A genome Tree of Life for the Fungi kingdom

JaeJin Choi[a,b,c,d] and Sung-Hou Kim[a,b,c,e,1]

[a]Department of Chemistry, University of California, Berkeley, CA 94720; [b]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [c]Department of Integrated Omics for Biomedical Sciences, Yonsei University, Seoul 03722, Republic of Korea; [d]Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, Republic of Korea; and [e]Center for Computational Biology, University of California, Berkeley, CA 94720

Contributed by Sung-Hou Kim, July 24, 2017 (sent for review July 7, 2017; reviewed by Se-Ran Jun and Charles R. Vossbrinck)

Fungi belong to one of the largest and most diverse kingdoms of living organisms. The evolutionary kinship within a fungal population has so far been inferred mostly from the gene-information–based trees ("gene trees"), constructed commonly based on the degree of differences of proteins or DNA sequences of a small number of highly conserved genes common among the population by a multiple sequence alignment (MSA) method. Since each gene evolves under different evolutionary pressure and time scale, it has been known that one gene tree for a population may differ from other gene trees for the same population depending on the subjective selection of the genes. Within the last decade, a large number of whole-genome sequences of fungi have become publicly available, which represent, at present, the most fundamental and complete information about each fungal organism. This presents an opportunity to infer kinship among fungi using a whole-genome information-based tree ("genome tree"). The method we used allows comparison of whole-genome information without MSA, and is a variation of a computational algorithm developed to find semantic similarities or plagiarism in two books, where we represent whole-genomic information of an organism as a book of words without spaces. The genome tree reveals several significant and notable differences from the gene trees, and these differences invoke new discussions about alternative narratives for the evolution of some of the currently accepted fungal groups.

fungal phylogeny | proteome tree | divergence tree | alignment-free method | feature frequency profile

Diversity of Fungi

Fungi form one of the largest eukaryotic kingdoms, with an estimated 1.5–5 million species. They form a diverse group with a wide variety of life cycles, metabolisms, morphogenesis, and ecologies, including mutualism, parasitism, and commensalism with many live organisms. They are found in all temperature zones of the Earth with diverse fauna and flora, and have a very broad and profound impact on the Earth's ecosystem through their functions of decomposing diverse biopolymers and other biological compounds in dead or live hosts, and of synthesizing diverse classes of biomolecules as foods for human and other eukaryotes (1–3). Whole-genome sequences of varying completeness of over 400 fungal species are available publicly at present. The genome size for the species ranges from about 2–180 million nucleotides and predicted proteome size ranges from about 2–35 thousand proteins.

Phylogeny Derived from "Gene Trees"

The evolutionary phylogeny, or kinship, among the fungi have been inferred almost exclusively from the gene-information–based tree ("the gene trees"), construction of which use, most commonly, the multiple sequence alignment (MSA) method on the gene or protein sequences encoded for a small number of highly conserved and orthologous genes (4, 5). Thus, strictly speaking, a gene tree may represent the combined phylogeny of the selected genes, but may not represent organisms, because each species cannot be represented by a small number of selected genes, but only be represented by whole-genome information of the species. This issue about the gene trees is analogous to predicting the similarity between two books by comparing a similar sentence, paragraph, or chapter subjectively selected to represent each book, calculate their similarity by sequence alignment, then project the similarity to estimate how similar the two books are. For gene tree construction, there are indications that the larger the number of homologous genes selected, the better the topology of the trees converge to a stable state (6–8), suggesting that, ultimately, whole-genome information, if available, may be compared to obtain a stable and, perhaps, reliable tree from which the evolutionary phylogeny of organisms can be inferred. However, the whole-genome sequences cannot be compared by the MSA method, because the overwhelming portion of the whole-genome sequences cannot be aligned by MSA.

"Genome Tree" of Fungi

Due to dramatic advances in whole-genome sequencing technology, a large number of whole-genome sequences—at varying degrees of completeness—of fungi are now available publicly. Such whole-genome information provides an opportunity to explore the genome-information–based tree ("the genome tree") constructed using several types of whole-genome information: whole-genome DNA sequence, transcriptome RNA sequence, proteome amino acid sequence, exome DNA sequences, or other genomic features. In this study, we use the whole-proteome sequences on the Feature Frequency Profile (FFP) method (9) (*Materials and Methods*) to estimate the similarity between two organisms without sequence alignment, then build a proteome-based genome tree ("proteome tree"; see the first section in

Significance

Fungi belong to one of the largest and most diverse groups of living organisms. The evolutionary kinship within a fungal population has so far been inferred mostly from the gene-information–based trees ("gene trees") constructed using a small number of genes. Since each gene evolves under different evolutionary pressure and time scale, it has been known that one gene tree for a population may differ from other gene trees for the same population, depending on the selection of the genes. We present whole-genome information-based trees ("genome trees") using a variation of a computational algorithm developed to find plagiarism in two books, where we represent a whole-genomic information of an organism as a book of words without spaces.

Author contributions: S.-H.K. designed research; JJ.C. and S.-H.K. performed research; JJ.C. contributed new programs/figures; JJ.C. and S.-H.K. analyzed data; and JJ.C. and S.-H.K. wrote the paper.

Reviewers: S.-R.J., University of Arkansas for Medical Sciences; and C.R.V., The Connecticut Agricultural Experimental Station.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The FFP programs for this study, written in GCC(C++), have been deposited in GitHub, https://github.com/jaejinchoi/FFP.

[1]To whom correspondence should be addressed. Email: sunghou@berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1711939114/-/DCSupplemental.

Results for the reason of choosing the proteome sequence). The FFP method is a variation of a computational algorithm developed to find semantic similarities or plagiarism between two books by the similarity of the "word-frequency profiles," each representing a book, as in, for example, the Latent Semantic Analysis method of natural languages (10). In the FFP method, the whole-proteome sequence is treated as a book of words without spaces and the FFPs of two "books-without-spaces" are compared to estimate the similarity between the two "books." The method has been successfully tested and applied previously to construct a proteome tree of prokaryotes using proteome sequences (11) and the whole-genome DNA-based tree of mammals (12). In this study a proteome tree is presented for the members of the fungi (plus protozoas) of known genome sequences, which have much greater diversity and complexity in phenotype, morphogenesis, and life cycle than prokaryotes or mammals.

Objectives

The objective of this study is to present notable differences at various clade levels between the two types of fungi trees constructed by the two fundamentally different methods, and to highlight relevant observations related to the placement and branching order of several clades that are relevant in inferring evolutionary relationship to other groups in the Fungi kingdom. It is hoped that such differences and observations may encourage new discussions among scientists in the field: (*i*) to revisit the current narratives for the evolution of clades derived from the view point of gene trees alone; and (*ii*) to debate the pros and cons of the current gene-based evolutionary model, which hypothesizes that a single mechanism of nucleotide substitution of individual genes represents the mechanism of evolution of an organism, vs. an alternative genome-based model, which hypothesizes that there may be multiple mechanisms that manifest in the divergence of whole-genome information. We also hope that this method, developed and tested for fungi, may be usable for the construction of a Tree of Life for all organisms.

Results

This section addresses choice of whole-genome information, overall features of a proteome tree of fungi and protists, protistan origin of Microsporidia, and a description of other notable differences and similarities between the proteome tree and current gene trees.

Choice of Genomic Information: Whole-Genome DNA Sequence vs. Transcriptome RNA Sequence vs. Proteome Amino Acid Sequence. Since there is no a priori criteria for the best descriptor to build the organism phylogeny, we took an empirical approach to find the best one among three types of descriptors in the public databases: whole-genome DNA sequence, transcriptome sequence, and proteome sequence. In addition, the "optimal feature lengths" of the three descriptors, the critical information needed for the FFP method that would give the most stable tree topology, was also empirically determined using the Robinson-Foulds metric (13) in the PHYLIP package (14). The results of the empirical searches showed that the proteome tree is most topologically stable among the three genome trees (described in more detail in *Materials and Methods*). Various features of the proteome tree of the Fungi kingdom are described below and compared with those of the gene trees based on various selected gene sets.

Overall Features of the Proteome Tree of Fungi.

Three major groups. In contrast to four (Ascomycota, Basidiomycota, Monokarya, and Microsporidia) to eight (Glomeromycota, Zygomycota, Basidiomycota, Ascomycota, Chytridiomycota, Neocallimatigomycota, Blastocladiomycota, and Microsporidia) major groups in the Fungi kingdom in the gene trees (3) (Fig. S1*B*), there are only three (Monokarya, Basidiomycota, and Ascomycota) earliest diverging and deepest branching major fungal groups in the proteome tree (Figs. 1 and 2 and Fig. S1*A*). The first major group (group I in Figs. 1 and 2) corresponds to Monokaryotic fungi and consists of three subgroups that do not appear to produce dikaryons during their life cycle: Cryptomycota, Chytridomycota, and Zygomycota. The second major group (group II) corresponds to Basidiomycota, which are dikaryon-producing fungi whose sexual spores are formed externally on small-pedestal fruiting bodies called basidia, and consists of Puccinomycotina, Ustilaginomycotina, and Agaricomycotina. The third major group (group III) corresponds to Ascomycota, which are dikaryon-producing fungi whose sexual spores are formed internally inside sacs called "asci" on top of fruiting bodies, and consists of Taphrinomycotina, Saccharomycotina, and Pezizomycotina. The three major groups appear to have branched out almost simultaneously from the common ancestor of all fungi (Fig. 2).

Protistan origin of Microsporidia. Microsporidia has been assigned as the basal group of all fungi in most gene trees (e.g., Fig. S1*B*). Surprisingly, in the proteome tree, the group is placed among the nonfungal unicellular eukaryotic organisms of paraphyletic protists "Protozoa" (marked "(a)" in Figs. 1 and 2, and Fig. S1*A*; see below for details).

Similar clading patterns, but different branching order of clades. Although the member compositions of the groups at the next to the deepest level of divergence in the proteome tree are similar to those in the gene trees, the branching orders of some of the groups are different, and more so at higher branching levels (compare Fig. S1 *A* and *B*; also, see below). Fig. 1 shows clade membership and branching order of the proteome tree, and Fig. S1*A* shows the taxon identifications of the fungi and protists in the tree. The statistical support calculated by Jackknife Monophyly Indices (15) and the relative branch lengths for various clades are shown in a simplified tree (Fig. 2).

Protistan Origin of Microsporidia. The Microsporidia is a eukaryotic group of spore-forming unicellular obligate parasites to a very wide range of animal hosts, including human. Several thousands of them are named, suggesting that there may be more than an order-of-magnitude more unnamed Microsporidia species in nature. Individual Microsporidia species usually infect one host species or a group of closely related taxa. They have very small genomes, and the gene trees place the group at or near the basal position of all fungi (e.g., Fig. S1*B*).

Although the supporters of the fungal origin of Microsporidia have been gaining the ground rapidly among mycologists, alternative origins cannot be ruled out completely. It has been difficult to infer the evolutionary history of Microsporidia due to its shifting positions in the gene trees depending on the genes selected to build the gene trees and evolutionary narratives to explain the shifts based on comparative genome sequences and biochemical data (for a review, see ref. 16). To interrogate the boundary between the fungal kingdom and protists (large, diverse, and paraphyletic/polyphyletic, unicellular, nonfungal microbial eukaryotes) and also to revisit the fungal origin of Microsporidia, a group of 71 protists, for which genome sequences are available, was included in this study.

In the proteome tree constructed for a population containing both fungi and protists, as in the gene trees, all members of Microsporidia in the study form a single clade, suggesting that they most likely evolved from a common ancestor. However, the clade is not located with other fungi, as in the gene trees, but located among the protists, such as Giardia, Trichomonas, Entamoeba, and, Trypanosomatiae, some of which, like the Microsporidia, also lack or lost mitochondria, but have much larger genomes than Microsporidia (Figs. 1 and 2 and Fig. S1*A*). This observation indicates that the proteome sequences of Microsporidia

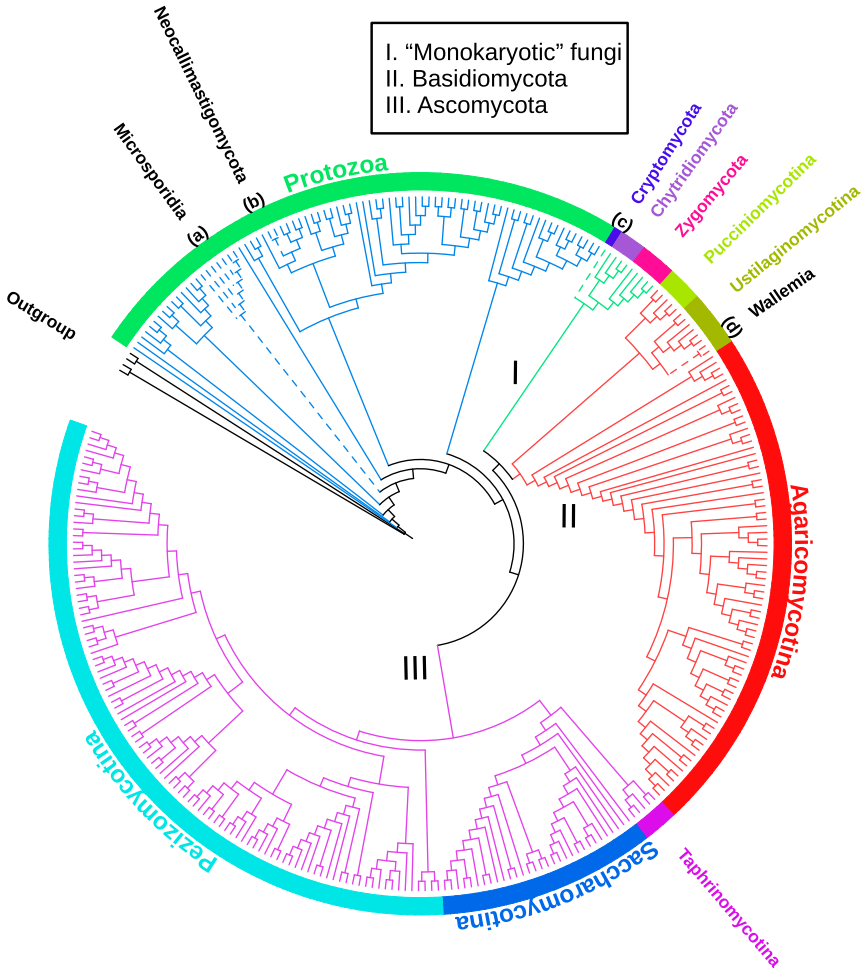

Fig. 1. A Circos (topological) representation of the proteome tree of Fungi kingdom. The branches of three major groups are colored in light green for group I (Monokaryotic fungi), red for group II (Basidiomycota), and purple for group III (Ascomycota). All protists are in blue. The branches of two sets of outgroups are in black. The names of nine groups at phylum level belonging to the three major groups are shown around the circle. The four marked (by lowercase alphabets in parentheses) groups with dotted-lined branches are the groups whose placements in the proteome tree are significantly different from those in the gene trees, as discussed in *Results*. The taxon identification numbers can be found in Fig. S1*A*, and their taxon names can be found in Table S1. For the identities of the outgroups, see *Materials and Methods*. The branch lengths are relative and not to scale. The figure was prepared using the Interactive Tree of Life (ITOL) (43).

are more similar to those of the protists than to those of fungi. The current narrative is that the very small genome sizes of Microsporidia have resulted from one or more steps of extreme reduction of much larger genomes of fungal origin (16). Most of these “evidences” are based on the sequence similarity of the proteins coded by one or limited number of genes (4, 5, 17, 18). However, the proteome tree suggests another narrative that the genomes of the Microsporidia may have a protistan origin rather than fungal origin [marked “(a)” in Figs. 1 and 2 and Fig. S1] and gone through similar extreme genomic reduction.

The protistan origin of Microsporidia was first shown by Vossbrink et al. (19) in their gene tree built using the DNA sequence of a small subunit ribosomal RNA gene, where they placed Microsporidia at the basal position of all eukaryotes they tested (including a few animals, plants, fungi, and protists). However, this proposal was “overturned” by the now-popular fungal origin of Microsporidia based on subsequent gene trees of certain protein-coding genes and narratives derived from biochemical and cellular observations, including the absence or loss of mitochondria, which is not critical to the fungal or protistan origin of Microsporidia (ref. 16 and references within). There was another gene-tree–based indication of grouping of Microsporidia at the basal position of all other eukaryotes: in this study, Thomarat et al. (20) observed, in table 1 of ref. 20, that of 99 gene trees built based on very carefully selected protein sequences of *Encephalitozoon cuniculi*, the first Microsporidida of a known genome sequence, a majority of their trees (80 of 99 gene trees) by the BIONJ method (*Materials and Methods*) placed *E. cuniculi* at the basal position of all eukaryotes (animals, fungi, and plants), thus presumably among protists, whereas the rest of the gene trees placed it at the basal positions of fungi (13 of 99), of fungi and animal (4 of 99), of animal (1 of 99), and at a “nonbasal” position (1 of 99). However, Thomarat et al. discounted their majority results supporting the protistan origin and took a minority results that supports the fungal origin by arguing the slower relative evolutionary change rate of the minority (13 of 99) genes. These observations, combined with the proteome tree, support the protistan origin of Microsporidia rather than the currently popular view of the fungal origin of Microsporidia.

More detailed kinship of Microsporidia among various clades of protists awaits whole-genome sequences of many more protists of diverse taxa, since all 71 protists in this study population are from a subgroup of nonphotosynthetic protists, Protozoa.

Other Notable Differences Between the Features of the Proteome Tree and the Gene Trees. There are a few discrepancies in cladding pattern of species or subspecies between the proteome tree

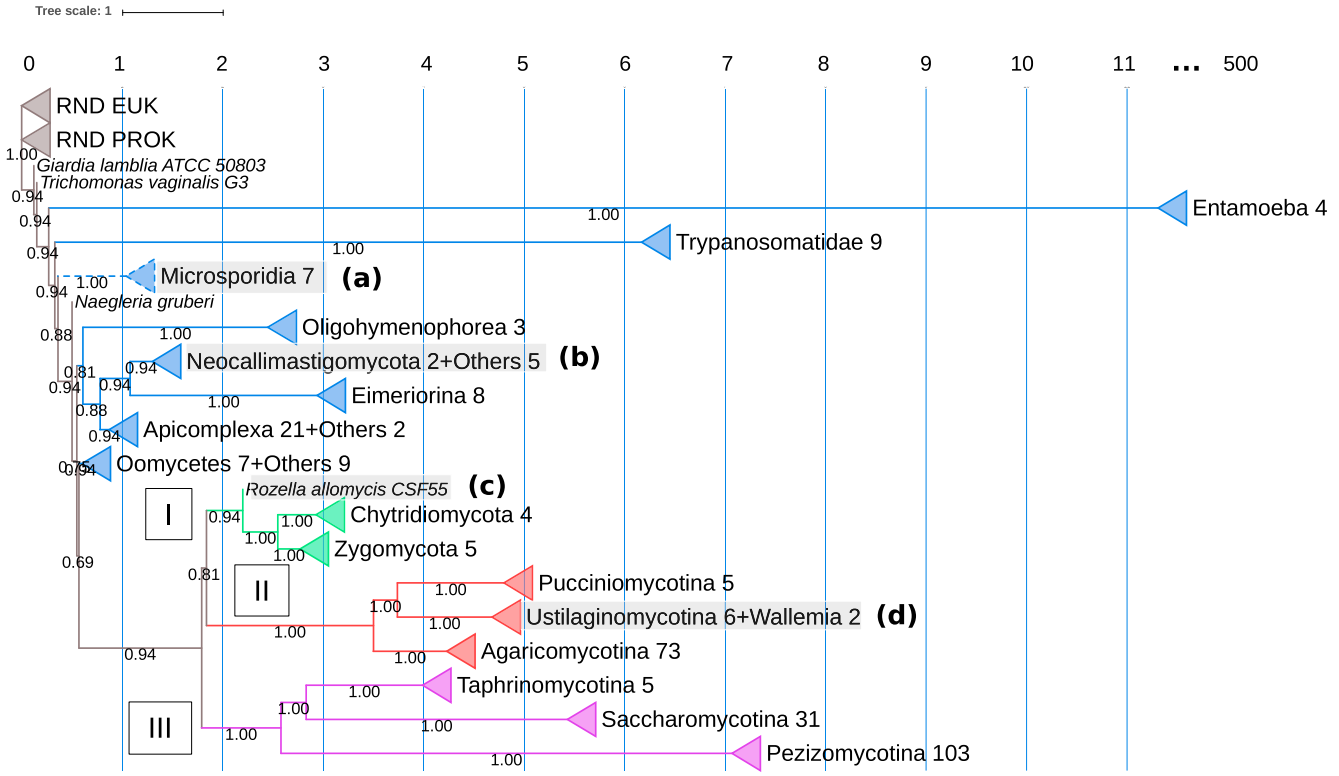

Fig. 2. Simplified proteome tree of Fungi and Protozoa. The figure shows the proteome tree collapsed at the phylum or equivalent levels with the relative branch lengths from one common ancestor of a clade to its previous common ancestor. (The branch lengths for the two outgroups and uncollapsed species are not shown.) For the statistical support of the collapsed groups, the Jackknife Monophyly Index (5) for each collapsed clade (except the two outgroups) are shown under the branch lines. The branch lengths calculated by JSD are normalized to 1,000 (the scale on top), which corresponds to 500 from the common ancestor of fungi and protists to the terminal leaves. The number of the members in a clade is indicated at the end of the clade name, and the four marked (by lowercase alphabets in parentheses) groups are the groups whose placements in the proteome tree are significantly different from those in the gene trees, as discussed in the *Results*. The clade colors correspond to those in Fig. 1. For the identities of the out-group, see *Materials and Methods*. The tree was constructed using ITOL (43).

and the gene trees (Fig. S1). Most discrepancies are found in the branching order of the clades. Beside the differences mentioned above, there are several other notable differences. Some of the differences may be partly due to the small number of species or taxa of known genome sequences available in public database, especially in group I and among the paraphyletic groups of protists. Four notable examples of differences as of this time are described below.

Neocallimastigomycota as a member of protists. Neocallimastigomycota (21) (represented by only two species of known genome sequences, *Piromycetes* and *Orpinomycetes*, in the proteome tree, both having AT-rich genomes) are found in the digestive tracts of herbivores, and reproduce in the stomach of ruminants. They have been classified as a member of a group containing Chytridiomycota in the gene trees (21–23) (Fig. S1*B*). However, in the proteome tree, it groups with a subgroup of protists that also has AT-rich genomes, such as Dictyostelium (24) and Acytostelium (25) [see "(b)" in Figs. 1 and 2 and Fig. S1*A*].

***Branch position of* Rozella.** *Rozella allomyces*, a member of Cryptomycota, is an obligate parasite to other fungi, and is often placed, in the gene trees, at the basal position to or as the sister group of Microsporidia, the basal group of all fungi in the gene trees (Fig. S1*B*). But in the proteome tree, where Microsporidia no longer belongs to Fungi, *Rozella* remains at the basal position of group I (containing Chytridiomycota and Zygomycota) [see "(c)" in Figs. 1 and 2, and Fig. S1*A*].

***Branching order of* Wallemia.** In the proteome tree, *Wallemia* in the phylum Basidiomycota groups with Ustilaginomycotina (Figs. 1 and 2 and Fig. S1*A*), but it groups with Agaricomycotina in some gene trees (26, 27) (Fig. S1*B*).

Branching order of Ascomycota. Many gene trees (Fig. S1*B*) show that Taphrinomyconita branched out first from the common ancestor of Ascomycota (major group III), but in our proteome tree the common ancestor of Taphrinomyconita and Saccharomycotina branched out first (Fig. 2).

Discussion

Gene Trees vs. the Proteome Tree. There is a fundamentally different assumption made in constructing the two trees. In the gene trees it is assumed that an organism can be represented by the DNA/protein sequences of a small number of selected genes that are common among the study organisms and assumed to have the information about the evolution of the organisms. But in the proteome tree the proteome sequence of all genes coding for the proteins of an organism is assumed to represent the organism and have the evolutionary information. This fundamental difference will be reflected in the differences in the two types of trees, because: (*i*) the gene trees show the phylogeny of the selected genes, while the proteome tree reveals the phylogeny of all protein-coding genes and thus, (*ii*) different evolutionary narratives need to be considered for macro- and microscale evolutionary events, such as massive genomic reduction, gain/loss of group of genes, long branch attraction (28), horizontal gene transfer (29), high evolutionary rate of certain genes, and others. There are two important additional differences. (*i*) The evolutionary model used to calculate the evolutionary "distances" between two organisms in building distance-based trees

is different. In gene trees, various DNA or amino acid substitution models are used to calculate the evolutionary distances, where different terms of the parameters describe the rate at which one nucleotide/amino acid replaces another during evolution of the selected genes. However, in genome trees, such as the proteome tree used in this study, a divergence model is used, where the divergence of whole-proteome sequences is used to calculate the evolutionary distance of the organisms. (*ii*) The types of genes compared are different. In the gene trees, an evolutionary distance between two organisms is estimated by the substitution rates between amino acid sequences of the selected genes common among the study organisms, and thus the vast majority of genes, most of which are not common among the study organisms, are ignored and assumed to not contribute to the evolution of the organisms. On the other hand, the genome trees assume that all genes, common as well as noncommon among the study organisms, may contribute to evolution, and the evolutionary distance between any two organisms is estimated by Jensen-Shannon divergence (JSD) of the FFPs.

Complete Proteome Sequence Information and FFP. The complete sequence information within a whole proteome can be described in more than one way. The most obvious way is a collection of the ordered amino acid sequences of all predicted proteins in the whole genome. Another way is a feature count profile (FCP) of the proteome, the counts of all computationally generated unique features, which are the overlapping short sequence-fragments of an optimal length (see first section of *Results*). Such features are generated by a sliding window of the optimal width along the entire length of each protein sequence of the proteome, where two neighboring features overlap completely, except for one terminal residue at the opposite end of each feature (9) (*Materials and Methods*). A collection of such almost completely overlapping and "deconstructed" fragments can be reassembled to "reconstruct" the whole-proteome sequence by a process similar to the "shotgun sequence assembly" method (30), with one important difference: the reconstructed whole-proteome sequence from FCP will be identical to the starting sequence without additional gaps or ambiguities. Since FFP is the same as FCP except that all counts are converted to frequencies in FFP, FFP is a convenient way of describing a whole-proteome sequence as a multidimensional vector that allows mathematical comparison of any given two whole-proteome sequences without alignment.

Comparison with Another Alignment-Free Method. A similar alignment-free proteome tree for the fungal proteome sequence, constructed using a different alignment-free method (composition vector or CV) (31, 32), reported that the CV tree supported largely the MSA-based gene trees. This conclusion is different from ours for several reasons. The CV tree: (*i*) was constructed using 82 fungal proteome sequences available in 2007, while we used 244 fungal proteome sequences; (*ii*) did not include any proteome sequences from protists to interrogate the phylogenic position of Microsporidia, as was done in our study; (*iii*) used a feature length of 7 based on bootstrap value in constructing the CVs, which is much shorter than the optimal length of 13 empirically determined for the convergence of the topology of the proteome trees (*Materials and Methods*); (*iv*) used CVs that are modified by subtracting a statistical background; and (*v*) used the correlation distance between their CV vectors to estimate branch distances, but we used JSD (33) in our proteome tree, which is more appropriate for measuring the similarity/difference between two probability distributions such as FFP vectors.

Materials and Methods

The Sources of Fungi Proteome Dataset and Taxonomic Names Used. All publicly available fungi proteome sequences used in this study are obtained from three databases: the National Center for Biotechnology Information (NCBI), MycoCosm of Joint Genome Institute (JGI) fungal portal, and Broad Institute database. We also included all of 71 protists, for which the genome sequences are available at the NCBI, to inquire whether Microspridia groups with protists or fungi. All fungi and protozoa data used were downloaded by July 2015. We manually excluded organelle-derived protein sequences from mitochondria, chloroplasts, or plastids. To remove any proteome sequences derived from largely incomplete genome sequences, we removed those proteomes with the reported number of peptides fewer than that of *E. romaleae SJ-2008*, a member of Microsporidia with, at present, the smallest and completely assembled genome sequence with 1,830 proteins. Altogether, our study dataset contains 244 unique fungal species and 71 protozoan species. For the outgroup we used two prokaryotes, one with a small proteome and the other with a large proteome. In addition, to test a unique feature of the FFP method, an alignment-free comparison that may be able to use a "synthetic" outgroup, the proteome sequences of two fungi were shuffled. For the Fungi+protist population, both outgroups gave the same proteome trees, proving that, when constructing "the proteome Tree of Life" of all organisms, which is in progress, the shuffled sequences can be used as the outgroup of the tree.

For NCBI data, we downloaded the Reference Sequence (RefSeq) database (34, 35), a curated nonredundant sequence database of genomes, transcripts, and proteins, using an FTP (File Transfer Protocol) site. For JGI data that are available publicly, we obtained from "MycoCosm," JGI fungal portal (36, 37), by choosing annotated "Gene Catalog proteins," if listed, or the largest protein file in the list. Finally, we downloaded two fungi data from the "Multicellularity" study by Broad Institute the (38). For the data used for this study and their sources, see Table S1.

All fungi taxonomic names as well as taxon identifier (taxonIDs) in this study are obtained from the NCBI taxonomy site (https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=advisors.) (39). When there is overlapping taxonIDs among the three databases, we chose the proteome sequences from the NCBI.

Whole-Genome DNA Tree vs. Transcriptome Tree vs. Proteome Tree. Fig. S2*A* reveals that the proteome tree is the best for our purpose, because: (*i*) it converges to the most topologically stable tree; and (*ii*) it remains topologically stable as evidenced by the persistently lowest Robinson-Foulds metric (13) for increasing feature lengths, starting from a feature length of about 11. Furthermore, Fig. S2*B* shows that: (*iii*) the distribution of the JSD (see below), used here as a measure of divergence distance, among the FFPs of the proteome sequence is more broadly spread and less skewed toward large divergence distances than in those of the FFPs of transcriptome or genome sequences, thus stabilizing the tree topology. For the above reasons, we chose the feature length of 13 as an optimal feature length to construct the proteome tree in this study.

Proteome Tree Construction Using Whole-Proteome FFP Method. The method has been described previously (9, 11). Briefly, the whole-proteome sequence of a fungus is scanned using a sliding window of an optimal length (9) (*Results*) and the number of occurrences of each unique feature (defined as the peptide sequence of the optimal length) is counted. A high-dimensional vector consisting of the ordered collection of all such counts is used to describe the whole-proteome sequence of the fungus. To normalize different proteome sizes we then convert the counts to frequencies to form FFPs of the proteome sequence of the fungus. Thus, the proteome sequence is equivalent to the FFP of the proteome (see *Discussion*). Finally, we calculate the JSD (33) between two given FFPs as a measure of the difference between the two fungal organisms, and construct the divergence matrix for all fungal pairs in the study population. This divergence matrix is then used to build a proteome tree using a neighbor-joining method called BIONJ (40). We have observed in our earlier studies (11, 12) that the trees constructed using the JSD matrix of FFPs has performed better than other "distances," such as Euclidean distances, cosine distances, or Jensen-Shannon distances (which are the square roots of JSDs) in producing stable trees.

Outgroup for the FFP-Based Proteome Tree. For the outgroup of our study, we tested two types of proteome sequences: the first is the proteome sequences of two prokaryotes of known genome sequence: *Candidatus Portiera aleyrodidarum BT-B-Hrs* (Gram-negative proteobacteria) with the smallest proteome size, 253 proteins, and *Ktedonobacter racemifer* DSM 44,963 (green nonsulfur bacteria) with the largest proteome size, 11,288 proteins, among the prokaryotes in this study; the second consists of two fungal proteome sequences of the smallest (a Microsporidia, *E. romaleae SJ-2008*, with 1,831 proteins) and the largest (a Basidiomycota, *Sphaerobolus stellatus*

SS14, with 35,274 proteins) proteome in the study population, but each proteome sequence is randomly shuffled using "Fisher-Yates shuffle" method (41, 42). In this shuffling process, the amino acid composition and the total length of the proteome of each randomized sequence remain the same as the original sequence.

We used these artificially constructed protein sequences to test their potential utility as an outgroup for a future proteome tree of all organisms. In both cases, the outgroup sequences form a clade, which is separated from all fungal and protozoan clades of the study population. No differences were found in the proteome tree topology and the member composition of clades regardless of whether we used all of them together, each separately, or unshuffled sequences of the two prokaryotes (results not shown).

Computer Code Availability. The FFP programs for this study, written in GCC (C++), will be deposited in GitHub: https://github.com/jaejinchoi/FFP.

ACKNOWLEDGMENTS. We thank Prof. John Taylor (University of California, Berkeley) for helpful advice and discussions, and Prof. Johannes Kusters (University Medical Center, Utrecht) for comments and suggestions. The genomic sequence data of JGI used in this work were produced by the US Department of Energy Joint Genome Institute (www.jgi.doe.gov/) in collaboration with the user community. This research was partly supported by a grant from World Class University Project, Ministry of Education, Science and Technology, Republic of Korea (to S.-H.K.); financial support from Research Initiative Program of Korea Research Institute of Bioscience and Biotechnology, Daejeon, Republic of Korea (JJ.C.); and a gift grant to University of California, Berkeley (to support JJ.C.).

1. Stajich JE, et al. (2009) The fungi. *Curr Biol* 19:R840–R845.
2. Taylor JW, et al. (2004) The fungi. *Assembling the Tree of Life*, eds Cracraft J, Donoghue MJ (Oxford Univ Press, New York).
3. Petersen JH (2013) *The Kingdom of Fungi* (Princeton Univ Press, Princeton, NJ).
4. James TY, et al. (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* 443:818–822.
5. Hibbett DS, et al. (2007) A higher-level phylogenetic classification of the Fungi. *Mycol Res* 111:509–547.
6. Moreira D, Philippe H (2000) Molecular phylogeny: Pitfalls and progress. *Int Microbiol* 3:9–16.
7. Heath T, Hedtke S, Hillis D (2008) Taxon sampling and the accuracy of phylogenetic analyses. *J Syst Evol* 46:239–257.
8. Hillis DM, Pollock DD, McGuire JA, Zwickl DJ (2003) Is sparse taxon sampling a problem for phylogenetic inference? *Syst Biol* 52:124–126.
9. Sims GE, Jun SR, Wu GA, Kim SH (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci USA* 106: 2677–2682.
10. Deerwester S, Dumais ST, Furnas GW, Landauer TK (1990) Indexing by latent semantic analysis. *J Am Soc Inf Sci* 41:391–407.
11. Jun SR, Sims GE, Wu GA, Kim SH (2010) Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proc Natl Acad Sci USA* 107:133–138.
12. Sims GE, Jun S-R, Wu GA, Kim S-H (2009) Whole-genome phylogeny of mammals: Evolutionary information in genic and nongenic regions. *Proc Natl Acad Sci USA* 106: 17077–17082.
13. Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. *Math Biosci* 53: 131–147.
14. Felenstein J (1989) PHYLIP-phylogeny inference package (version 3.2). *Cladistics* 5:163–166.
15. Siddall ME (1995) Another monophyly index: Revisiting the Jackknife. *Cladistics* 11: 33–56.
16. Keeling PJ, Slamovits CH (2004) Simplicity and complexity of microsporidian genomes. *Eukaryot Cell* 3:1363–1369.
17. James TY, et al. (2013) Shared signatures of parasitism and phylogenomics unite Cryptomycota and microsporidia. *Curr Biol* 23:1548–1553.
18. Keeling PJ, Luker MA, Palmer JD (2000) Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. *Mol Biol Evol* 17:23–31.
19. Vossbrinck CR, Maddox JV, Friedman S, Debrunner-Vossbrinck BA, Woese CR (1987) Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature* 326:411–414.
20. Thomarat F, Vivarès CP, 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.
21. Gruninger RJ, et al. (2014) Anaerobic fungi (phylum Neocallimastigomycota): Advances in understanding their taxonomy, life cycle, ecology, role and biotechnological potential. *FEMS Microbiol Ecol* 90:1–17.
22. Griffith GW, et al. (2010) Anaerobic fungi: Neocallimastigomycota. *IMA Fungus* 1:181–185.
23. Youssef NH, et al. (2013) The genome of the anaerobic fungus *Orpinomyces* sp. strain C1A reveals the unique evolutionary history of a remarkable plant biomass degrader. *Appl Environ Microbiol* 79:4620–4634.
24. Eichinger L, et al. (2005) The genome of the social amoeba *Dictyostelium discoideum*. *Nature* 435:43–57.
25. Urushihara H, et al. (2015) Comparative genome and transcriptome analyses of the social amoeba *Acytostelium subglobosum* that accomplishes multicellular development without germ-soma differentiation. *BMC Genomics* 16:80.
26. Padamsee M, et al. (2012) The genome of the xerotolerant mold *Wallemia sebi* reveals adaptations to osmotic stress and suggests cryptic sexual reproduction. *Fungal Genet Biol* 49:217–226.
27. Zalar P, Sybren de Hoog G, Schroers H-J, Frank JM, Gunde-Cimerman N (2005) Taxonomy and phylogeny of the xerophilic genus *Wallemia* (Wallemiomycetes and Wallemiales, cl. et ord. nov.). *Antonie van Leeuwenhoek* 87:311–328.
28. Bergsten J (2005) A review of long-branch attraction. *Cladistics* 21:163–193.
29. Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* 9:605–618.
30. Staden R (1979) A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* 6:2601–2610.
31. Qi J, Luo H, Hao B (2004) CVTree: A phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res* 32:W45–W47.
32. Wang H, Xu Z, Gao L, Hao B (2009) A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evol Biol* 9:195.
33. Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory* 37:145–151.
34. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I (2014) RefSeq microbial genomes database: New representation and annotation strategy. *Nucleic Acids Res* 42: D553–D559.
35. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33:D501–D504.
36. Grigoriev IV, et al. (2014) MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Res* 42:D699–D704.
37. Grigoriev IV, et al. (2012) The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res* 40:D26–D32.
38. Broad Institute (2012) Origins of Multicellularity Sequencing Project and Fungal Genome Initiative Broad Institute of Harvard and MIT. Available at https://www.broadinstitute.org/annotation/genome/multicellularity_project/Credits.html. Accessed July 2015.
39. Sayers EW, et al. (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 40:D13–D25.
40. Gascuel O (1997) BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 14:685–695.
41. Fisher RA, Yates F (1948) *Statistical Tables for Biological, Agricultural and Medical Research* (Oliver and Boyd, London).
42. Knuth DE (1973) Seminumerical algorithms. *The Art of Computer Programming* (Addison-Wesley, Boston), 3rd Ed.
43. Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 44:W242–W245.
44. R Core Team (2016) R: A language and environment for statistical computing. Available at https://www.r-project.org/. Accessed April 2016.
45. Gribaldo S, Philippe H (2002) Ancient phylogenetic relationships. *Theor Popul Biol* 61: 391–408.