
Aspergillus oryzae[b]	NITE	12,074	21/12/2005	http://www.nite.go.jp/index-e.html	Ascomycete
Botrytis cinerea	BROAD	16,448	05/10/2005	http://www.broad.mit.edu/annotation/fungi/fgi/	Ascomycete
Candida glabrata[b]	NCBI	5181	08/10/2005	ftp://ftp.ncbi.nih.gov/genomes/Fungi/	Hemi-Ascomycete
Candida lusitaniae[b]	BROAD	5941	12/01/2006	http://www.broad.mit.edu/annotation/fungi/fgi/	Hemi-Ascomycete
Chaetomium globosum	BROAD	11,124	24/06/2005	http://www.broad.mit.edu/annotation/fungi/fgi/	Ascomycete
Coccidioides immitis[b]	BROAD	10,457	23/01/2006	http://www.broad.mit.edu/annotation/fungi/fgi/	Ascomycete
Cryptococcus neoformans[b]	NCBI	6475	26/10/2005	ftp://ftp.ncbi.nih.gov/genomes/Fungi/	Basidiomycete
Debaryomyces hansenii[b]	NCBI	6893	04/08/2005	ftp://ftp.ncbi.nih.gov/genomes/Fungi/	Ascomycete
Fusarium graminearum[b]	BROAD	11,640	11/09/2003	http://www.broad.mit.edu/annotation/fungi/fgi/	Ascomycete
Kluyveromyces lactis[b]	NCBI	5331	04/08/2005	ftp://ftp.ncbi.nih.gov/genomes/Fungi/	Hemi-Ascomycete
Magnaporthe grisea[b]	BROAD	12,841	08/12/2005	http://www.broad.mit.edu/annotation/fungi/fgi/	Ascomycete
Neurospora crassa[b]	BROAD	10,620	28/02/2005	http://www.broad.mit.edu/annotation/fungi/fgi/	Ascomycete
Phanerochaete chrysosporium[b]	JGI	10,048	"1.0"	http://genome.jgi-psf.org/	Basidiomycete
Rhizopus oryzae	BROAD	17,467	29/11/2005	http://www.broad.mit.edu/annotation/fungi/fgi/	Zygomycete
Saccharomyces bayanus[b]	MIT	9424	15/12/2004	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes/S_bayanus/MIT/	Hemi-Ascomycete
Saccharomyces castellii	WashU	4677	17/12/2004	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes/S_castellii/WashU/	Hemi-Ascomycete
Saccharomyces cerevisiae[b]	NCBI	5873	23/01/2006	ftp://ftp.ncbi.nih.gov/genomes/Fungi/	Hemi-Ascomycete
Saccharomyces kluyveri	WashU	2968	04/10/2003	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes/S_kluyveri/WashU/	Hemi-Ascomycete
Saccharomyces kudriavzevi	WashU	3768	08/07/2003	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes/S_kudriavzevii/WashU/	Hemi-Ascomycete
Saccharomyces mikatae	MIT	9057	15/12/2004	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes/S_mikatae/MIT/	Hemi-Ascomycete
Saccharomyces paradoxus[b]	MIT	8955	15/12/2004	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes/S_paradoxus/MIT/	Hemi-Ascomycete
Schizosaccharomyces pombe[b]	NCBI	5093	04/08/2005	ftp://ftp.ncbi.nih.gov/genomes/Fungi/	Archeo-ascomycete
Sclerotinia sclerotiorum[b]	BROAD	14,522	03/10/2005	http://www.broad.mit.edu/annotation/fungi/fgi/	Ascomycete
Stagonospora nodorum[a]	BROAD	16,597	13/05/2005	http://www.broad.mit.edu/annotation/fungi/fgi/	Ascomycete
Trichoderma reesei[b]	JGI	9997	"1.2"	http://genome.jgi-psf.org/	Ascomycete
Ustilago maydis[b]	BROAD	6522	20/11/2003	http://www.broad.mit.edu/annotation/fungi/fgi/	Basidiomycete
Yarrowia lipolytica[b]	NCBI	6666	04/08/2005	ftp://ftp.ncbi.nih.gov/genomes/Fungi/	Hemi-Ascomycete
Mean		8890			

[a] The 23 species kept for building the reference tree. [b] The 21 genomes kept for tree comparisons.

artefactually lacking in one genome. In order to minimize the number of ortholog clusters lost by such an artefact, we investigated variation in the number of single-copy clusters by removing one fungal genome at a time from the initial complete data set (Table 1). As expected, removing highly divergent genomes (e.g., the zygomycete *Rhizopus oryzae*) from the data set resulted in an increase in the number of single-copy clusters, as more orthologs were then shared by all the remaining genomes (data not shown). In addition, removing some other genomes, although closely related to others in the data set, also increased the number of single-copy clusters, indicating that these genomes most probably contained numerous errors in protein prediction. Therefore, for the subsequent phylogenetic analysis, we kept only species that did not drastically decrease the single-copy cluster number. Following this procedure, we reduced our initial data set, which contained 30 fungal genomes, to a data set with 23 genomes (Table 1). Furthermore, because we needed a reliable reference tree to perform the topological comparisons, we had to keep only nodes that were supported and consistent with commonly accepted relationships among fungal species. We thus removed two species, *Aspergillus oryzae* and *Stagonospora nodorum*, as their placement in the 23 genome phylogeny was uncertain and had low support (see online Appendix 1; http://www.systematicbiology.org). The subsequent tree topology comparisons were thus performed using 21 genomes (Table 1).

Phylogenetic Analyses

Protein sequences in the orthologous clusters were aligned using ClustalW 1.38 (Thompson et al., 1994) with default settings. We kept only the alignment sections that were unambiguously aligned and without gaps. In order to detect such regions, we used the program Gblocks 0.91b (Castresana, 2000) with the minimum number of sequences set to 16, the minimum number of flanking positions set to 20 and no gaps allowed; otherwise, default settings were assumed. For the aligned orthologous sequences in each cluster, we determined which amino acid substitution model best fit the data using ProtTest 1.4 (Abascal et al., 2005), which estimates the likelihood of each model and the parameter values under a maximum likelihood framework. The AIC-1 criterion was used to rank the 80 different evolution models tested by ProtTest 1.4. We used the information on the chosen model and its parameters to infer the corresponding maximum likelihood gene trees with PHYML v2.4.4 (Guindon and Gascuel, 2003). A bootstrap analysis with 100 replicates was performed for each tree in order to assess the support for the nodes. The majority rule criterion was used to obtain the consensus trees from the bootstrap analysis with the program Consense in the PHYLIP 3.66 package (Felsenstein, 1989). We thus obtained 246 gene trees, one for each individual orthologous data set.

In order to obtain a reference tree, the individual alignments best fitted by the same evolutionary model were concatenated with a custom-made Perl script. From the model selection analysis performed using ProtTest, we found that almost all the orthologs in the data set were best approximated by either the WAG or the rtREV models. We therefore built two concatenated alignments, one for each group of sequences (see Results). We inferred a maximum likelihood phylogenetic tree from the two concatenated alignments, one for the WAG and one for the rtREV concatenations, following the same method that was used to infer the individual trees. In this case, we obtained 250 bootstrap replicates to determine the node support values. Also, we tested a mixture-model approach under a Bayesian framework to infer the reference tree, without assuming an a priori model, and compared the resulting topology with the likelihood reference trees previously obtained assuming the WAG and the rtREV models. For the mixture-models analysis we used the PhyloBayes 2.3 package (Lartillot and Philippe, 2004), which implements a Bayesian Monte Carlo Markov chain (MCMC) sampler with a Dirichlet process (using the cat option). Two simultaneous chains were run for 74,390 and 74,666 generations, respectively. We used the program bcomp included in PhyloBayes2.3 to check for convergence: if the largest discrepancy across the bipartitions (maxdiff) of two or several chains is less than 0.1, this indicates a good convergence level (PhyloBayes 2.3 manual). We obtained a maxdiff value <0.1 in our chain comparisons. We used the readpb program in the PhyloBayes 2.3 package to obtain the summary and consensus trees. We tried different burnin values discarding 100, 500, and 1000 trees and we tested different sampling intervals, choosing trees every 10, 5, or 2 trees. In all cases the consensus tree obtained was the same as the likelihood trees (see Results, Fig. 1).

Tree Topology Comparisons

The congruence of every individual phylogeny with respect to the reference tree obtained by concatenation was assessed using three different indices. First, we estimated an overall topological score (Nye et al., 2006), which provides a measure of the distance between two trees in terms of topology. The algorithm used is specifically designed to compare trees produced using different genes for the same set of species, as in our case. The java applet (distributed at http://www.mrc-bsu.cam.ac.uk/personal/thomas/phylo_comparison/comparison_page.html) matches the branches in the two trees that have a similar partition of leaf nodes and finds an optimum 1-to-1 correspondence map. To obtain the overall score, first every pair of edges is assigned a score that reflects the topological similarity of the branches and then the branches in the trees are paired up to optimize the global score (Nye et al., 2006). Furthermore, we assessed the performance of each gene by visually inspecting the node support of the gene trees with topological scores higher than 90% (online Appendix 2; http://www.systematicbiology.org).

The commonly used Robinson-Foulds symmetric distance was also employed to assess the topological similarity between trees (Robinson and Foulds, 1981).

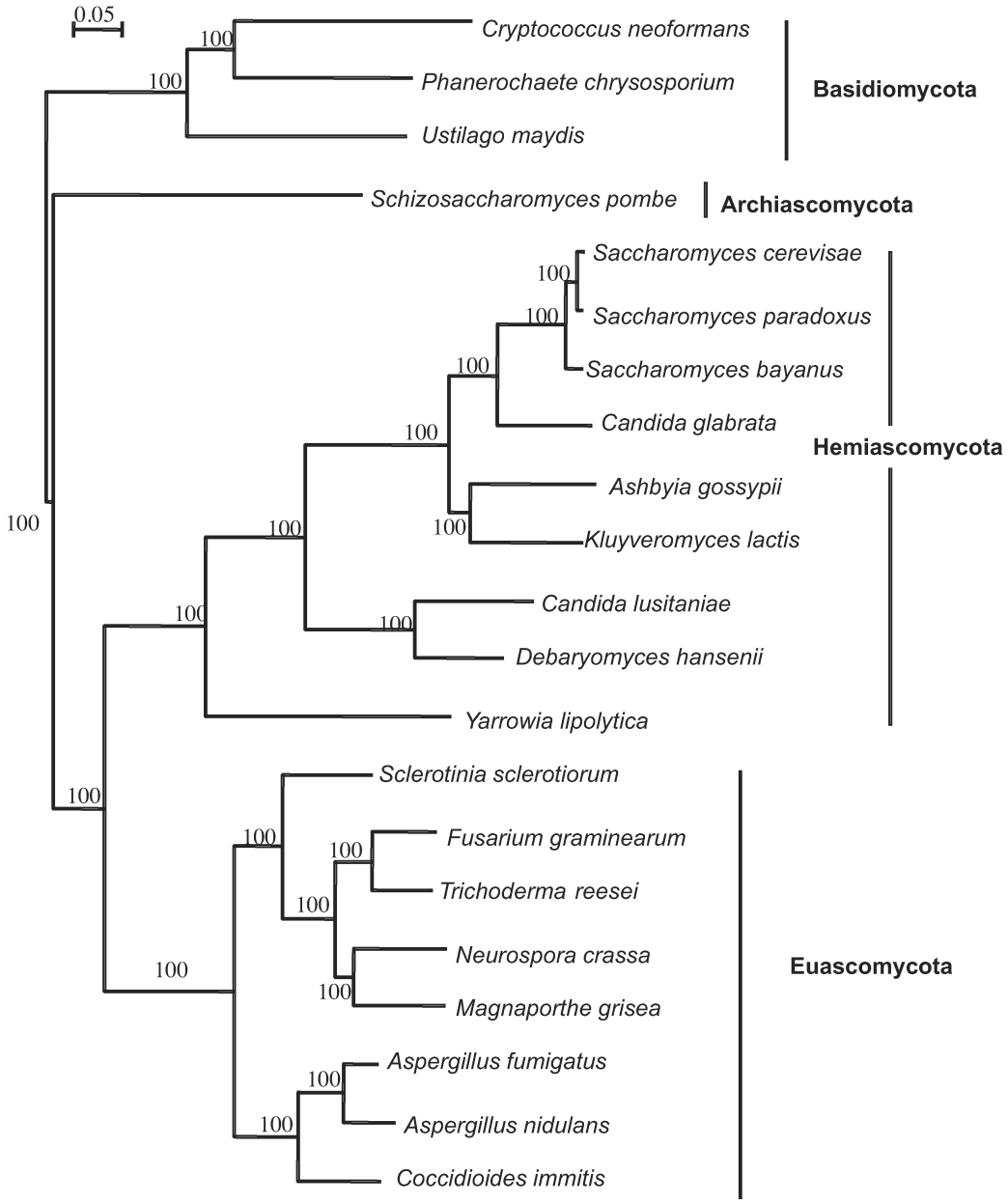

FIGURE 1. Phylogeny of the concatenated data set using the 246 single-copy orthologs extracted from 21 genomes (Table 1). Indicated supports are percentages over 250 bootstraps. Bayesian support for all nodes is also consistent with a 100/100 bootstrap value.

This metric considers all the possible branches that could exist on the two trees. Each branch divides the set of species into two groups, with one group connected to one end of the branch and the other group connected to the other end. A partition of the full set of species can be created with respect to the tree topology. For each tree, a list is created with all the partitions therein. The symmetric distance between two trees is the count of how many partitions there are among their respective lists that are not shared between the two.

Finally, the last metric we used to compare our trees was the Kuhner-Felsenstein distance (Kuhner and Felsenstein, 1994). Unlike the two previous methods, the latter distance takes into account the branch lengths in the trees. The method starts with all the possible partitions in each tree, then assigns a value, which is 0 if the partition is absent from the tree, or the length of the branch if the partition is present in the tree. The distance, called the branch score, corresponds to the sum of squares of differences between the lists of two trees. If the same branch is present in the two trees, it will contribute the square of the difference between the two branch lengths. A branch that is absent from both trees contributes zero to the total sum.

Statistical Analyses

Statistical tests were performed with the JMP 5 software (SAS Institute 1995). Because the Arcsin-square-root-transformed topological score did not significantly deviate from normality (Shapiro-Wilk W test), we used an ANOVA to test the effect of the number of variable sites, the gene size and the percentage of variable sites on the topological score. For the chi-square test on the 19 GO categories, we pooled those containing a few genes in order to have at least five genes per category. Gene size (i.e., alignment length) was highly correlated with the number of variable sites ($r = 0.959$; $P < 0.00001$); they were therefore analyzed in separate ANOVAs.

We also estimated the average evolutionary rate of the individual 246 genes, with and without assuming a molecular clock, using PAML 4_OSX_Intel (Yang, 1997). This approach was undertaken to investigate whether the best-performing genes evolved with more clock-like rates of evolution than other less informative genes. We used the program codeml in the PAML package to estimate the average substitution rate across the alignments. For all genes, we inferred the evolutionary rates in two ways: (i) by assuming a global molecular clock (i.e., all branches in the tree have the same rate), and (ii) by allowing each branch in the tree to have an independent evolutionary rate. The two competing hypotheses are nested and can thus be compared by a likelihood ratio test (LRT) with degrees of freedom equal to $n - 2$, where n is the number of species (sequences; Yang, 1997).

RESULTS

Clusters of Single-Copy Orthologs

We recovered different numbers of single-copy orthologous gene clusters for the three different data sets (containing 30, 23, and 21 genomes, respectively; Table 1). For the 30 fungal genomes, there were 275,948 predicted proteins that were compared using an all versus all BLASTP search and the subsequent Tribe-MCL analysis allowed us to identify 17,956 clusters, from which only 43 contained exactly one protein in each genome (single-copy clusters). For the data set with 23 fungal genomes, we identified 219 single-copy clusters of putative orthologous proteins, with alignments ranging from 111 to 1708 amino acids in length (mean length: 522 aa (amino acids), median length: 447 aa). The average identity within the 219 clusters ranged from 23% to 83% (mean: 52%, median: 52%). For the 21 remaining genomes, a total of 246 clusters of single-copy orthologs were retrieved. The alignments ranged in length from 111 to 2197 amino acids (mean length: 569 aa, median length: 482 aa). Average identity within the 246 clusters ranged from 24% to 84% (mean: 51%, median: 51%).

Phylogenies

Maximum likelihood and Bayesian methods produced the same topologies (not shown). We therefore present below only the trees obtained using maximum likelihood. In the case of the 23-genome data set, most of the protein-coding genes (200/219) were best fitted either by the WAG (Whelan and Goldman, 2001; 163 genes) or the rtREV (Dimmic et al., 2002; 37 genes) evolution models. We therefore concatenated all the 163 gene alignments fitted by the WAG model on one hand, and all the 37 gene alignments fitted by the rtREV model on the other hand, to infer the maximum likelihood gene tree for each of these two concatenated data sets. The two concatenated data sets yielded the same topology with similar node supports. The tree inferred from the WAG concatenated data set is shown in online Appendix 1. All the nodes were well supported and in agreement with previously published phylogenies (Fitzpatrick et al., 2006; James et al., 2006; Robbertse et al., 2006), except for the placement of *S. nodorum* and the relationships among the three *Aspergillus* species.The phylogenetic position of *S. nodorum* was also difficult to estimate with accuracy in previous studies (Fitzpatrick et al., 2006; Robbertse et al., 2006).

To understand the poor support for the placement of *S. nodorum*, we looked at all the single-gene phylogenies. Three main positions were recovered for *S. nodorum*, as indicated in online Appendix 1. The first position was found in 49 trees, the second position in 76 trees, and the third one in 73 trees. Other placements were recovered for 21 genes. We also looked at all individual gene phylogenies for the relationships among the three *Aspergillus* species. Of the 219 protein-coding genes, 203 supported the three species as monophyletic. Among these 203 genes, 84 supported the closest relationship for *A. nidulans* and *A. oryzae*, 81 genes for *A. fumigatus* and *A. oryzae*, and 38 genes for *A. nidulans* and *A. fumigatus*.

Because we needed a reliable reference tree, we removed the latter two genomes from our data set. For the 21-genome data set, out of 246 protein clusters, 116 were best approximated by the WAG model, and 122 by the rtREV model. The reference tree inferred from the concatenation of the 122 alignments fitted by the rtREV model is shown in Figure 1. The tree inferred from the concatenation of the 116 alignments fitted by the WAG model yielded exactly the same topology. All the nodes were supported with 100% bootstrap proportions and were consistent with previously published phylogenies (James et al., 2006; Fitzpatrick et al., 2006; Robbertse et al., 2006; Kuramae et al., 2006).

Phylogenetic Performance of Individual Genes

For the 246 trees obtained for the 21-genome data set, the congruence of every individual phylogeny with respect to the reference tree obtained by concatenation (Fig. 1) was assessed using three different indices: the Nye topological score, the Robinson-Foulds distance, and the Kuhner-Felsenstein distance (see Materials and Methods). All three measures were significantly correlated (Fig. 2). The topological score and the Robinson-Foulds distance in particular had a very high correlation coefficient, whereas there was more variability in the relationship between the topological score and the Kuhner-Felsenstein distance (Fig. 2b). This makes

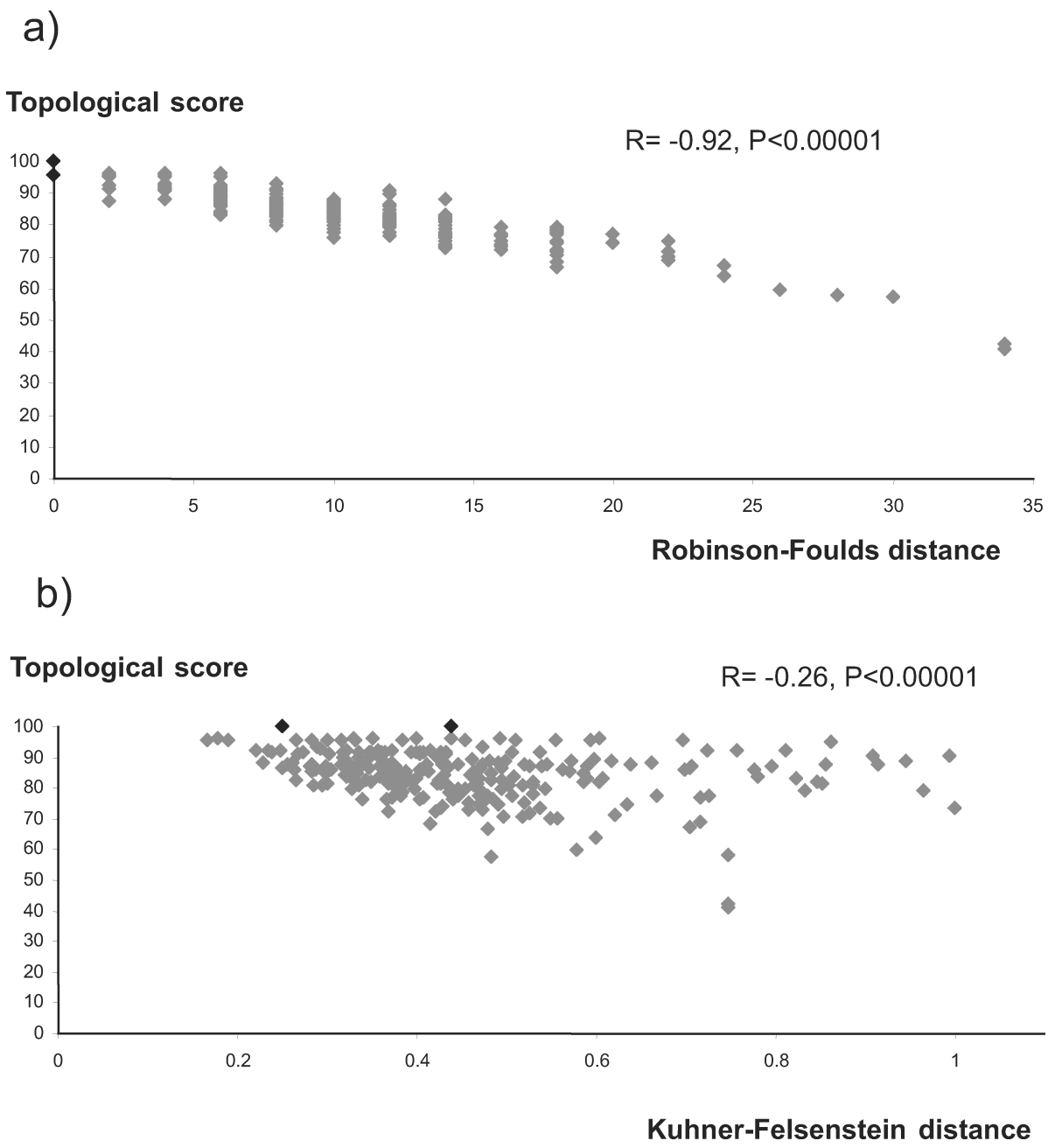

FIGURE 2. Plot of the topological score as a function of the Robinson-Foulds distance (a) and the Kuhner-Felsenstein distance (b). The two genes with a topological score of 100% are represented in black.

sense because the first two measures are based solely on topology, whereas the Kuhner-Felsenstein distance also takes branch lengths into account. The genes performing well, as assessed from the Kuhner-Felsenstein distance, also had a good topological score, but the reverse was less often true. In particular, the two genes having a topological congruence of 100% did not have the lowest Kuhner-Felsenstein distances. Because we were more interested in retrieving correct phylogenetic relationships than branch length estimates, we mainly refer to the topological score hereafter.

The distribution of the topological score for the 246 genes from the 21-genome data set approximated a Gaussian curve skewed towards the low topology scores (Fig. 3). As seen above, only two genes, MS456 and MS277, yielded exactly the same topology as the concatenated tree (see online Appendix 2 for a list of the 59 genes with a topological score at least equal to 90%, which corresponded to topologies exhibiting discrepancies with the reference tree for one to three nodes). Most of the bootstraps in these individual robust phylogenies were higher than 70%, although a few

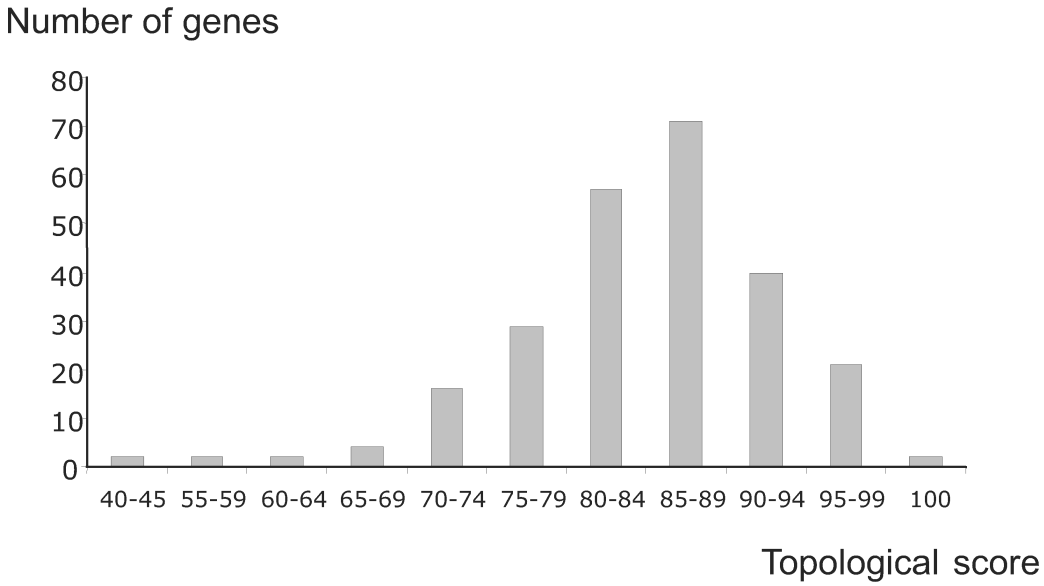

FIGURE 3. Distribution of the topological score for the 246 individual gene phylogenies with 21 species.

nodes per tree had lower support (online Appendix 2). Nodes with low support were sometimes those that differed from the reference tree, but not necessarily: some nodes that were not recovered in the reference tree could appear well-supported in single-gene trees, whereas nodes recovered in the reference tree could be poorly supported in some single-gene trees.

In order to investigate whether the good phylogenetic performance of the two genes MS465 and MS277 was dependent on the sample of species used, we computed topological scores on subsets of the 21 species. For each subset, we aligned the corresponding subset of sequences and followed the same procedure implemented for the full set analyses. We first computed topological scores on two nonoverlapping subsets of respectively 10 and 11 species, maximizing the dispersion of the chosen species in the tree in each subset. More genes having a 100% topological score were retrieved in these subsets than for the 21-species data set. The topological scores of the 246 genes were highly significantly correlated between the two nonoverlapping subsets ($r = 0.5$; $P > 0.00001$). MS456 had again a 100% topological score in the two independent subsets. MS277 also had a high topological score, although not 100% in both subsets. We then computed the topological scores on 10 subsets of 15 species. The subsets were chosen for minimizing the overlap of species among them. The distribution of the topological scores of the two best-performing genes, MS465 and MS277, was compared to the distribution of all the genes across all the subsets (Fig. 4). The distributions showed that these two genes had good phylogenetic performance compared to the other genes, independent of the sample. Some variability existed in the topological scores among the different subsets, but MS277, and especially MS456, always retained a high degree of phylogenetic power (Fig. 4). Similar results were obtained using the Robinson-Foulds and the Kuhner-Felsenstein distances (online Appendices 3 and 4; http://www.systematicbiology.org).

In order to further test whether the two genes, MS456 and MS277, can be of general utility for phylogenetic inference, we built a tree using all sequences available in GenBank that had a significant BLAST hit to either one of the query sequences. *Homo sapiens* was chosen as an outgroup. The resulting trees included 36 species (e.g., online Appendix 5 for MS456; http://www.systematicbiology.org), were well supported and consistent with phylogenetic relationships known from previous studies (James et al., 2006; Fitzpatrick et al., 2006). Only *Puccinia graminis* had a placement inconsistent with firmly established relationships, but with low bootstrap support (37%).

It is noteworthy that the protein-coding genes most commonly used in fungal phylogenies, such as the β-tubulins, the elongation factor EF-1α, the γ-actin, heat shock proteins, chitinases, chitin synthases, RNA polymerases, dehydrogenases, and histones were not found in the list of the best-performing genes (online Appendix 2). Some of the latter genes were not even present in the single-copy ortholog clusters shared by most genomes, as they appeared to have paralogs in some of them (e.g., some tubulins, elongation factors, chitinases, and dehydrogenases). We computed the topological scores for the elongation factor EF-1α, β-tubulin, and γ-tubulin. Only 20 species had orthologs of EF-1α among our 21-species data set, and we choose the best hits among the paralogs present in some species for β-tubulin and γ-tubulin. The topological scores were poor for these three genes, even when they were concatenated: 80.3% for EF-1α, 79.4% for β-tubulin, 84.6% for γ-tubulin, and 85.2% for the tree resulting from their concatenation. Among the 246 single-copy orthologs, the number of genes that had a better topological score than the widely used β-tubulin, γ-tubulin, and EF-1α genes was thus respectively 49.2%,

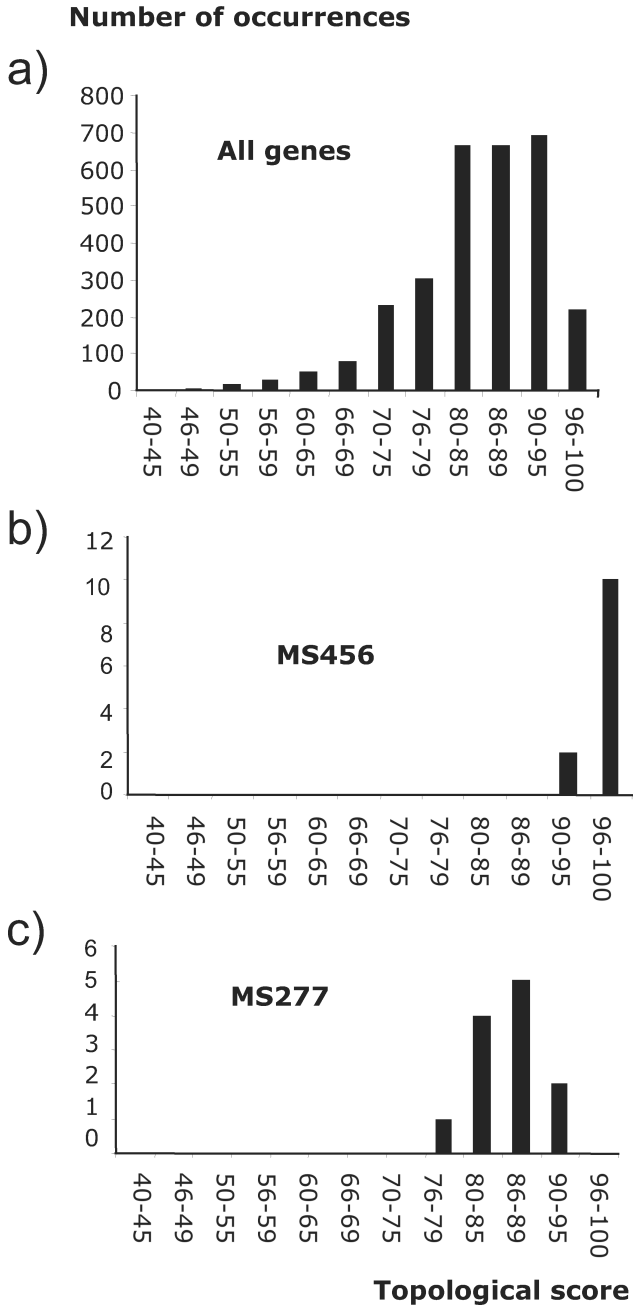

FIGURE 4. Distribution of the number of occurrences for given topological scores on 10 subsets of 15 species for the 246 individual gene phylogenies (a), and for the two genes yielding a 100% topological score: MS456 (b) and MS277 (c).

77.6%, and 53.6%. Even when combined, 50.8% of the genes had individually better topological scores.

Minimal Number of Concatenated Genes Required to Produce a Robust Phylogeny

Based on the results from the topological scores described above, we ranked the genes according to the congruence of their topologies to that of the reference tree. We concatenated 2, 3, 5, 10, 15, 20, and 25 single-gene alignments from the 21-genome data set, corresponding to the individual genes that provided the most congruent topologies with respect to the reference tree. Concatenating the two best-performing genes was sufficient to obtain a topology identical to the reference tree, with all bootstraps higher than 70%. Concatenation of the top three genes was sufficient to recover bootstrap supports higher than 80% for all the nodes in the tree. However, even with 25 genes, not all nodes had 100% bootstrap supports.

Characteristics of the Best-Performing Genes

Some of the gene ontology (GO) function categories (Ashburner et al., 2000) were significantly overrepresented in the 59 best-performing genes compared to the 246 orthologs. These overrepresented categories corresponded to biological regulation, organelle, membrane, organelle part, and binding (Fig. 5; χ^2 test, $P = 0.004$). According to the annotation as in *Saccharomyces cerevisiae*, the two best-performing genes identified in this study are (i) MS277, a protein required for processing of 20S pre-rRNA in the cytoplasm and it associates with pre-40S ribosomal particles; and (ii) MS456, a component of the hexameric MCM complex, which is important for priming origins of DNA replication (online Appendix 2).

The ANOVAs detected a significant effect of the number of variable sites ($F_{1,241} = 25.82$, $P < 0.0001$; Fig. 6) and gene size ($F_{1,241} = 28.51$, $P < 0.0001$) on the topological score. Interestingly, the two genes yielding the same topology as the concatenated data sets had intermediate numbers of variable sites and gene sizes (Fig. 6). All the genes longer than 700 base pairs yielded topological scores higher than 88%. However, there was high variability in the phylogenetic performance for a given gene size, and topological scores higher than 90% could be recovered at any gene size.

The phylogenetic informativeness of a gene is likely related to its evolutionary rate and it has been suggested that a given gene can be effectively used to solve a soft polytomy if it evolves at an optimal rate at a relevant time scale (Townsend, 2007). We tried to determine whether the genes we identified as the most phylogenetically informative (MS456 and MS277) evolved under an optimal rate (i.e., correlated with phylogentic performance) across all the branches in the phylogeny. We estimated the average substitution rate of the 246 genes, with and without assuming the molecular clock. The two models were compared by means of a LRT and the assumption of a molecular clock was rejected for all 246 genes. The average evolutionary rates computed without assuming the clock were not significantly correlated with the topological score, the Robinson-Foulds, or the Kuhner-Felsenstein distance. The average rates estimated for the 246 genes ranged between 0.62 and 2.88 expected total number of amino acid substitutions per amino acid site across the phylogeny. The average substitution rates of the two best genes were 0.88 for MS277 and 1.52 for MS456, but within the same rate interval the topological score varied considerably, from 100% for the best genes

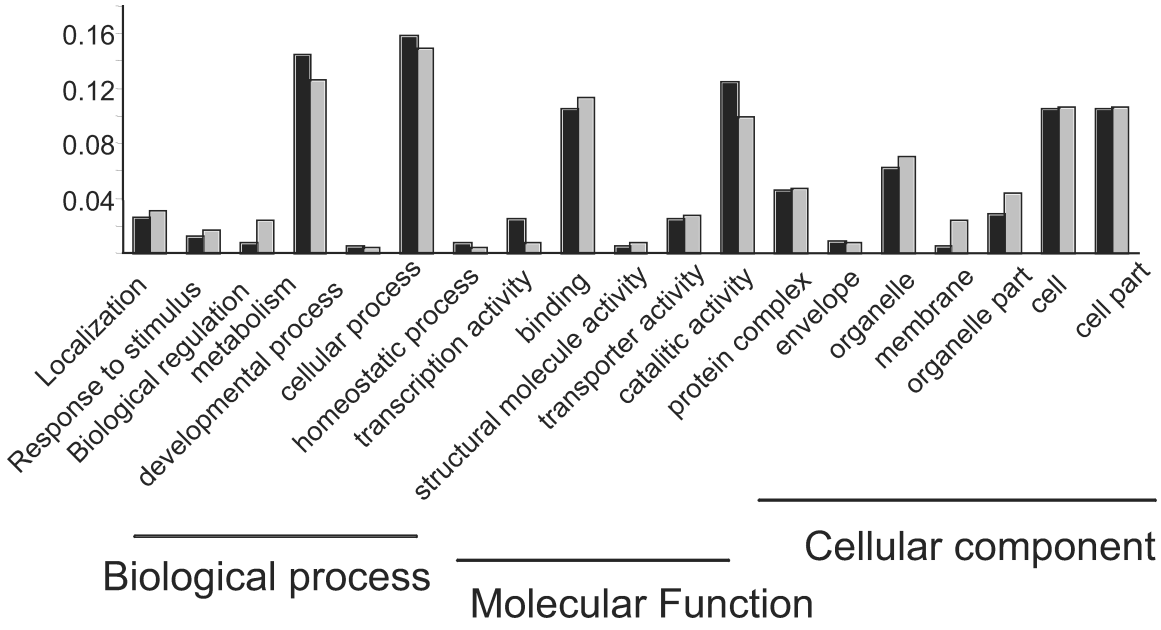

FIGURE 5. Proportions of the putative functions, according to the Gene Ontology (GO) classification, among the 246 clusters of orthologs (in black) and among the 59 best-performing genes (in grey).

to 57% for the worst-performing gene in that rate range. We did not find an optimal average rate at which these genes evolve, and there is no correlation between the average rate and the phylogenetic performance estimated by the topological and branch length distances used here.

FUNYBASE

A Web site is available with open access to the clusters of orthologous sequences: http://genome.jouy.inra.fr/funybase/funybase_result.cgi. The aligned clusters of orthologous sequences can be downloaded either for the data set containing 30 fungal genomes or for the data set including the 21 genomes ultimately retained. For each cluster, the mean identity of the sequences, the amino acid substitution model that best fit the data, and the topological score are available. The database was named FUNYBASE, after FUNgal phyLogenomic dataBASE.

DISCUSSION

Comparison of the Methods Used across Studies for Identifying Genes with High Phylogenetic Performance

In this study we show that there is high variability in the phylogenetic performance of single genes and that

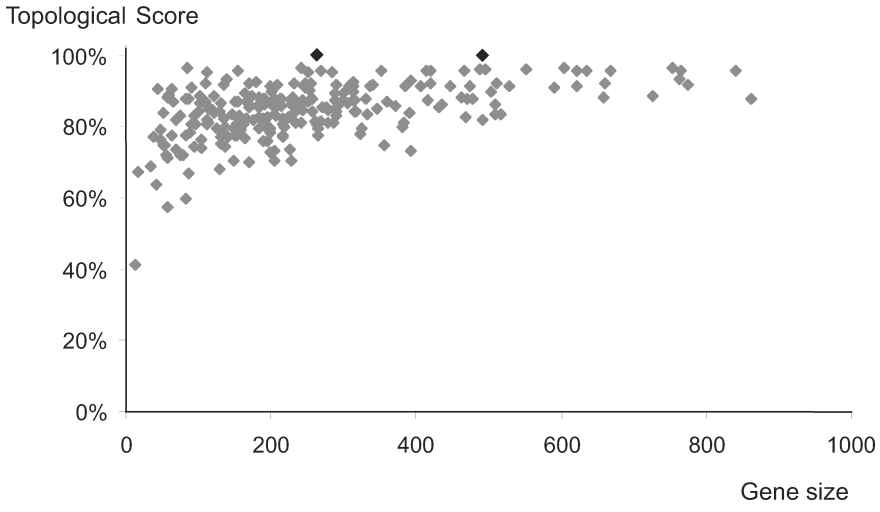

FIGURE 6. Plot of the topological score as a function of gene size for the 246 individual gene phylogenies with 21 species. The two genes with a topological score of 100% are represented in black.

accurate phylogenies can be obtained using only a few of the best-performing genes. Similar conclusions have been reached by previous studies on different taxa using bacteria (Konstantinidis et al., 2006), yeasts (Rokas et al., 2003), and vertebrates (Graybeal, 1994; Cummings et al., 1995; Zardoya and Meyer, 1996; Springer et al., 2001; Mueller, 2006). In contrast, a recent study aiming at identifying the best genes for fungal phylogenies concluded that the concatenation of 40 to 45 proteins was needed to correctly resolve the fungal tree of life (Kuramae et al., 2007). There are important differences in terms of the methods employed in each of these studies with respect to the analysis conducted here and we discuss below how it could have affected the conclusions.

First, our work and most of the previous studies differ in the methods followed to obtain the clusters of shared and putatively orthologous genes, from which the alignments and trees are inferred. We decided to use the MCL clustering methods because we wanted to recover the maximum possible number of orthologous gene clusters but doing so under sufficiently stringent conditions to avoid hidden paralogs. The trade-off involved in this operation is best handled with MCL because its specificity can be finely tuned with respect to the data set at hand (Costa et al., 2005; Brohee and van Helden, 2006). We chose a value of the inflation parameter that had been shown to produce an optimal number of clusters containing shared single-copy genes (Enright et al., 2002; Dujon et al., 2004; Robbertse et al., 2006). Robbertse et al. (2006) also used this approach on fungi and made a thorough comparison of the number of clusters inferred with respect to different values of the inflation parameter. They found an optimal value at which the orthologs recovered were reliably identified and clustered. Other studies used rather ad hoc methods to derive clusters of orthologous proteins, either identifying each gene family with a single representative in each genome (Fitzpatrick et al., 2006) or by inferring orthology based on the eukaryotic ortholog protein database KOG (http://www.ncbi.nlm.nih.gov/COG/grace/shokog.cgi); (Kuramae et al., 2006, 2007). We consider that these latter methods are not designed to optimize the number of inferred orthologous clusters: methods may be too liberal when the only criterion to propose orthology is that proteins are present as single-copy in all compared genomes, as hidden paralogy can be a serious concern; on the contrary, methods can be too conservative when they are based on more general ortholog databases, for instance, the KOG database, which to date is built using only two fungal species (*S. cerevisiae* and *S. pombe*) and require similarity with other more distantly related eukaryotes, as in this case they are likely to miss many orthologs that are exclusively shared among fungi. Furthermore, we showed that the inclusion of some genomes drastically decreased the number of shared orthologous clusters, most probably due to numerous errors in protein predictions, and this had never been evaluated. In the end we recovered 219 clusters of orthologs in the 23-species data set and 246 single-copy orthologs in the 21-species data set, which represents only ca. 2.5% of the mean number of proteins found in the 30 fungal species (Table 1). This small number of single-copy genes shared by all fungal genomes is in agreement with previous studies: 531 single-copy clusters of orthologs shared among 25 eukaryotic genomes (Kuramae et al., 2006); 70 single-copy clusters of orthologs shared among 33 fungal genomes (Kuramae et al., 2007); 153 shared among 42 fungi (Fitzpatrick et al., 2006); and 854 shared among 17 Ascomycota (Robbertse et al., 2006). We expect the reliability of orthology to differ among studies, and a comparison among the orthologous groups obtained by different studies might be warranted, although it lies outside of the scope of the present work. Finally, it is interesting to note that a recent study has demonstrated that the OrthoMCL algorithm (Li et al., 2003) performs better than the Tribe-MCL method for the automatic detection of ortholog clusters in multiple eukaryotic genomes (Chen et al., 2007). The latter study was performed with default values for the inflation parameter, so it is not possible to make a direct comparsion with our approach. However, for future work it will probably be simpler to directly evaluate the ortholog clusters produced by OrthoMCL than performing a fine-tuning of the Tribe-MCL inflation parameter.

The choice of methods used to infer the trees does not seem to affect the results. In this study, maximum likelihood and Bayesian methods produced the same topologies. Other studies have also reported similarly convergent results regardless of the methods used to infer trees. Model choice, in the case of likelihood and Bayesian methods, is almost always done as we did here, by statistically testing the fit of different models to the data and choosing the best fitting model. It is important to note that in our case two models were consistently chosen as fitting best our data: the WAG (Whelan and Goldman, 2001) and the rtREV (Dimmic et al., 2002) models, similar to previous findings (Robbertse et al. 2006). The fact that those two models are among the most complex available to date explains to some degree their performance. However, it was somewhat puzzling to find that WAG and rtREV each fitted best to roughly half of the clusters when we analyzed the 21-species data set (116 fitted by WAG and 122 fitted by rtREV), as opposed to the analysis of the 23-species data set, where WAG best fitted 163 clusters and rtREV best fitted only 37 clusters. We speculate that both models perform roughly similarly well in fitting our data, and that choosing one over the other is especially difficult with current statistical tests. Both models have 189 adjustable parameters and though derived from very different data, the rtREV model may take a mathematical form that makes it equivalent to the WAG model (Dimmic et al., 2002). It has been suggested that model averaging may capture the best features of different models in limit cases where model choice is especially difficult to decide (Posada and Buckley, 2004).

A very important difference among the studies that aimed at identifying genes with high phylogenetic performance was the comparison of the phylogenies obtained from individual genes and a reference tree. We

employed three metrics to estimate the distances, both in terms of topology (Robinson and Foulds, 1981; Nye et al., 2006) and branch length (Kuhner and Felsenstein, 1994). These metrics are strict measures of distance between trees (Felsenstein, 2004); all three produced similar results and were correlated. In contrast, most of the previous studies did not use indices for topological comparison. The first studies obtained individual gene trees by analyzing partitions of different mitochondrial genes or subsets of nucleotides extracted from the mitochondrial genome and compared those individual phylogenies to a reference tree that was not based on whole-genome data sets: Cummings et al. (1995) used the contraction/decontraction method (Robinson-Foulds, 1981); Springer et al. (2001) measured the variation in bootstrap support level; Zardoya and Meyer (1996) used the KH (Kishino and Hasegawa, 1989) and Templeton's (1983) tests; and Mueller (2006) compared the trees by counting the number of recovered expected branches. Among the studies using a reference tree based on whole-genome sequence, some were not explicit about the method used (Robbertse et al., 2006; Kuramae et al., 2006); in one case the YAPTP test was used (Fitzpatrick et al., 2006), which only tells if the topologies under comparison are more similar between them than expected by chance (Creevey et al., 2004); in another study the correlation of the branch lengths between single gene and reference trees was measured (Rokas et al., 2003). Finally, Kuramae et al. (2007) used Pearson's correlation coefficient between genetic distance matrices. This latter study is very close to ours in terms of both objectives and model organisms, but uses very different methods and reaches drastically different conclusions. Only 70 clusters of orthologs were found to be shared among their 33 fungal genomes, based on the KOG database; the reference tree included species with ambiguous placements, such as *S. nodorum* and *A. oryzae*; phylogenetic performance was measured as a correlation of distance matrix and was thus not based on topological comparison. This may explain why Kuramae et al. (2007) concluded that the concatenation of 40 to 45 proteins were needed to recover their reference tree, whereas we found that only a few genes were sufficient. We did a BLAST search of the five genes with the highest correlation coefficient to the reference tree found by Kuramae et al. (2007) and checked their topological score in FUNYBASE. Among the five proteins, four corresponded to our clusters MS34, FG684, MS400, and FG705, with topological scores of 84.2%, 82%, 78.8%, and 76.9%, respectively. These very low topological score values indicate that those proteins ranked as the best-performing ones by Kuramae et al. (2007) actually yield trees with very low topological similarity with the tree based on whole genome sequences. Further, among the five proteins ranked as the best-performing ones by Kuramae et al. (2007), the last one corresponded to the cluster FG3837, for which no ortholog was detected in nearly half of the genomes in our data sets. We therefore advocate that future studies aiming at finding genes with high phylogenetic performance use only genomes with reliable protein prediction, use MCL to infer clusters of single-copy orthologs, use a reliable reference tree, and use a strict topological metric; e.g., the topological score (Nye et al., 2006).

Another novelty of our approach compared to previous ones is the test of the consistency of phylogenetic performance among samples. If the genes producing a tree similar to the reference tree for a set of species performed less well on another set of species, they would be of limited interest. However, we showed that the phylogenetic performance of the genes, as measured by indices of topological similarity, did not depend strongly on the chosen set of species. The phylogenetic performance must therefore be an intrinsic property of the genes, most probably related to factors such as evolutionary rates, number of informative sites, and demographic and selective histories.

It would in fact be convenient to find parameters from which phylogenetic performance could be predicted, because it would remove the necessity of comparison with full-genome data sets. We found that both gene size and the total number of variable sites were significantly correlated and were moderately good predictors of the phylogenetic performance of individual genes (see also Galtier, 2007). All the genes longer than 700 bp yielded good topological scores. However, the two best-performing genes had intermediate gene sizes and there was a great variability in the phylogenetic performance for a given gene size, indicating that other factors have an impact on phylogenetic performance of single genes. Because topological scores higher than 90% could be recovered at any gene size, the best approach may be to combine several genes with high phylogenetic performance, regardless of their size. The GO functional categories that corresponded to biological regulation, membrane, organelle part, and binding were overrepresented in the 59 best-performing genes relative to the 246 single-copy genes identified, but this cannot serve either as a strong a priori predictor of phylogenetic performance. The evolutionary rate did not appear useful to detect genes with high phylogenetic performance either. It may be that the demographic and selective histories of the genes are more relevant. Assessing phylogenetic performance using similar approaches as ours, or as the one proposed by Townsend (2007), therefore appears essential to detect the genes that produce the most accurate phylogenies.

The Best-Performing Genes Identified for Fungal Phylogenies

The phylogeny of fungal species inferred here using a genome-wide sampling of orthologous proteins had the same topology as those found in previously published studies, where complete genome sequences were also used, either by concatenating common orthologs or by employing supertree methods (Fitzpatrick et al., 2006; Robbertse et al., 2006; Kuramae et al., 2006). In those studies, the placement of *S. nodorum* and the relationships among the three *Aspergillus* species were also weakly supported. Three different placements of *S. nodorum* occurred with similar frequency in the single gene trees. Because the second and third placements (online Appendix 1) corresponded to the closest nodes

subsequent to the first placement, we believe that the discordance between the different gene trees is probably best explained by rapid speciations that prevented accumulation of synapomorphies (Rosenberg and Nordborg, 2002; Rokas and Carroll, 2006) or caused lineage sorting. The lack of resolution of the nodes among the *Aspergillus* species can also be due to rapid speciation (Rosenberg and Nordborg, 2002; Rokas and Carroll, 2006) or to lineage sorting (Pollard et al., 2006). Alternatively, duplications and differential loss of paralogs, or horizontal gene transfers, may be responsible for the observed conflicting phylogenetic signals. *Aspergillus oryzae* has indeed been subject to a large-scale genome expansion compared to the two other *Aspergillus* species, due to either duplication or lateral gene transfers (Machida et al., 2005). The difficult placement of the *Aspergillus* and *S. nodorum* species confirms that some clades are refractory to resolution even with complete-genome data sets.

Full-genome data sets nevertheless seem to provide reliable information on the evolutionary relationships of most species. However, if genes are to be sequenced in order to build a phylogeny, only a few can be practically sequenced. Single-gene analyses are, however, dependent on the genes having an evolutionary history that reflects that of the entire organism, which is often not the case. By comparing the topologies of the individual gene trees with that of the reference tree, based on the full genomes, we found a high variability in the phylogenetic performance of individual genes. Some genes performed better than others, and this seemed to be an intrinsic property of the genes, not dependent on the sample. It was remarkable, and not necessarily expected, that two single genes produced exactly the same topology as the reference tree. Our results show that two well-chosen genes were sufficient to recover a robust phylogenetic tree that reflected the whole-genome relatedness among the species. By extending our analysis to 36 genomes, we showed that the best-performing genes found with our approach can be of general utility for fungal phylogenies, although five to six genes among the best-performing genes should probably be used to resolve all the nodes. The genes commonly used for inferring fungal phylogenies often had paralogs in some genomes and/or were lacking from others and had furthermore poor phylogenetic performance. The currently standing fungal tree of life therefore seems to have been built using marker genes that perform suboptimally for phylogenetic inference. The validity of current classifications based on single commonly used markers (e.g., ITS, elongation factors, tubulins) may warrant a careful revision.

We note that the genes ranked here as the best-performing ones may not be suitable for shallower taxonomic scales. Nevertheless, deep-level clades within the Dicaryomycota are the most difficult to resolve (Lutzoni et al., 2004; James et al., 2006) and are those for which our best-performing genes should be the most useful (Liu and Hall, 2004). For shallower-level phylogenetics, the FUNYBASE allows the rapid mining of single-copy genes present in all fungal genomes and will therefore be useful for finding suitable genes at the desired taxonomic scale. It is, however, interesting to note that previous studies found roughly the same ranking in the performance of mitochondrial protein-coding genes for recovering the expected phylogeny of teleosteans as for recovering the expected phylogeny of tetrapods and mammals (Zardoya and Meyer, 1996; Miya and Nishida, 2000). The genes we ranked among the best-performing genes for resolving relationships among the major fungal groups should be useful for many purposes, economizing costs and improving accuracy for large-scale phylogenies, in particular for building the tree of life. Basal nodes in the fungi, and in particular within the Dicaryomycota, are indeed still poorly supported (James et al., 2006) and need further analyses using genes with high phylogenetic performance at this scale.

CONCLUSION

Phylogenetic sequence analysis of three to six genes represents currently the most favourable approach for inferring relationships among species. Our evaluation of the best-performing genes describing relationships between species from the two main fungal phyla shows that this approach is highly reliable, as it provides results similar to those obtained with whole-genome data sets. The choice of genes, however, is critical because the performance of the different protein-coding genes in deriving a reliable phylogeny was highly variable. We expect that the methodology developed here based on fungal genomes will guide the selection of genes to use in future phylogenetic studies in fungi and will also be applied to other groups of organisms as more genomes become available.

ACKNOWLEDGMENTS

We thank Gilles Fisher, Bernard Dujon, and Françis Martin for sharing data and Joëlle Amsellem, Emmanuel Quevillon, Pierre Nicolas, Damien de Vienne, Sébastien Ollier, and Manuela López-Villavicencio for helpful discussions. We thank Cécile Ané, Jack Sullivan, David Fitzpatrick, and anonymous referees for useful comments on a previous version of the manuscript. This study was funded by the French Bureau des Ressources Génétiques (BRG 2005–2008), an "ACI Jeunes Chercheurs" (2003–2006), and an "ANR Blanc" (ANR-06-BLAN-0201). G. A. acknowledges CNRS and U. PSUD postdoctoral grants.

REFERENCES

Abascal, F., R. Zardoya, and D. Posada. 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

Altschul, S., T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene Ontology: Tool for the unification of biology. Nat. Genet. 25:25–29.

Brohee, S., and J. van Helden. 2006. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7:488.

Castresana, J. 2000 Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Chen, F., A. J. Mackey, J. K. Vermunt, and D. S. Roos. 2007 Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE. 2:e383.

Costa, G. G. L., L. A. Digiampietri, E. H Ostroski, and J.C. Setúbal. 2005. Evaluation of graph-based protein clustering methods. Pages 95-11 *in* Proceedings of the Fifth Brazilian Symposium on Mathematical and Computational Biology (BIOMAT2005), Rio de Janerio, Brazil.

Creevey, C. J., Fitzpatrick, D. A., Philip, G. K., Kinsella, R. J., O'Connell, M. J., Pentony, M. M., Travers, S. A., Wilkinson, M. and J. O. McInerney. 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proc. R. Soc. Lond. B Biol. Sci. 271:2551–2558.

Cummings, M. P., S. P Otto, and J. Wakeley. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12:814–822.

Dimmic, M. W., J. S. Rest, D. P. Mindell, and R. A. Goldstein. 2002. rtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J. Mol. Evol. 55:65–73.

Dujon, B., D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. de Montigny, C. Marck, C. Neuveglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J. M. Beckerich, E. Beyne, C. Bleykasten, A. Boisrame, J. Boyer, L. Cattolico, F. Confanioleri, A. de Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J. M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G. F. Richard, M. L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wesolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker, and J. L. Souciet. 2004. Genome evolution in yeasts. Nature 430:35–44.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acid Res. 30:1575–1584.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Fitzpatrick, D., M. Logue, J. Stajich, and G. Butler. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol. Biol. 6:99.

Galtier, N. 2007. A model of horizontal gene transfer and the bacterial phylogeny problem. Syst. Biol. 56:633–642.

Gilks, W. R., B. Audit, D. de Angelis, S. Tsoka, and C. A. Ouzounis. 2005. Percolation of annotation errors through hierarchically structured protein sequence databases. Math. Bio. 193:223–234.

Graybeal, A. 1994. Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. Syst. Biol. 43:174–193.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Hawksworth, D. L. 1991. The fungal dimension of biodiversity: Magnitude, significance, and conservation. Mycol. Res. 95:641–655.

James, T. Y., F. Kauff, C. L. Schoch, P. B. Matheny, V. Hofstetter, C. J. Cox, G. Celio, C. Gueidan, E. Fraker, J. Miadlikowska, H. T. Lumbsch, A. Rauhut, V. Reeb, A. E. Arnold, A. Amtoft, J. E. Stajich, K. Hosaka, G. H. Sung, D. Johnson, B. O'Rourke, M. Crockett, M. Binder, J. M. Curtis, J. C. Slot, Z. Wang, A. W. Wilson, A. Schussler, J. E. Longcore, K. O'Donnell, S. Mozley-Standridge, D. Porter, P. M. Letcher, M. J. Powell, J. W. Taylor, M. M. White, G. W. Griffith, D. R. Davies, R. A. Humber, J. B. Morton, J. Sugiyama, A. Y. Rossman, J. D. Rogers, D. H. Pfister, D. Hewitt, K. Hansen, S. Hambleton, R. A. Shoemaker, J. Kohlmeyer, B. Volkmann-Kohlmeyer, R. A. Spotts, M. Serdani, P. W. Crous, K. W. Hughes, K. Matsuura, E. Langer, G. Langer, W. A. Untereiner, R. Lucking, B. Budel, D. M. Geiser, A. Aptroot, P. Diederich, I. Schmitt, M. Schultz, R. Yahr, D. S. Hibbett, F. Lutzoni, D. J. McLaughlin, J. W. Spatafora, and R. Vilgalys. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature 443:818–822.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.

Konstantinidis, K. T., A. Ramette, and J. M. Tiedje. 2006. Toward a more robust assessment of intraspecies diversity using fewer genetic markers. App. Env. Microbiol. 72:7286–7293.

Koonin, E. V. 2005. Orthologs, Paralogs, and evolutionary genomics. Annu. Rev. Genet. 39:309–338.

Kuhner, M. K., and J. Felsenstein.1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol.11:459–468.

Kuramae, E. E., Robert, V., Echavarri-Erasun, C. and T. Boekhout. 2007. Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: An example from the fungal kingdom. BMC Evol. Biol. 7:134.

Kuramae, E. E., Robert, V., Snel, B., Weiss, M. and T. Boekhout. 2006 Phylogenomics reveal a robust fungal tree of life. FEMS Yeast Res. 6:1213–1220.

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Li, L., C. J. Stoeckert Jr., and D. S. Roos. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Liu, Y. J., and B. D. Hall. 2004. Body plan evolution of ascomycetes, as inferred from an RNA polymerase II phylogeny. Proc. Natl. Acad. Sci. USA 101:4507–4512.

Lutzoni, F., F. Kauff, C. J. Cox, D. McLaughlin, G. Celio, B. Dentinger, M. Padamsee, D. Hibbett, T. Y. James, E. Baloch, M. Grube, V. Reeb, V. Hofstetter, C. Schoch, A. E. Arnold, J. Miadlikowska, J. Spatafora, D. Johnson, S. Hambleton, M. Crockett, R. Shoemaker, G.-H. Sung, R. Lücking, T. Lumbsch, K. O'Donnell, M. Binder, P. Diederich, D. Ertz, C. Gueidan, K. Hansen, R. C. Harris, K. Hosaka, Y.-W. Lim, B. Matheny, H. Nishida, D. Pfister, J. Rogers, A. Rossman, I. Schmitt, H. Sipman, J. Stone, J. Sugiyama, R. Yahr, and R. Vilgalys. 2004. Where are we in assembling the fungal tree of life, classifying the fungi, and understanding the evolution of their subcellular traits? Am. J. Bot. 91:1446–1480.

Machida, M., K. Asai, M. Sano, T. Tanaka, T. Kumagai, G. Terai, K. I. Kusumoto, T. Arima, O. Akita, Y. Kashiwagi, K. Abe, K. Gomi, H. Horiuchi, K. Kitamoto, T. Kobayashi, M. Takeuchi, D. W. Denning, J. E. Galagan, W. C. Nierman, J. J. Yu, D. B. Archer, J. W. Bennett, D. Bhatnagar, T. E. Cleveland, N. D. Fedorova, O. Gotoh, H. Horikawa, A. Hosoyama, M. Ichinomiya, R. Igarashi, K. Iwashita, P. R. Juvvadi, M. Kato, Y. Kato, T. Kin, A. Kokubun, H. Maeda, N. Maeyama, J. Maruyama, H. Nagasaki, T. Nakajima, K. Oda, K. Okada, I. Paulsen, K. Sakamoto, T. Sawano, M. Takahashi, K. Takase, Y. Terabayashi, J. R. Wortman, O. Yamada, Y. Yamagata, H. Anazawa, Y. Hata, Y. Koide, T. Komori, Y. Koyama, T. Minetoki, S. Suharnan, A. Tanaka, K. Isono, S. Kuhara, N. Ogasawara, and H. Kikuchi. 2005. Genome sequencing and analysis of Aspergillus oryzae. Nature 438:1157–1161.

Miya, M., and M. Nishida. 2000. Use of mitogenomic information in teleostean molecular phylogenetics: A tree-based exploration under the maximum-parsimony optimality criterion. Mol. Phylogenet. Evol. 17:437–455.

Mueller, R. L. 2006. Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic anlysis. Syst. Biol. 55:289–300.

Nye, T. M. W., P. Lio, and W. R. Gilks. 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics 22:117–119.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: Evidence for incomplete lineage sorting. PLoS Genet. 2:1634–1647.

Posada, D., and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793–808.

Robbertse, B., J. B. Reeves, C. L. Schoch, and J. W. Spatafora. 2006. A phylogenomic analysis of the Ascomycota. Fung. Genet. Biol. 43:715–725.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rokas, A., and S. B. Carroll. 2006. Bushes in the tree of life. PLoS Biol. 4:1899–1904.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Rosenberg, N. A., and M. Nordborg. 2002. Genealogical theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3:380–390.

Schaffer, R. 1975. The major groups of Basidiomycetes. Mycologia 66:1–18.

Springer, M. S., R. W. DeBry, C. Douady, H. M. Amrine, O. Madsen, W. W. de Jong, and M. Stanhope. 2001. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. Mol. Biol. Evol. 18:132–143.

Templeton, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of human and the apes. Evolution 37:221–244.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 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.

Townsend, J. P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.

Yang, Z. 1997. PAML: A program for package for phylogenetic analysis by maximum likelihood. CABIOS 15:555–556.

Zardoya, R., and A. Meyer. 1996. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Mol. Biol. Evol. 13:933–942.

First submitted 14 January 2008; reviews returned 24 March 2008; final acceptance 7 May 2008

Associate Editor: Cécile Ané