Synthetic and Systems Biotechnology 4 (2019) 225–231

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: http://www.keaipublishing.com/synbio

Resolution of deep divergence of club fungi (phylum Basidiomycota)

Hongliang Mao, Hao Wang*

T-Life Research Center, Department of Physics, Fudan University, Shanghai, 200433, PR China

ARTICLE INFO

Keywords:
Fungi
Basidiomycota
Phylogenetics
Phylogenomics
CVTree

ABSTRACT

A long-standing question about the early evolution of club fungi (phylum Basidiomycota) is the relationship between the three major groups, Pucciniomycotina, Ustilaginomycotina and Agaricomycotina. It is unresolved whether Agaricomycotina are more closely related to Ustilaginomycotina or to Pucciniomycotina. Here we reconstructed the branching order of the three subphyla through two sources of phylogenetic signals, i.e. standard phylogenomic analysis and alignment-free phylogenetic approach. Overall, beyond congruency within the frame of standard phylogenomic analysis, our results consistently and robustly supported the early divergence of Ustilaginomycotina and a closer relationship between Agaricomycotina and Pucciniomycotina.

1. Introduction

The club fungi (phylum Basidiomycota), encompasses more than 30,000 described species distributed in almost all terrestrial and some aquatic (freshwater and marine) habitats [1], form the second largest phylum in the Kingdom Fungi. Basidiomycota is typically characterized by the presence of basidia (singular: basidium), swollen terminal cells of hyphae bearing sexual spores. The most familiar members of this phylum are edible mushrooms and jelly fungi, as well as puffballs, stinkhorns, shelf and bracket fungi. Other members include pathogens that attack crops and animals, endophytes that enhance host growth, and wood-rotting decomposers in forests that play important role in the carbon cycle. Overall, Basidiomycota have huge impacts on human culture, economy and ecosystem functions [2].

The introduction of molecular techniques since early 1990s has well established that the Basidiomycota has been strongly supported as the sister clade of the phylum Ascomycota ([3,4]; see summary in Ref. [5]). Furthermore, within Basidiomycota, three major clades, i.e. the subphyla Agaricomycotina, Pucciniomycotina and Ustilaginomycotina, have also gained strong support from diverse analyses ([6,7]; see summary in Ref. [5]). Agaricomycotina includes two thirds of described Basidiomycota, including mushrooms, jelly fungi, basidiomycetous yeasts, wood decayers, litter decomposers, and ectomycorrhizal fungi, along with important pathogens of timber, vegetable crops, and human [8]. The vast majority of species in Pucciniomycotina and Ustilaginomycotina are parasitic plant rusts [9] and smuts [10], respectively. However, the order of branching of the three subphyla is still unknown [8,11,12]. On this issue, two hypotheses have been proposed, but none of them have gained comprehensive support yet.

The hypothesis of Pucciniomycotina early suggests that the basal split differentiated Pucciniomycotina from the group of Ustilaginomycotina + Agaricomycotina. This hypothesis was proposed in middle 1990s based on investigation of variation of ultrastructure of septal pores and spindle pole bodies [13] and was observed in 18 ribosomal RNA (rRNA) molecular phylogenetics [6]. This relationship has also been recovered in several other later rRNA or protein coding genes analyses, and been believed to be compatible with ultrastructural characters, cell wall biochemistry and 5S rRNA secondary structure (see summary in Ref. [8]). However, none of these early molecular analyses based on single or a few loci could provide statistically sound support to this hypothesis.

Alternatively, the early divergence of Ustilaginomycotina which takes Ustilaginomycotina as the basal lineage and groups Agaricomycotina and Pucciniomycotina has also been recovered by some recent protein-coding gene phylogenomics, e.g. Ref. [14]. However, these works have not provided convincing statistical support at this specific relationship either.

Recent sequencing technique revolution has generated genomes and especially gene repertoire of a considerable of fungal organisms distributed in broad taxa range, which has stimulated waves of genome-wide phylogenomic analyses in the kingdom fungi (see e.g. Refs. [15,16]). Many of these analyses have used a strategy of concatenating multiple orthologous genes into a single supermatrix and then analyzing the supermatrix using standard tree construction methods (called standard phylogenomics hereafter). Concatenation of multiple loci in many cases generated trees with high confidence indices, e.g. bootstrap

Peer review under responsibility of KeAi Communications Co., Ltd.
* Corresponding author. T-Life Research Center, Department of Physics, Fudan University, Shanghai, 200433, People's Republic of China.
E-mail address: ynwh.km@gmail.com (H. Wang).

https://doi.org/10.1016/j.synbio.2019.12.001
Received 13 August 2019; Received in revised form 18 November 2019; Accepted 4 December 2019

value (BP) and posterior probability (PP). In 2012, Ebersberger and colleagues' work first supported the Pucciniomycotina early hypothesis with high confidence values through standard phylogenomics performed on a dataset of 152 protein coding genes [17].

It has been well-documented that resulting trees of different standard phylogenomic studies often conflict in topology (see brief reviews in Refs. [18,19]). Incongruence may derived from every step in the reconstruction, from dataset selection to perturbation in analytical procedure to analytical bias of tree reconstruction methods (see summary in Ref. [18]). Specifically, using different datasets and/or using different methods to clean data on the same dataset may generate conflict trees with strong statistical support in terms of confidence indices like BP and Bayesian PP [20]. Because no prior knowledge about the true phylogeny is available, the only way to assess the validity of reconstructions lies in the consistency of resulting phylogenetic trees. The consensus is that the topology recovered by more independent reconstructions is more likely to be reliable. Therefore, consistent evolutionary reconstruction based on qualified independent methodologies is crucial in sorting out true species trees.

The accessibility of genomic information from broad taxa has made it feasible to extract phylogenetic information from genomic elements other than the alignment of rRNA and/or protein-coding genes. These efforts are expected to greatly improve our understanding of evolution by providing evidence from dimensions other than standard phylogenomics that based on alignment and concatenation of genes. Complementary to standard phylogenetic analysis, alignment-free sequence comparison methods have been advanced significantly during genome sequencing revolution (see a brief summary in Ref. [21]). Composition vector tree (CVTree) [22] is one of these booming methods. Instead of exploiting site substitution, CVTree profiles difference of sequences through the effective frequency of short strings. As a good independent verification of standard phylogenetic reconstruction, this method and its variants have been successfully used to resolve the evolutionary relationships of many major divisions of life, including virus [23], archaea and bacteria [24], fungi [25] and animals [26].

Here we tackle the early branching within Basidiomycota. We have performed a standard phylogenomic analysis includes 171 orthologous protein coding gene groups from 91 fully sequenced fungal genomes, and built a CVTree based on the whole-gene repertoire of these fungi. Results of the two analyses all consistently and robustly support that a sister relationship between Agaricomycotina and Pucciniomycotina and the early divergence of Ustilaginomycotina.

2. Materials and methods

2.1. Sequence data

The gene annotations (coding sequence and peptide sequences) of the 216 sequenced fungi genomes (87 Basidiomycota and 129 Ascomycota) were downloaded from DOE-JGI website (JGI Fungi Portal: http://genome.jgi.doe.gov/programs/fungi/index.jsf). Annotations of two strains of *Cryptococcus gattii* (strain r265_1 and w276_1) were downloaded from Broad Institute (http://www.broadinstitute.org/scientific-community/data).

2.2. Identification of single-copy orthologous gene clusters

We collected whole-gene sets of 218 sequenced Dikarya genomes (89 Basidiomycota and 129 Ascomycota) and generated 119,208 gene clusters using OrthoMCL version 5 program [27] with default parameter settings. We then selected clusters meeting the following requirements [1]: member genes of clusters occurred at most once in each of the 218 organisms [2]; member genes occurred in at least 80% of the Basidiomycetes genomes [3]; member genes occurred in at least 50% of the Ustilaginomycetes and Pucciniomycetes genomes, respectively [4]; Every cluster had single-copy in every species. We required [3] because currently available genomes in the three sub-phyla are quite unbalanced (67 in Agaricomycotina, but only 14 in Pucciniomycotina and 8 in Ustilaginomycotina; see Table S1). The above requirements [2–4] selected gene clusters with reasonable taxon-sampling in the three subphyla of Basidiomycota. Finally a total of 171 gene clusters were used for phylogenomic reconstruction (Supplementary Table S1).

Based on current version of gene annotation of *Postia placenta MAD-698-R*, its orthologus genes occurred in only 52 of the 171 single-copy gene clusters while genes from all other genomes occurred in at least 133 clusters. Because this genome was underrepresented in our final single-copy gene clusters and a closely related genome *Postia placenta MAD-698-R-SB12* was included in this analysis, we excluded *Postia placenta MAD-698-R* from further analysis.

2.3. Outgroups selection

We selected 3 Ascomycota organisms *Saitoella complicata*, *Pyrenophora tritici-repentis* and *Aplosporella prunicola* as outgroup species because they showed the best gene representation in the orthologs set (the three species had 122, 103 and 102 genes occurred in the 171 clusters, respectively) and represented the Ascomycota well (from two subphyla). Our final dataset consisted of 171 single-copy orthologous genes clusters from 91 genomes (88 Basidiomycota +3 Ascomycota).

2.4. Construction of supermatrix

Peptide sequences of member genes of the 171 clusters were aligned using MUSCLE (version 3.8.31 [28]). The resultant alignments were manually inspected and concatenated to yield an alignment consisting of on average 72,251 residues per sequence amounting to 196,810 aligned positions. TrimAl (version 1.2 [29]) was run with parameter setting of "-autometed1" to exclude low quality regions and the reduced supermatrix of 34,658 columns were used for standard phylogenomic reconstruction.

2.4.1. Standard phylogenomic analysis

Model selection using ProtTest program (version 3.3 [30], with "-all-matrices" and "-all-distributions") showed the combination of LG + G + I + F fitted our dataset best according to both the AIC and BIC criterion, where G, I and F present that parameters of Gamma distribution, proportion of the invariable sites and amino acid frequencies were estimated from the alignment, respectively. Besides the top-ranking LG model, we also built trees with the second- and third-ranking models (WAG + G + I and JTT + G).

For three models, we built the Maximum likelihood (ML) tree using RAxML (version 7.8.5 [31]), the Bayesian (Bayes) tree using MrBayes version 3.2.2 [32], and the Neighbor-joining (NJ) tree was constructed using the MEGA package (version 6.0 [33]). Bayesian tree search was conducted in 2 parallel runs (4 chains) for 1 million generations and trees were sampled every 500 generations, with the first 25% discarded as burn-in. The NJ tree was conducted with missing data treatment option of 'pairwise deletion'. We also built Maximum Parsimony (MP) tree using MEAG version 6.0, invoking 'used all sites' and 'Subtree-Pruning-Regrafting' heuristic tree searching method. All reconstructions called bootstrapping 1000 replicates.

We used the CONSENSE program in the PHYLIP package (version 3.69, http://evolution.genetics.washington.edu/phylip.html) to generate the majority rule consensus tree joining ML, NJ, MP and Bayes trees.

To evaluate statistic support for all three possible relationships among the three subphyla of Basidiomycota, we used CONSEL package version 1.2 [34] to perform tree topology selection and calculate p-values for topologies. Based on the 171-gene ML tree (Fig. S3), we shuffled the relationships between the three subphyla to generate three topologies (Table S4, column 1), and estimated per site log likelihoods for the three topologies by RAxML under the LG + G + I + F model,

then conducted tests with CONSEL.

2.5. Detection of genes with sequence bias

We used TrSpEx [35] to evaluate the effects of potential sequence bias. In TrSpEx, the major variable that helped to identify long-branch (LB) genes was LB score, which measured for each taxon the percentage deviation from the average pairwise distance between taxa. Patristic distance (PD) was the total length of branches that linked two taxa in a given gene tree. From the mean value of pairwise PD of taxon i (PD_i) to all other taxa and the average pairwise PD across all taxa in the tree (PD_a), LB score of taxon i was calculated as ($PD_i - PD_a$)/PD_a). Two indices derived from LB score, i.e. mean value of upper quartile of LB scores and standard deviation of LB scores, were taken as measures of taxa with the longest branches and heterogeneity, respectively. The rationale of TreSpEx was that genes that had unexpectedly high values of upper quartile of LB scores or standard deviation of LB scores were taken as having long branches. For the 171 gene clusters, we drew density plots of the two indices and took genes located in right shoulder regions (regions at which the curve were deviated from normal distribution) as candidate long-branch genes (Figs. S12a–S12c). TrSpEx estimated saturation of genes using the linear regression of PD against uncorrected distances. The uncorrected distance was simply the number of different residues in sequences without considering multiple substitutions. A low R2 by the regression indicates that the uncorrected distance could not explain a gene's PD, which means the gene may had multiple substitutions. Therefore, genes were taken as being affected by saturation if they had unexpectedly low values of R^2 or slope of best-fit regression line. For the 171 gene clusters, we drew the distributions of R^2 and the slope and took genes located in the left shoulder regions as candidates with saturation (Fig. S12d - S12df). We used stringent and loose criteria to define shoulder regions of these distributions.

2.6. CVTree construction

The CVTree phylogenetic reconstruction method was first described in 2004 [22]. This method profiled organisms using the so-called composition vectors (or CVs). A CV of length K was basically the distribution of the frequency of oligopeptides of length K in the whole-protein set of an organism, but modified through the subtraction of a background distribution of frequency generated by a Markov model of order (K − 2). The distance between two organisms was then calculated by (1 − C)/2, where C was the correlation between two organisms, which was determined by taking the projection of one CV on another, i.e. the cosine of the angle between them. Specifically, if two vectors were identical, they had the highest correlation of C = 1; while if they had no components in common, C = 0 and the two vectors were orthogonal to each other. Subsequently, distance matrix could be calculated and the phylogeny was constructed by standard NJ method. In short, CVTree was a distance based method but estimated sequence distance in a new way.

CVTrees base on whole-gene sets and sub-gene sets were constructed using the CVTree webserver version 2.0 (tlife.fudan.edu.cn/cvtree [36]), with amino acids sequences as input, K-tuple length = 7 and bootstrapping 100 replicates. The sub-gene sets selection was performed using an in-house Perl script.

2.7. Gene ontology analysis

The GO annotations of gene clusters were retrieved and visualized using blast2GO (version 2.5 [37]) based on their best annotated BLASTP hit (e-value = 1e-10) in the nr database. Gene-enrichment analysis of yeast genes were performed at DAVID website (http://david.abcc.ncifcrf.gov/), choosing Saccharomyces Genome Database as background.

3. Results

3.1. Single-copy orthologous genes

We conducted standard phylogenomics analysis using conserved single-copy genes. Our ortholog-selection methods identified 171 single-copy orthologous gene clusters with relative balanced taxonomic distribution in sequenced Basidiomycota organisms (see Methods). These clusters included in 14,633 genes from 91 Dikarya (88 Basidiomycota and 3 Ascomycota outgroups) genomes (Supplementary Table S1). The average length of peptide sequences of these genes was 450, with the maximum and minimum length of 2,859 and 49, respectively (Supplementary Fig. S1). The number of genomes contained in these clusters ranged from 70 to 91, with an average of 86 (95%). Not considering outgroup organisms, the number of gene clusters belong to one genome ranged from 133 (*Malassezia sympodialis*) to 171 (10 genomes), with an average of 161 (94%).

We performed structural, conservative, and functional annotation investigations of these clusters. Genes in 166 clusters had at least one known protein domains (Pfam version 27.0 search using pfam_scan.pl downloaded from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/, e-value = 1e-5, Table S2). 124 of the 166 clusters hit genes with functional annotation, i.e. genes were not annotated as putative or hypothetical proteins, in NCBI NR database (http://blast.ncbi.nlm.nih.gov/) in terms of BLASTP search with e-value = 1e-10, and requiring the best high-scoring segment pairs cover at least 50% of query fungi genes. 20 clusters had homologs in animals or plants, suggesting their conservation in eukaryotic evolution.

The blast2GO [37] analysis on the 171 clusters showed that they belonged to a wide spectrum of functional categories (Fig. S2). Yeast (*Saccharomyces cerevisiae S288C*) genome had so far the best gene ontology (GO) annotations in the kingdom fungi. 78 out of the 171 clusters could find single-copy member genes in this genome. Functional annotation analysis of these genes using DAVID functional annotation Tool (http://david.abcc.ncifcrf.gov/) indicated that genes related to functions of ribosome (rRNA processing and ribosome biogenesis) were overrepresented (FDR correction) in our data set (Table S3).

3.2. Standard phylogenomics

We built a supermatrix of 34,658 aligned sites based on amino acids alignments of the 171 single-copy orthologous clusters (Supplementary dataset 1; see Methods). On average, missing data account for 6% of concatenated genes. ML, NJ, MP, and Bayes methods were used to construct trees from the supermatrix. For each of 4 tree construction method, we built trees using the top 3 substitution models (LG + *G* + *I* + *F*, WAG + *G* + *I* and JTT + *G*), and used the best supported trees (Figs. S3–S6) for consensus tree construction. Fig. 1 showed the higher-than-order-level relationships of the phylum obtained by majority rule consensus analysis of the above 4 phylogenies. The full phylogeny is shown in Fig. S7. All four methods consistently supported the early divergence of Ustilaginomycotina and grouping of Agaricomycotina + Pucciniomycotina with high BPs (ML: 94%; NJ: 100%; MP: 98%) and PP (1.0).

3.3. The topology is robust to perturbation in analytical procedure

To evaluate the effects of data treatments and sequence bias on the topology, we constructed several sub-gene-sets based on the 171 gene clusters and compared resulting phylogenetic hypotheses of these datasets to our main phylogenetic trees. First, we generated a dataset without missing genes. Because of uneven taxon representation of the three subphyla, we required the dataset contained at least 6 and 2 genomes from Pucciniomycotina and Ustilaginomycotina, respectively; and we also required that at least one outgroup species should be selected. We first randomly selected 9 genomes that met the above

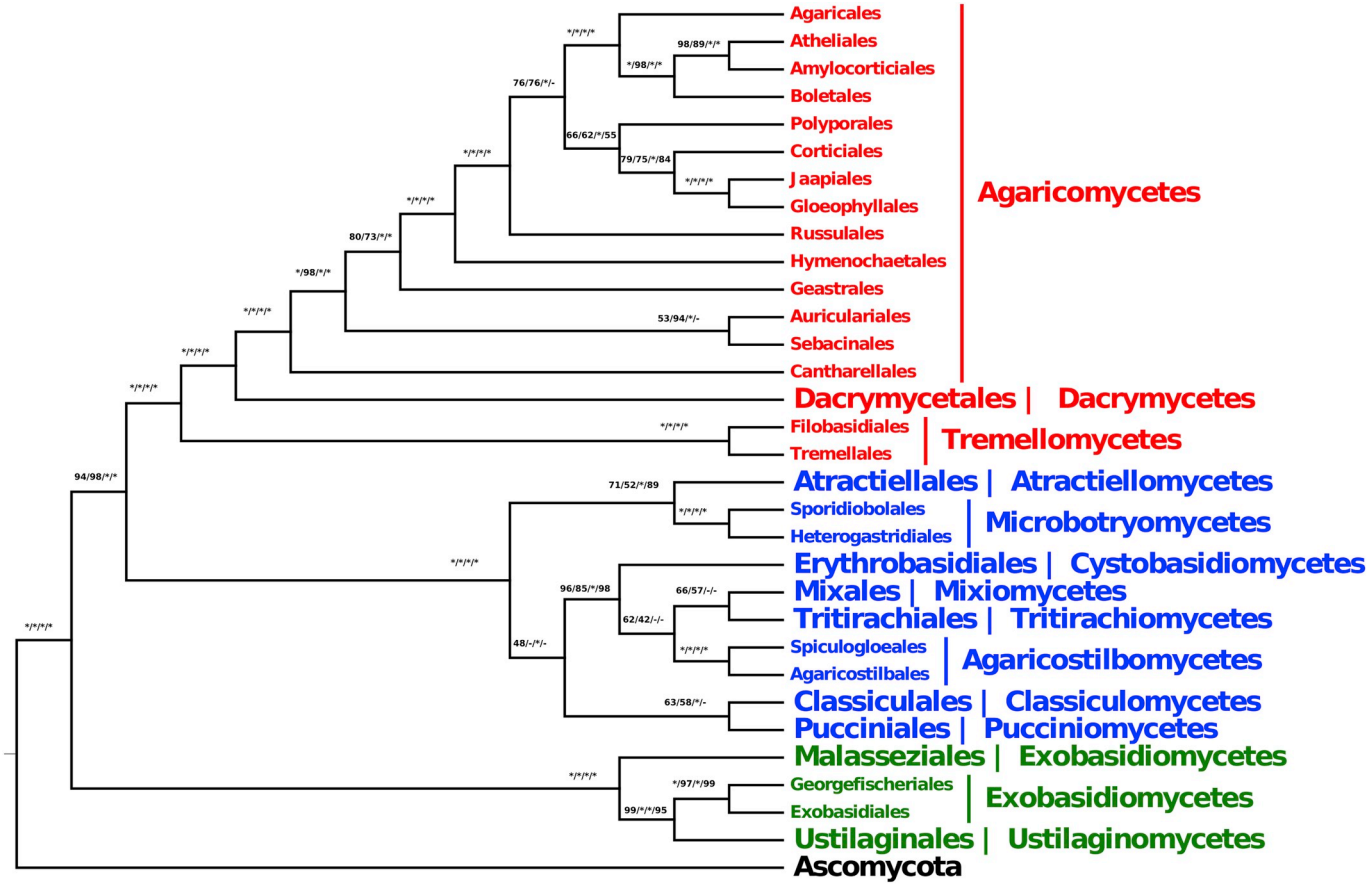

Fig. 1. The most likely evolutionary relationship involving the three subphyla of Basidiomycota. The tree shows higher-than-order level relationships of the majority rule consensus of 4 standard phylogenomic reconstructions of 91 genomes using 171 gene clusters. Statistical supports are numbers attached to nodes: from left to right, BP of ML, BP of MP, PP of Bayes and BP of NJ. Bootstrap values are reported as percentages. "*" indicates BP = 100% or PP = 1. "-" indicates the node is not recovered by corresponding tree reconstruction method.

conditions, than identified all clusters in which the 9 genomes occurred and recording all species that occurred every of these clusters, at last we kept genes that were from these species. The final dataset was composed of 75 clusters of which each cluster contained identical 42 species, including 30 Agaricomycotina, 8 Pucciniomycotina, 3 Ustilaginomycotina and 1 outgroup species, respectively. The resulting supermatrix contained 21,922 sites (Supplementary dataset 2).

Second, we excluded poorly aligned genes. For each of the alignment of 171 gene clusters, we first removed columns of which amino acids identity < 50%, then checked length of each retained sequence and removing those contained less than 50% alignment coverage. At last, we obtained a gene set of high alignment quality which consisted of 64 clusters with on average 85 species occurring in each cluster. The resulting supermatrix contained 6,920 sites (Supplementary dataset 3).

In the third evaluation we selected gene clusters with strong phylogenetic signals in terms of bootstrap consensus trees showed high average support across all internodes, according to suggestions in Ref. [18]. There were 41 and 13 clusters of which the average BP across all internodes of ML consensus trees ≥ 60% and 70%, respectively. The supermatrices of the two datasets had 20,376 and 8,011 sites, respectively (Supplementary dataset 4 and 5).

All ML trees based on aforementioned data treatments (Figs. S8–S11) were highly consistent with our major phylogeny (Fig. 1), and most trees supported that Ustilaginomycotina was the most early divergent among the three clades at rather high bootstrap values. The only exception was the tree based on the 13 clusters that had ≥ 70% average internode BP (Fig. S11). This phylogeny recovered the same topology, but the BP at the Agaricomycotina + Pucciniomycotina clade was only 58%. Overall, these evaluations suggested that the branching order recovered by the 171 gene clusters was quite stable to these treatments.

3.4. The topology is robust to sequence bias

We also evaluated the effects of sequence bias on the relationships recovered above. Here bias means long branch attraction and mutational saturation. TreSpEx [35] calculates the distributions of some indices for a multiple-gene dataset and takes genes that located at skewed regions of these distributions as biased genes. The indices includes [1] mean value of upper quartile of LB score, LB score measures how different the distance between two terminal nodes is from the average pair-wise distance of all terminal nodes [2]. standard deviation of LB score [3], slope of the linear regression between patristic and uncorrected pairwise distances of gene clusters and [4] R^2 of linear regression between patristic and uncorrected distances of gene clusters (see Methods; also see Ref. [35] for detail explanation of these indices). We used TreSpEx to detected genes that potentially cause long branch attraction and mutational saturation (Table S6), removed these genes and reconstructed ML phylogeny using remaining unbiased genes. Under both stringent and loose criteria (Fig. S12), the trees of all tests supported the early divergence of Ustilaginomycotina (Fig. S13), which suggested that this relationship might not be resulted from long-branch attraction and/or mutational saturation.

Relationship between three lineages could generate three topologies in total. To estimate statistical support for other possible alternative relationships among the three subphyla, we performed approximately

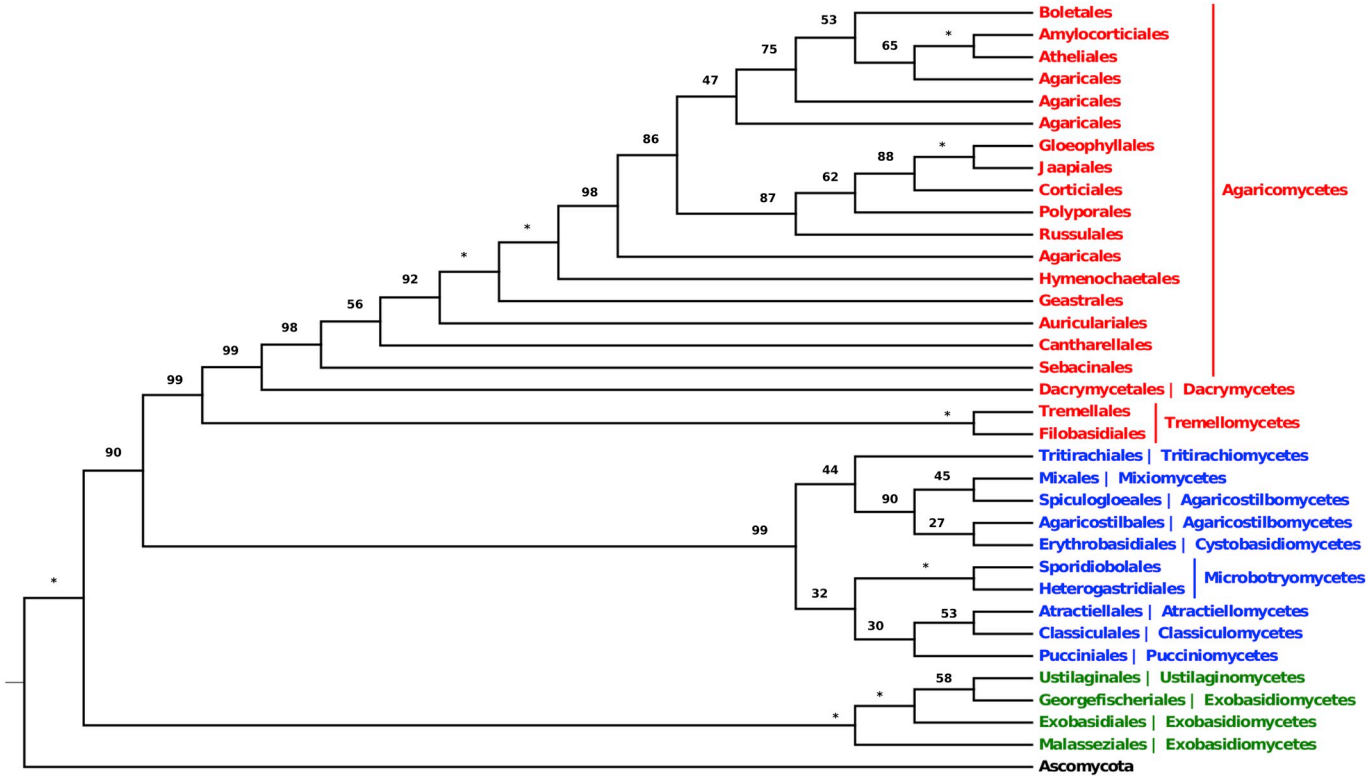

Fig. 2. The inferred CVTree of the 91 fungi genomes. This tree is obtained using whole-gene repertoires of these genomes with K = 7. Bootstrap values are reported as percentages. "*" indicates BP = 100%. Only higher-than-order level relationships are shown here. See Fig. S14 for the full CVTree.

unbiased test for the three possible sub-phyla level topologies by site bootstrapping as implemented in CONSEL version 1.2 [34]. This software identifies the top ranking topology for alternative tree hypotheses under the likelihood criterion by calculating p-values. We tested our dataset using default parameters and the result showed that ((Agaricomycotina, Pucciniomycotina), Ustilaginomycotina) as the best fitting topology (Table S4). All other tests equipped in CONSEL also support this result (Table S5).

3.5. CVTree phylogeny and robustness

The CVTree (Fig. 2 and Fig. S14) based on whole-protein sets of the 91 genomes and K = 7 (see Ref. [38] for details of determination of K) supported the early divergence of Ustilaginomycotina and the sister relationship between Agaricomycotina and Pucciniomycotina at high statistical confidence (BP = 90%). This result is different from our previous work in 2009 (*25*). At that time only 12 Basidiomycota genomes were available and the CVTree grouped Agaricomycotina and Ustilaginomycotina as sister clades with a low BP of 35% (Fig. 1 of [25]). These results highlight the importance of broad taxon sampling for resolving deep nodes.

To evaluate whether this result was sensitive to the selection of gene set, we generated sub-gene sets by random sampling without replacement to a certain percentage of genes from each of the 91 whole-gene sets and constructed CVTrees with these datasets. We changed the percentage from 60% to 95%, with 5% increase at each step, and repeated the sampling 100 times at each percentage value. All datasets strongly support the early divergence of Ustilaginomycotina (Table S7). These results suggest that this topology recovered by CVTree is not sensitive to the selection of genes.

3.6. Other highly supported deep relationships within basidiomycota

Both Standard phylogenomics and CVTree could not resolve relationships within subphylum Pucciniomycotina with high confidence. However, our results provided insights in early divergence of Ustilaginomycotina and Agaricomycotina. Both standard phylogenomics and CVTree suggested that Exobasidiomycetes was not monophyletic groups. In current taxon sampling, the basal split of Ustilaginomycotina occurred at the order Malasseziales and other organisms, including two orders currently considered as members of class Exobasidiomycetes. Our dataset include two species within genus *Malassezia*, which is monotypic within the family Malasseziaceae, which is itself monotypic within the order Malasseziales if excluding environmental samples. Therefore the two species represent this order quite well. The relationships of Exobasidiales (represented by one species), Georgefischeriales (one species) and Ustilaginomycetes (3 genus from order Ustilaginales) were inconsistent between standard phylogenomics and CVTree: the former grouped Exobasidiales and Georgefischeriales and placed Ustilaginomycetes as its sister clade with quite high confidence indices (Fig. 1), while the later recovered the sistership between Exobasidiales and the poorly supported (BP = 58%) Georgefischeriales + Ustilaginomycetes clade (Fig. 2).

Both standard phylogenomics and CVTree agreed that the deep divergence of subphylum Agaricomycotina was (Tremellomycetes, (Dacrymycetes, Agaricomycetes)), which is consistent with several previous multi-locus analysis ([2]; see review in Ref. [8]). Within the Agaricomycetes, our standard phylogenomics strongly supported Cantharellales as the sister clade of all other Agaricomycetes species (Fig. 1). In contrast, CVTree placed Sebacinales as basal lineage and grouped Cantharellales with the rest of Agaricomycetes, but the later grouping only gained rather low BP support (Fig. 2). We note that the early divergence of Agaricomycetes has proven difficult and the topologies recovered by previous works depended heavily on dataset and method chosen (see for example [2]). Though our standard phylogenomics of study seems in favor of the early divergence of Cantharellales, this hypothesis has not gain comprehensive support yet, and more analyses are required to confidently resolve this question.

4. Discussions

In contrast to the popular hypothesis that Pucciniomycotina represents the earliest diverged lineage among the three subphyla of Basidiomycota, our results consistently support the early divergence of Ustilaginomycotina. This result is strongly supported by all phylogenetic reconstructions (Figs. 1 and 2, Fig. S3-S7 and S14) and is robust to dataset treatments, sequence bias and gene selection (Fig. S8-S11 and S13, Table S7).

A recent research discovered that Ngaro elements, a Basidiomycota specific group of Tyrosine Recombinase-encoding retrotransposons, only occurred in Pucciniomycotina and Agaricomycotina, but was absent in all Ustilaginomycotina genomes investigated [39]. The most parsimonious explanation of this pattern is that the origin of fungal Nagro elements happened in the common ancestor of Agaricomycotina and Pucciniomycotina after the differentiation of Ustilaginomycotina, which supports the early divergence of Ustilaginomycotina from another independent source of evidence.

Many practices of phylogenomic/phylogenetic analyses seek to reconstruct evolutionary history of a set of organisms using one dataset, either consisting of an individual locus or a concatenation of multiple loci in the genomes. The belief that the topology of all nodes can be resolved simultaneously using one or few datasets means that there exists one dataset of which phylogenetic signals are strong enough for all nodes. However, the existence of such a dataset is not self-evident, especially when the investigated organisms had complex evolutionary history. Recent works have explicitly suggested nodes that underwent highly heterogeneous evolutionary history should be treated separately by using different datasets (see e.g. Refs. [17,18] for two examples about fungi). Though the efforts of pursuing the omnipotent dataset should by no means be banned, we believe that a node-by-node strategy represents a realistic choice in evolutionary reconstructions, at least at current stage. According to this, instead of discussing all nodes with high confidence values (we did obtained many such nodes, see Figs. 1 and 2), we carefully evaluate the deepest nodes in the club fungi phylogeny in this research.

When multiple datasets are used, the next question is to assess which relationship is valid for a specific node because different datasets may give conflicting reconstruction, and this problem of incongruence has been extensively documented by numerous works (see review in Ref. [19]). Since no objective criterion is available, consistency is the only way to assess the validity of phylogenies. In practice, consistency within dataset is usually evaluated and the dataset resulting in consistent trees with different tree reconstruction methods and/or data treatments are considered better than the dataset resulting in inconsistent trees. However, the problem still remains. Even when congruence is reached within one dataset, it is still possible that other datasets congruent to a different topology. For example, both our dataset and the Basidiomycota_1 of [17] showed very high level of self-inconsistency, but they support conflicting topologies regarding the divergence of the three subphyla.

When this happens, the principle of consistency requires that phylogenetic reconstruction based on independent sources of the evidence should be introduced to judge conflicting phylogenetic hypotheses. Here we introduce CVTree, a qualified method that extracts phylogenetic signals in a different way from standard phylogenomics. The CVTree method constructs phylogeny in two steps similar to standard NJ method. That is [1], generate a distance matrix representing pairwise relationship between studied organisms and [2] tree construction using the NJ algorithm. The core feature of CVTree lies in the first step. It generates distance matrix through an alignment-free strategy: CVTree measures the difference between sequences in term of the effective occurrence of short strings instead of site substitution in standard phylogenetics. This feature makes this method measure sequence similarity without introducing any parameters. The parameter-like “K-tuple length” is not a free parameter but decided by genome size investigated (*38*). The second step of CVTree is standard and bears the full set of advantages and shortcomings of NJ algorithm. Despite limitations of either standard phylogenomic or CVTree analysis when used independently, the early divergence of Ustilaginomycotina seems to be a converged conclusion of both methods. In conclusion, this work has first recovered with high statistical confidence the early divergence of Ustilaginomycotina, and beyond that, this relationship have gained strong supports from multiple lines of evidence.

Acknowledgements

The authors would like to thank Dr. Guanghong Zuo for valuable suggestions. The authors dedicate this article to the late Professor Bai-Lin Hao of Fudan University. This work was supported by the National Basic Research Program of China (973 Project; Grant No. 2013CB834100) and the National Natural Science Foundation of China (Grant No. 11474068). Authors thanks the support of the State Key Laboratory of Applied Surface Physics and the Department of Physics, Fudan University, China.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2019.12.001.

References

[1] Hibbett DS, Binder M. Evolution of marine mushrooms. Biol Bull 2001;201:319–22.
[2] Matheny PB, Wang Z, Binder M, Curtis JM, Lim YW, Nilsson RH, Hughes KW, Hofstetter V, Ammirati JF, Schoch CL, Langer E, Langer G, McLaughlin DJ, Wilson AW, Froslev T, Ge ZW, Kerrigan RW, Slot JC, Yang ZL, Baroni TJ, Fischer M, Hosaka K, Matsuura K, Seidl MT, Vauras J, Hibbett DS. Contributions of rpb2 and tef1 to the phylogeny of mushrooms and allies (Basidiomycota, Fungi). Mol Phylogenetics Evol 2007;43:430–51.
[3] Bruns TD, Vilgalys R, Barns SM, Gonzalez D, Hibbett DS, Lane DJ, Simon L, Stickel S, Szaro TM, Weisburg WG, et al. Evolutionary relationships within the fungi: analyses of nuclear small subunit rRNA sequences. Mol Phylogenetics Evol 1992;1:231–41.
[4] Lutzoni F, Kauff F, Cox CJ, McLaughlin D, Celio G, Dentinger B, Padamsee M, Hibbett D, James TY, Baloch E, Grube M, Reeb V, Hofstetter V, Schoch C, Arnold AE, Miadlikowska J, Spatafora J, Johnson D, Hambleton S, Crockett M, Shoemaker R, Sung GH, Lucking R, Lumbsch T, O'Donnell K, Binder M, Diederich P, Ertz D, Gueidan C, Hansen K, Harris RC, Hosaka K, Lim YW, Matheny B, Nishida H, Pfister D, Rogers J, Rossman A, Schmitt I, Sipman H, Stone J, Sugiyama J, Yahr R, Vilgalys R. Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits. Am J Bot 2004;91:1446–80.
[5] Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, Eriksson OE, Huhndorf S, James T, Kirk PM, Lucking R, Thorsten Lumbsch H, Lutzoni F, Matheny PB, McLaughlin DJ, Powell MJ, Redhead S, Schoch CL, Spatafora JW, Stalpers JA, Vilgalys R, Aime MC, Aptroot A, Bauer R, Begerow D, Benny GL, Castlebury LA, Crous PW, Dai YC, Gams W, Geiser DM, Griffith GW, Gueidan C, Hawksworth DL, Hestmark G, Hosaka K, Humber RA, Hyde KD, Ironside JE, Koljalg U, Kurtzman CP, Larsson KH, Lichtwardt R, Longcore J, Miadlikowska J, Miller A, Moncalvo JM, Mozley-Standridge S, Oberwinkler F, Parmasto E, Reeb V, Rogers JD, Roux C, Ryvarden L, Sampaio JP, Schussler A, Sugiyama J, Thorn RG, Tibell L, Untereiner WA, Walker C, Wang Z, Weir A, Weiss M, White MM, Winka K, Yao YJ, Zhang N. A higher-level phylogenetic classification of the Fungi. Mycol Res 2007;111:509–47.
[6] Swann EC, Taylor JW. Phylogenetic perspectives on Basidiomycete systematics - evidence from the 18s ribosomal-rna gene. Can J Bot 1995;73:S862–8.
[7] Bauer R, Begerow D, Sampaio JP, Weib M, Oberwinkler F. The simple-septate basidiomycetes: a synopsis. Mycol Prog 2006;5:41–66.
[8] Hibbett DS. A phylogenetic overview of the Agaricomycotina. Mycologia 2006;98:917–25.
[9] Aime MC, Matheny PB, Henk DA, Frieders EM, Nilsson RH, Piepenbring M, McLaughlin DJ, Szabo LJ, Begerow D, Sampaio JP, Bauer R, Weiss M, Oberwinkler F, Hibbett D. An overview of the higher level classification of Pucciniomycotina based on combined analyses of nuclear large and small subunit rDNA sequences. Mycologia 2006;98:896–905.
[10] Begerow D, Stoll M, Bauer R. A phylogenetic hypothesis of Ustilaginomycotina based on multiple gene analyses and morphological data. Mycologia 2006;98:906–16.
[11] James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, Lumbsch HT, Rauhut A, Reeb V, Arnold AE, Amtoft A, Stajich JE, Hosaka K, Sung GH, Johnson D, O'Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schussler A, Longcore JE, O'Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies DR, Humber RA, Morton JB, Sugiyama J, Rossman AY, Rogers

JD, Pfister DH, Hewitt D, Hansen K, Hambleton S, Shoemaker RA, Kohlmeyer J, Volkmann-Kohlmeyer B, Spotts RA, Serdani M, Crous PW, Hughes KW, Matsuura K, Langer E, Langer G, Untereiner WA, Lucking R, Budel B, Geiser DM, Aptroot A, Diederich P, Schmitt I, Schultz M, Yahr R, Hibbett DS, Lutzoni F, McLaughlin DJ, Spatafora JW, Vilgalys R. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature 2006;443:818–22.

[12] Zhao RL, Li G, Sanchez-Ramirez S, Stata M, Yang Z, Wu G, Dai Y, He S, Cui B, Zhou J, Wu F, He M, Moncalvo J, Hyde KD. A six-gene phylogenetic overview of Basidiomycota and allied phyla with estimated divergence times of higher taxa and a phyloproteomics perspective. Fungal Divers 2017;84:43–74.

[13] McLaughlin DJ, Frieders EM, Lu HS. A microscopist's view of heterobasidiomycete phylogeny. Stud Mycol 1995;38:91–109.

[14] Leonard G, Richards TA. Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. P Natl Acad Sci USA 2012;109:21402–7.

[15] Robbertse B, Reeves JB, Schoch CL, Spatafora JW. A phylogenomic analysis of the Ascomycota. Fungal Genet Biol 2006;43:715–25.

[16] Fitzpatrick DA, Logue ME, Stajich JE, Butler G. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol 2006;6:99.

[17] Ebersberger I, Simoes RD, Kupczok A, Gube M, Kothe E, Voigt K, von Haeseler A. A consistent phylogenetic backbone for the fungi. Mol Biol Evol 2012;29:1319–34.

[18] Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 2013;497:327–31.

[19] Salichos L, Stamatakis A, Rokas A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol Biol Evol 2014;31:1261–71.

[20] Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Worheide G, Baurain D. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 2011;9:e1000602.

[21] Cheng J, Cao F, Liu Z. AGP: a multimethods web server for alignment-free genome phylogeny. Mol Biol Evol 2013;30:1032–7.

[22] Qi J, Wang B, Hao BI. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J Mol Evol 2004;58:1–11.

[23] Gao L, Qi J. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol Biol 2007;7:41.

[24] Hao BL, Gao L. Prokaryotic branch of the Tree of Life: a composition vector approach. J Syst Evol 2008;46:258–62.

[25] Wang H, Xu Z, Gao L, Hao BL. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evol Biol 2009;9:195.

[26] Yuan J, Zhu Q, Liu B. Phylogenetic and biological significance of evolutionary elements from metazoan mitochondrial genomes. PLoS One 2014;9:e84330.

[27] Li L, Stoeckert Jr. CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 2003;13:2178–89.

[28] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–7.

[29] Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 2009;25:1972–3.

[30] Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 2005;21:2104–5.

[31] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–3.

[32] Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 2001;17:754–5.

[33] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 2013;30:2725–9.

[34] Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 2001;17:1246–7.

[35] Struck TH. TreSpEx-detection of misleading signal in phylogenetic reconstructions based on tree information. Evol. Bioinform. Online 2014;10:51–67.

[36] Xu Z, Hao B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res 2009;37:W174–8.

[37] Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005;21:3674–6.

[38] Li Q, Xu Z, Hao B. Composition vector approach to whole-genome-based prokaryotic phylogeny: success and foundations. J Biotechnol 2010;149:115–9.

[39] Muszewska A, Steczkiewicz K, Ginalski K. DIRS and Ngaro retrotransposons in fungi. PLoS One 2013;8:e76319.