Marine Genomics 36 (2017) 33–39

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Genomics/technical resources

A phylogenetic framework for the kingdom Fungi based on 18S rRNA gene sequences

Pablo Yarza [a,*], Pelin Yilmaz [b], Katrin Panzer [c,d], Frank Oliver Glöckner [b,d], Marlis Reich [c,*]

[a] Ribocon GmbH, Fahrenheitstrasse 1, 28359 Bremen, Germany
[b] Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany
[c] University of Bremen, Molecular Ecology, BreMarE, Leobener Str. 5, 28359 Bremen, Germany
[d] Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany

ARTICLE INFO

Article history:
Received 14 March 2017
Received in revised form 28 April 2017
Accepted 15 May 2017
Available online 1 June 2017

Keywords:
SILVA
SSU
Reference phylogenetic tree
Alignment
Community analysis
Fungi

ABSTRACT

The usage of molecular phylogenetic approaches is critical to advance the understanding of systematics and community processes in the kingdom Fungi. Among the possible phylogenetic markers (or combinations of them), the 18S rRNA gene appears currently as the most prominent candidate due to its large availability in public databases and informative content. The purpose of this work was the creation of a reference phylogenetic framework that can serve as ready-to-use package for its application on fungal classification and community analysis. The current database contains 9329 representative 18S rRNA gene sequences covering the whole fungal kingdom, a manually curated alignment, an annotated and revised phylogenetic tree with all the sequence entries, updated information on current taxonomy, and recommendations of use. Out of 201 total fungal taxa with more than two sequences in the dataset, 179 were monophyletic. From another perspective, 66% of the entries had a tree-derived classification identical to that obtained from the NCBI taxonomy, whereas 34% differed in one or the other rank. Most of the differences were associated to missing taxonomic assignments in NCBI taxonomy, or the unexpected position of sequences that positioned out of their theoretically corresponding clades. The strong correlation observed with current fungal taxonomy evidences that 18S rRNA gene sequence-based phylogenies are adequate to reflect genealogy of Fungi at the levels of order and above, and justify their further usage and exploration.

1. Introduction

Fungi are a highly diverse group of eukaryotic organisms that inhabit a wide array of habitats. Current estimates of the species richness in the fungal kingdom exceed the number of 1.5 Mio (Hawksworth, 2012). However, only 100,000 fungal species have been so far described (Stajich et al., 2009) and their classification is a difficult task due to their large heterogeneity. Morphological identification of fungi, for example, is often hampered by the scarcity of discriminatory taxonomic characters. An additional concern is the morphological transition of many fungal species as a response to changing environmental conditions (Rayner and Coates, 1987; Slepecky and Starmer, 2009) or as part of their life-cycle. Thus, many fungi have only been described based on their asexual morph or the connection to the related sexual morph was not perceived. All these points supported the proliferation of polyphyletic genera and groups.

The incorporation of molecular phylogenetic data enabled robust genealogical inferences that revolutionized fungal classification. Sequence-based phylogenetic studies resolved, thus, the relation of anamorphs-teleomorphs, synonyms or wrongly identified species (Kurtzman et al., 2011; Wijayawardene et al., 2014; Reblova et al., 2016). It further shed light on fungal diversity by detection of new taxa in the tree topology such as the *Cryptomycota* (Lara et al., 2010; Jones, 2011) or the *Archaeorhizomycetes* (Porter et al., 2008; Rosling et al., 2011). For community ecology, phylogeny-based approaches also gain in popularity as they help to understand the processes that generate variation in diversity, identity and co-occurrence such as trait evolution (James et al., 2006; Powell et al., 2009; Reich et al., in press), biogeography (Wu et al., 2000; Teeling et al., 2005; Ghikas et al., 2010) or non-random community processes (Enquist et al., 2002; Horner-Devine and Bohannan, 2006; Forest et al., 2007; Panzer et al., 2015b).

Multi-locus approaches or the use of entire genomes are nowadays the standard for inferring deep evolutionary processes in Fungi as they maximize phylogenetic information compared to the use of a single marker gene. An obstacle is here the limited availability of data: the concatenated alignment of the AFTOL project (Assembling the Fungal Tree of Life; http://www.aftol.org/) includes data from only 214 taxa. Phylogenetic genome comparisons have been done so far only for up

* Corresponding authors.
E-mail addresses: pyarza@ribocon.com (P. Yarza), reich@uni-bremen.de (M. Reich).

http://dx.doi.org/10.1016/j.margen.2017.05.009

to 82 fungal species (Wang et al., 2009). Furthermore, the datasets are biased towards taxa of the Dikarya. For the classification of single taxa from environmental samples or for ecological community studies that are mainly based on single marker genes and are rich in undescribed fungal species, these methods are not applicable.

The most prominent fungal marker genes are the Internal Transcribed Spacers (ITS) and the 28S and 18S rRNA gene sequences (Feau et al., 2011). They have all their pros and cons (for detailed discussion, see Reich and Labes (in press)): while the ITS is the best candidate for species barcoding, it lacks the phylogenetic power (Schoch et al., 2012), why classification of environmental sequences with ITS is often limited to kingdom or phylum level (Nilsson et al., 2016). In contrast, the 18S rRNA gene sequence rarely resolves fungal taxa to species or genus level but is an important phylogenetic marker for a reliable classification of undescribed fungal taxa. Additionally, it is represented by the highest sequence number in public databases (Quast et al., 2013; Panzer et al., 2015b) covering all major fungal groups compared to the 28S rRNA gene sequence. Unfortunately, only few phylogenetic datasets are publicly available such as the group-specific alignments of the AFTOL projects (http://www.aftol.org/data.php) and of the PHYMYCO database (Mahe et al., 2012) or group specific phylogenetic reference trees (Hinchliff et al., 2015; Panzer et al., 2015a). Until now, a 18S rRNA gene sequence phylogenetic reference dataset that covers all major groups of Fungi is missing.

The present study wishes to facilitate current and future mycological research by providing a fungal specific 18S rRNA gene sequence phylogenetic tree based on a quality controlled selection of sequence entries, supplemented with a curated taxonomy, as well as intron-free and manually improved alignments. The alternative taxonomic classification extracted from the topology of the calculated consensus tree was compared to the currently accepted classification, to elucidate whether the 18S rRNA gene sequence phylogenetic analysis represents a suitable approach for the advance of fungal taxonomy.

2. Materials and methods

2.1. Sequence dataset and alignment

Initially, 71,787 high-quality and aligned eukaryotic 18S rRNA gene sequences were downloaded in ARB format (Ludwig et al., 2004) from the SILVA SSU Ref database (Quast et al., 2013) release 111 (https://www.arb-silva.de/documentation/release-111/). A stringent quality control was applied as follows: only fungal sequences according to the SILVA taxonomy with a taxonomic annotation to genus level were retained. The alignment was manually curated to further improve the positional orthology, as an important prerequisite for phylogenetic reconstruction (Ludwig and Klenk, 2001; Peplies et al., 2008). This tedious task was facilitated by the ARB editor which takes into account the secondary structure of the molecule (helix and loop information), with the help of a positional variability filter previously calculated for the whole eukaryotic domain (i.e. eukaryotic SAI (sequence associated information) filter). Re-allocation of misplaced bases solved the most problematic areas. In addition, manually-identified introns were cut out from >400 sequence entries. For example, the sequence entry with accession no. AB479213, with a sequence length of 2124 bp, contained an intron of 441 bp starting at sequence position 1401. Sequences with gaps longer than 20 bp or with mismatching stretches longer than 20 bp were detected and rejected from the study. Finally, in a neighbour joining tree reconstruction 266 sequences were detected to form very long branches, or with high probability for being contaminations or having unreliable annotations, and were consequently removed from the study. The final working dataset had 9329 manually-aligned 18S rRNA gene sequences of Fungi, of high quality in terms of sequence length (average = 1708) and minor amount of ambiguities (average = 0.4).

2.2. Phylogenetic analysis

2.2.1. Software

The phylogenetic reconstruction was performed in ARB, version 6.0.2. The maximum likelihood inferences run on standalone RAxML 8.0.20 (Stamatakis, 2014), and the trees were afterwards imported into ARB.

2.2.2. Core and extra datasets

The 9329 sequences were used to create a preliminary neighbour joining reconstruction, followed by a de-replication using the ARB OTU (Operational Taxonomic Unit) tool. This functionality detects homogenous groups within an existing tree topology and suggests a sequence representative within each clade (i.e. the sequence with least mean distance to the other group members). More detailed information is available at the ARB's official documentation at http://help.arb-home.de/di_clusters.html. We configured the tool to detect all possible subtrees of at least two members and with a maximum per-cluster distance of 4%. As a result, 1098 sequences entries were retained. After several preliminary tests with draft tree reconstructions, we defined a robust outgroup of 37 high quality sequence entries belonging to protists of the lineages *Cryptophyta* and *Amoebozoa*. The accession numbers of these sequence entries are: AB240962, AB240960, AF019062, AF114438, AF508267, AF508269, AM051195, AM051200, AM901344, AM901358, AM901370, AY642712, AY642724, AY642713, AY665097, AY919695, AY919710, AY919794, DQ310197, EF195737, EF195738, GQ375264, GQ397474, GQ397473, GQ905499, GU808311, GU808317, HM135080, HM363628, HQ191419, HQ219344, JF730765, JQ223010, JQ226537, S81337, U07412, X57162. Including both the representative selection and the outgroup, these 1135 sequence entries constituted the 'core dataset'. The remaining 8194 fungal sequences were considered as the 'extra dataset'.

2.2.3. Filters

A 50% maximum frequency filter was created for the 1135 sequences in the core dataset. As a result, the full SILVA's SSU alignment of 50,000 positions was reduced to 1696 positions. The filtering of highly variable positions is a standard procedure to remove singularities that in many cases correspond to artifacts and not to real evolutionary events, for example due to sequencing errors. This noise removal further ensures the positional orthology and optimizes the phylogenetic inference. By applying the filter, it is assumed that some of the true hypervariability will be filtered out. However, this usually affects the lowest taxonomic levels, and seems to be negligible in our case due to the non-redundant nature of the core dataset.

2.2.4. Preliminary trees

The 50%-filtered core dataset (1135 sequences) was submitted to three distinct treeing calculations: (1) A neighbour joining (NJ) tree was created in ARB using the Jukes Cantor correction, (2) a maximum parsimony tree (MP) was built in ARB using the previous NJ topology as a template followed by global parsimony optimization, and (3) a maximum likelihood tree (ML) was created with the external software RAxML using the GTR model with GAMMA correction.

2.2.5. Final consensus tree

A consensus tree topology was created using the three preliminary topologies (NJ, MP and ML) as source. The creation of the consensus tree topology is an iteratively process of selecting the "best" branch found within a pool of branches (derived from source trees) and adding it to the final tree according to the criteria given in http://help.arb-home.de/consense_algo.html. The remaining 8341 sequences were then inserted one at a time by phylogenetic placement (Fig. 1), using ARB Parsimony, which infers the most parsimonious place in the tree where this sequence could fit in. Each insertion was done with a filtering option that considered for calculation only the positions existing on that

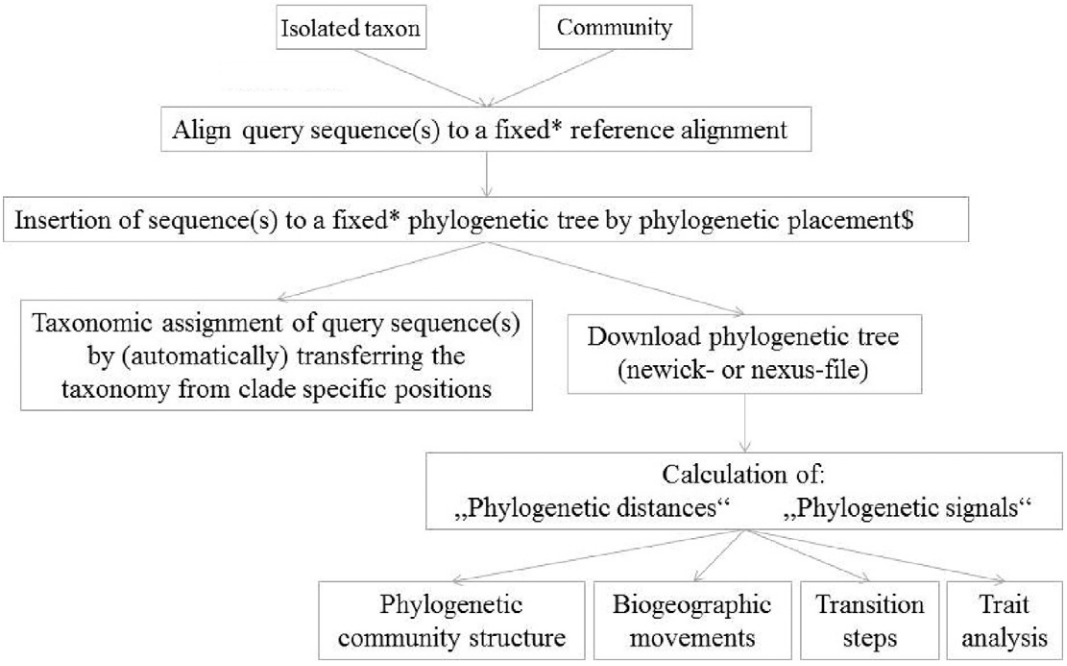

Fig. 1. Scheme demonstrating the necessary work steps to integrate the phylogenetic reference tree into the overall workflow. Different applications are possible including a simple classification of sequences or as basis for phylogeny-driven ecological analysis. *sequences are aligned/inserted without changing alignment structure/topology of the tree; $phylogenetic placement accommodates the evolutionary history of genes on a gene-by-gene basis.

particular sequence, and also excluded bases beyond the 18S rRNA gene sequence boundary.

2.2.6. Clade recognition

211 fungal high taxa, distributed in 150 orders, 40 classes, 12 subphyla and 9 phyla according to the NCBI (National Center for Biotechnology Information) taxonomy (Federhen, 2012), were checked for monophyly (i.e. clades) in the new 18S rRNA gene sequence-based phylogenetic reconstruction. Clades were recognized when at least two thirds of the members of a taxon were present. This proportion, somewhat similar to thresholds of related work (Yarza et al., 2014; Yilmaz et al., 2015), was used to identify a representative core set of sequence entries. In such cases a group in the tree was created, and named accordingly, to position each taxon. The members of a taxon that fell apart from the expected monophyletic core set were considered as outliers. These were subsequently regarded as intruders if they affiliated within another existing clade, or simply outliers without a clear taxonomic affiliation when falling in an outer position (e.g. basal lonely branch). Outliers and intruders may occur due to poor local tree resolution, but often they are indicating true issues related to bad sequence annotations (e.g. contaminated samples), or even taxonomic misclassifications.

2.3. Taxonomic inspection

Each of the 9329 sequence entries in the working dataset, excluding the outgroup, were assigned to an order, class, subphylum and phylum according to two distinct taxonomy sources: the NCBI taxonomy, as a provider of the accepted taxonomy of Fungi, and the alternative classification derived from the 18S rRNA-based phylogenetic tree. The NCBI taxonomy was acquired during the months of December 2016 and January 2017. To allow more detailed comparisons, those sequence entries with missing assignments in NCBI taxonomy in one or several ranks were completed with status ‘unassigned’. For example, *Lecophagus longisporus* (acc no. AB014400) with no order, no class and no subphylum assignments in NCBI was classified as phylum *Ascomycota*, subphylum unassigned, class unassigned and order unassigned. Concerning the tree-based taxonomy, the group name was copied into a database field within the ARB environment for all the 9329 fungal sequences. Hereby, clades within clades conformed this taxonomic classification. Areas of the tree with less recognized clades originated sequence entries with a missing assignment into one or more ranks. In those cases the ‘unassigned’ status was used as well. For example, sequence acc. no. AY123321 showed the following classification according to the tree: phylum *Basidiomycota*, subphylum *Pucciniomycotina*, class *Microbotryomycetes*, order unassigned. Both the NCBI taxonomy and the tree-based taxonomy were thoroughly compared for all sequences.

2.4. Data deposition

A public repository, https://www.arb-silva.de/no_cache/download/archive/publications/fungi_18S/, has been created to give access to the following materials: (i) the whole alignment of the 9329 fungal sequence entries in multi-fasta format, (ii) the reference phylogenetic tree in newick format, and (iii) the ARB database which contains the curated fungal dataset, with the tree, filter and new database fields.

3. Results

3.1. A toolkit for fungal systematics

The principal result of this work is a collection of curated materials that constitute a specialized toolkit for fungal systematics based on 18S rRNA gene sequences (Table 1). The two most relevant materials comprise a manually supervised alignment of 9329 aligned sequence entries of good quality, covering a significant portion of the fungal diversity, and a single phylogenetic tree where all sequences are contained. The phylogenetic tree includes clade information to facilitate its topology exploration. A revised taxonomy is available that contains the taxonomy derived from the clades in the tree (i.e. their names and hierarchical classification). This taxonomy path has a fixed number of levels, four in this case (order, class, subphylum and phylum), to allow precise comparisons between different taxonomic classifications. In those cases where the tree could not provide assignment into a certain rank, i.e. by the lack of the corresponding clade, the term unassigned was used (see Materials and methods). All materials can be accessed

Table 1
Description of files generated in this work. Files available at the project's repository (https://www.arb-silva.de/no_cache/download/archive/publications/fungi_18S/).

Aligned_FungiRef111.fasta	Aligned dataset in multi fasta format, taxonomy included in header lines. The taxonomy is based on the tree topology.
FungiRef111.newick	Whole tree (9366 sequences, includes outgroup) in newick format. Includes tree-based taxonomic information.
Core_trees.zip	Collection of preliminary trees and core consensus tree created with the representative set of 1135 sequence entries. Files included: *datasetCONSENSUS_core_tree.newick, ML_core_tree.newick, MP_core_tree.newick* and *NJ_core_tree.newick*.
table_sequence.xls	Sequence list including Acc No. for easy mapping, and comparison of NCBI taxonomy as of January 2017 to tree taxonomy as deduced from the tree clades.
table_taxon.xls	List of recognized taxonomic lineages in the tree. Report of monophyly, matched entries, clade members, clade intruders and outliers.
FungiRef111.arb	ARB file including alignment, reference tree and taxonomy.
fasta_fungiref111.eft	ARB export filter for fasta format. This is the export filter used to export the sequence dataset from the ARB database to a multi fasta format output file.
Taxonomic_assessment_in_ARB.pdf	A brief "how-to"-manual for phylogenetic placement and automatic taxonomic classification of inserted sequences.

independently to facilitate the user's particular bioinformatics set up (Table 1). The tree-derived taxonomy is an available feature not only via the ARB database (as new taxonomy field "tax_fungiref"), but directly through the newick and table files as well (Table 1).

For ARB-software users, these materials are all self-contained within a single database file that is a curated and modified version derived from the official SILVA SSU Ref 111 release. The 50% variability filter '18S_fungi' used for the tree calculation is included as well as a SAI of the ARB database. For 355 sequence entries of this database introns were detected and removed, and therefore the primary data changed with respect to the original submission. To avoid misunderstandings, the accession number database field (acc) has been edited by appending the word "_modified" (e.g. AB233335_modified), and the database field (remarks) reads "had_introns" for all of them.

3.2. Tree highlights

The three independent phylogenetic reconstructions made with the core representative dataset and distinct algorithms (NJ, MP, ML - see Materials and methods) yielded topologies of overall acceptable similarity. To reflect this with a value, 713 nodes (63%) of the 1133 total nodes in the core consensus tree (see Materials and methods) were shared by at least two of the three topologies, and 420 nodes (37%) were distinctly resolved by all the three topologies. These preliminary trees and the consensus core tree have been supplied in the project repository for transparency. On the light of the surprisingly high correlation between the final consensus phylogeny (i.e. all extra sequences added) and the fungal taxonomy (see below), we decided to not show a strict consensus topology, which means, adding multifurcations in every place where at least one tree had differed. Such a conservative topology would compromise too much the agreement between taxonomy and phylogeny (data not shown), which is needed for the purpose of our tree: the use for fungal classification and phylogeny-based ecological studies.

The phylogenetic tree had a strong correlation with the current fungal classification. From 211 initial taxa listed, 10 were represented by a single sequence and 201 had two or more sequences in the dataset (Repository, table_taxon.xls). Among the latter, 22 showed poly- or paraphyletic patterns and 179 were recognized as monophyletic clades (i.e. with at least two thirds of all members included, see Materials and methods) (Repository, table_taxon.xls). These coherent clades represented 8 out of 9 phyla (89%), 11 out of 12 subphyla (92%), 37 out of 40 classes (92%), and 123 out of 150 orders (82%) (Repository, table_taxon.xls). Despite their internal taxonomic coherence, many of the subphylum and phylum-level clades appeared interconnected by short branches. This is an indication of low resolution at this most ancient tree zone. For example, although there is a tendency which groups *Basidiomycota*, *Ascomycota* and *Entorrhizomycota* together, and separates them from e.g. *Chytridiomycota*, the short branches interconnecting them indicate that this particular branching pattern has indeed low support. The complete phylogenetic tree in newick format can be accessed over the repository (Repository, FungiRef111.newick).

3.3. Taxonomic considerations

A detailed comparison of the NCBI taxonomy and the tree-derived taxonomy revealed that among the 9329 fungal sequence entries (100%), in 6168 cases (66%) the two taxonomies were fully consistent, while the other 3161 ones (34%) had distinct taxonomic paths (Table 2). For a full report see the table "table_sequence.xls" in the repository. The discrepancies occurred due to one or more of the following possible reasons:

(1) "*Dead*" *taxa*. Most of the cases (89% of the total number of distinct taxa) occurred when the tree rendered unassigned status where the NCBI taxonomy did have an assignation. We used the term "dead" to refer to such a loss on the new taxonomy (Table 2). Most often, the dead taxa occurred when the original taxa (i.e. as given by NCBI taxonomy) were not recognized as clades in the tree. For example, the 672 sequences belonging to phylum *Mucoromycota* did not form a monophyletic clade and were then regarded as unassigned phylum. Another possibility is when the sequences, now behaving as outliers (i.e. out from its theoretically expected affiliation – see Materials and methods), got positioned in a "lonely" area with no recognized clades, for example *Physoderma maculare* (acc no. DQ536489), out of the phylum-level clade *Blastocladiomycota*.

(2) "*Born*" *taxa*. Another source of discrepancy (8% of the total number of distinct taxa) appeared when sequences with incomplete taxonomic classification according to NCBI (i.e. unassigned status), in at least one of the four high ranks considered, were affiliated in a recognized tree clade and thereby obtained the corresponding taxonomic classification. We refer to them as "born" taxa, as they indicate a gain in taxonomic information (Table 2). In total, 909 sequence entries accounted for one or more unassigned levels. 325 sequences had no order level assignment, 561 sequences without class level assignment, 259 sequences with no subphylum level assignment, and 1 sequence had missing phylum level assignment. In total there were 1146 unassigned taxa according to NCBI taxonomy for the working sequence dataset. Among them, the tree resolved 272 ones ("born" status) spanning 149 total sequence entries. For example, *Mastodia tesellata* (acc no. FN668947) without order-level assignment but unambiguously positioned within the order *Verrucariales* in our tree.

(3) "*Renamed*" *taxa*. And finally, some differences (3% of the total number of distinct taxa) occurred when the sequence fell in a different clade (of the same rank) than the specified by NCBI taxonomy, and thus the sequence acquired the new clade name. These sequences are also recognized as outliers (see above) and we called them "renamed" taxa (Table 2). For example, *Helicoon richonis* (acc no. AY856952) affiliated in the order *Pleosporales* instead of the expected assignment within *Orbiliales*.

Table 2
The reference NCBI taxonomy was compared against the alternative taxonomy derived from the phylogenetic tree. A general comparison (Full path) registered the differences using the whole taxonomic data at once. Here, two states were considered: "equal" or "distinct". In addition, the two classifications were compared rank by rank (phylum, subphylum, class and order) for a detailed report. In this case, four states were considered: "equal" (i.e. no change in the rank assignation), "renamed" (i.e. the two taxonomies disagree in terms of the taxonomic assignation), "born" (i.e. an assignation is provided where NCBI had unassigned status), and "dead" (i.e. the tree taxonomy rendered unassigned status where NCBI had a valid assignation). Total paths indicate the number of individual taxonomic assignations observed (i.e. each sequence records four assignments, one per taxonomic rank).

Status		Full path	Phylum	Subphylum	Class	Order	Total paths
Total		9329	9329	9329	9329	9329	37,316
Equal		6168 (66%)	8588	9117	7940	8238	33,883
Distinct		3161 (34%)	741	212	1389	1091	3433 (100%)
	Dead	–	741	104	1304	924	3073 (89%)
	Born	–	0	108	74	90	272 (8%)
	Renamed	–	0	0	11	77	88 (3%)

4. Discussion

4.1. Lights and shadows of a reference phylogenetic tree for the kingdom Fungi

In this study we have been cautious on creating a non-redundant dataset of sequences to create a robust core phylogenetic reconstruction, and this has been proven to be a beneficial approach for a global phylogenetic reconstruction of Fungi. However, a minority of taxa might have been pruned in excess with bad consequences. This is a possible explanation for the unexpected position of class *Basidiobolomycetes* out of the subphylum *Entomophthoromycotina*. Only seven sequences were available for the whole class *Basidiobolomycetes*, six of the genus *Basidiobolus* and one intruder of the genus *Conidiobolus*, and they were all >99% similar to each other. Accordingly, during the data reduction process only one sequence was selected to represent the whole class *Basidiobolomycetes*. Furthermore, this class has been shown to be a basal lineage in *Entomophthoromycotina* (Gryganskyi et al., 2013). Therefore, the most plausible explanation for this unexpected affiliation is that this single sequence had not enough weight to reveal the correct position of the ancient lineage *Basidiobolomycetes*.

Another limitation of the current tree topology is that the branching pattern at the highest ranks appeared somewhat fuzzy, to judge by the low support in terms of short branches of the distinct phyla. The poor resolution at this most ancient part is a complex issue, and is a common phenomenon in comprehensive SSU-based trees made with large sequence datasets (Ludwig et al., 2011). The so called bush-shaped topologies often appear where the dataset cannot be clustered with enough significance by the methodology employed.

Thus, having in mind that phylogenetic trees are dynamic entities that change according to the quality and availability of the data (Ludwig and Klenk, 2001), we do not obstinate on finding the greatest resolution possible to resolve the genealogy of Fungi unambiguously with a single marker gene. Our efforts have rather been concentrated on the creation of a curated reference dataset and a useful tool. We have created a phylogenetic tree based on a selection of 9329 fungal 18S rRNA gene sequences of high quality, following state-of-the-art data curation and treeing methodologies. The topology of our tree provided excellent resolution at the four and highest taxonomic ranks explored: >80% of all the lineages considered appeared supported by the tree as monophyletic clades. This strong correlation with current fungal taxonomy evidences that 18S rRNA gene sequence-based phylogenies are adequate to reflect genealogy of fungi at the levels of order and above.

Historically, the criteria to characterize and classify taxa are subjected to the technical limitations. These, including sequence-data availability, hamper recognizing their phenotypic and genotypic features, as well as their phylogeny, unambiguously. This inherent subjectivity (or lack of objective criteria), accumulated through many years of fungal systematics, is well exemplified in areas of the tree topology where the taxonomy showed poor stratification. This is yet a feeling that the branches in a local tree topology could sometimes be better understood by means of more clades, for example, more orders within a class, or more classes within a phylum. We found this problem in groups such as the class *Entomophthoromycetes* that is currently composed of one single order *Entomophthorales*, represented by 56 sequences in the dataset. The internal sequence identity values registered were about 80%. This heterogeneity, together with the remarkable paraphyletic pattern of the genus *Conidiobolus* (Gryganskyi et al., 2013) suggest that the current *Entomophthoromycetes* could be split into several orders.

4.2. Applications of the phylogenetic tree

A phylogenetic tree serves as a placeholder for the different taxa that can be recognized as clades. In this regard, the placement of a sequence into a preexisting topology can directly reveal its taxonomic affiliation. The significance of this assignation is limited by at least four main factors: (1) Full length and reliable sequences with manually curated alignments display optimal conditions, whereas short sequences and/or with bad alignments can add extra distance which distorts their phylogenetic position. (2) Only optimized tree topologies that present a reliable phylogeny should be used as references for taxonomic assessment. (3) Well annotated trees are necessary as in origin all phylogenetic trees come as raw topologies where the clades' limits have to be defined. Therefore, the tree becomes useful for further taxonomic assignations only when the distinct taxa have been recognized as coherent clades, and annotated following a standardized nomenclature. And then, the more annotated clades exist in a tree, the richer classification schema is available for a better taxonomic assessment, by clade membership. (4) Consistent and up-to-date taxonomic annotation of the reference sequences and the tree clades is essential, due to the rapid changes in taxonomy. The phylogenetic tree presented here has been created taking all these premises into account, and is closer to what we believe is a reference tree for Fungi.

The applications of this tree as a tool are mainly two: taxonomic assignation and phylogeny-based diversity assessments (Fig. 1). For both cases, the raw query sequences can be inserted into the reference tree using a phylogenetic placement method: The first step would be the alignment of the query sequences according to the profile followed by their insertion into the pre-existing tree topology. The insertion step should not alter the underlying tree topology to prevent potential negative effects caused when including partial or low quality sequences. Several programs exist that allow phylogenetic placement for example ARB (Ludwig et al., 2004), RAxML (Berger et al., 2011; Stamatakis, 2014) and pplacer (Matsen et al., 2010). They differ in the phylogenetic criterion used to place the query sequences. All developed programs are nowadays suitable for a fast and accurate insertion of thousands of query sequences into the reference tree.

4.2.1. Taxonomic assignment

For the classification of fungal taxa, very often a sequence similarity-based approach is taken. In such cases, the classification of the query sequence is based on a reference dataset of known origin (e.g. BLAST

(Altschul et al., 1990)). However, if the query is only distantly related to the taxa represented by the reference dataset, the query might be erroneously forced into an alignment with a known sequence or return an uninformatively broad collection of best matching hits. Application of the lowest common ancestor rule will result for the latter case into the classification on a very high taxonomic level like phylum or kingdom (Reich et al., in press). In contrast, the presence of a query sequence on a certain branch of an annotated phylogenetic tree gives precise information about the evolutionary relationship of that sequence to other sequences in the tree. We provide here curated taxonomy to the order level and above as this level is supported for all fungal groups present in our tree. However, for some fungal groups, classification beyond the order level is possible but needs manually inspection of the tree and placement of the newly inserted sequence.

4.2.2. Phylogeny-based diversity assessment

One obstacle for high throughput sequencing based studies is the integration of all generated sequences into the reference tree. Often, the generated sequences were clustered into OTUs using the sequences of a reference tree as seed. However, this approach introduces a bias by the nonuniform rates of sequence divergence across clades and inaccurate a priori clustering thresholds. In addition, they should only be used when all taxa of the community are represented in the tree. Instead, phylogeny-based distance measures are a more robust approach as they exploit the degree of divergence between different sequences: a query sequence placed deep in the tree can indicate how the query is distantly related to other sequences, whereas the corresponding taxonomic name would simply indicate membership in a large taxonomic group. This becomes of special importance when working with environmental community samples composed by a large number of undescribed or cryptic taxa. Branch lengths are used for example to calculate phylogenetic signals for trait conservatism analysis (Webb et al., 2008), diverse distance matrices for exploring the community structure and diversity (Lozupone and Knight, 2005; Hamady et al., 2010), to measure transition steps of organisms (Alverson et al., 2007) or for study biogeographical movements (Wu et al., 2000; Teeling et al., 2005) (Fig. 1).

5. Conclusions

Fungal taxonomy is a very active field where nomenclature, classification and characterization of taxa can rapidly change according to the availability of new technologies and information that may contribute to the discovery of their key features. However, the taxonomy-data transfer from the original scientific articles (i.e. primary taxonomy resources) to the sequence databases is an unpaved road: whole teams of expert curators are required to keep track of changes (Federhen, 2012; Yilmaz et al., 2014). Such a limitation motivates that errors on taxonomic information tend to accumulate fast. Then, since many sequence entries share the same classification (mainly on high ranks), the errors propagate with a snow-ball effect to the rest of subordinate repositories and end users.

To improve this situation, we used the strategy of creating a representative dataset of manageable size where data-curation could be intensively applied. The meticulous data curation and phylogenetic analysis performed has proven the validity of 18S rRNA-based approaches to infer the hierarchical classification of a fungal taxon into high ranks. This motivates us to advance in the understanding of tree topologies inspired by the Candidate Taxonomic Unit concept initially developed for Prokaryotes (Yarza et al., 2014).

Finally, the work accomplished here (alignment, dataset, tree, taxonomy) transcends its application as a fungal reference dataset and tool. First, the refined alignment has been returned back to the SILVA project for its inclusion into the SILVA seed. Second, the NCBI taxonomy and Mycobank teams have been informed about the 'alternative' taxonomy derived from the phylogenetic tree that has solved some of the gaps in the taxonomic classification. Third, the data has been additionally shared within the Open Tree of Life initiative (https://tree.opentreeoflife.org). In conclusion, the return of curated data back to higher repositories closes a circle of quality management, promoting a more efficient integration and propagation of these improvements to the whole community.

6. Acknowledgements

We thank Ingrid Ugussi Vukman for partly processing the taxonomic annotation and quality control of the sequences.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Alverson, A.J., Jansen, R.K., Theriot, E.C., 2007. Bridging the Rubicon: phylogenetic analysis reveals repeated colonizations of marine and fresh waters by thalassiosiroid diatoms. Mol. Phylogenet. Evol. 45, 193–210.
Berger, S.A., Krompass, D., Stamatakis, A., 2011. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. Syst. Biol. 60, 291–302.
Enquist, B.J., Haskell, J.P., Tiffney, B.H., 2002. General patterns of taxonomic and biomass partitioning in extant and fossil plant communities. Nature 419, 610–613.
Feau, N., Decourcelle, T., Husson, C., Desprez-Loustau, M.L., Dutech, C., 2011. Finding single copy genes out of sequenced genomes for multilocus phylogenetics in non-model Fungi. PLoS ONE 6.
Federhen, S., 2012. The NCBI taxonomy database. Nucleic Acids Res. 40, D136–D143.
Forest, F., Grenyer, R., Rouget, M., Davies, T.J., Cowling, R.M., Faith, D.P., Balmford, A., Manning, J.C., Proches, S., van der Bank, M., Reeves, G., Hedderson, T.A.J., Savolainen, V., 2007. Preserving the evolutionary potential of floras in biodiversity hotspots. Nature 445, 757–760.
Ghikas, D.V., Kouvelis, V.N., Typas, M.A., 2010. Phylogenetic and biogeographic implications inferred by mitochondrial intergenic region analyses and ITS1-5.8S-ITS2 of the entomopathogenic fungi *Beauveria bassiana* and *B. brongniartii*. BMC Microbiol. 10.
Gryganskyi, A.P., Humber, R.A., Smith, M.E., Hodge, K., Huang, B., Voigt, K., Vilgalys, R., 2013. Phylogenetic lineages in *Entomophthoromycota*. Persoonia 30, 94–105.
Hamady, M., Lozupone, C., Knight, R., 2010. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J. 4, 17–27.
Hawksworth, D.L., 2012. Global species numbers of fungi: are tropical studies and molecular approaches contributing to a more robust estimate? Biodivers. Conserv. 21, 2425–2433.
Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R., Gude, K., Hibbett, D.S., Katz, L.A., Laughinghouse, H.D., McTavish, E.J., Midford, P.E., Owen, C.L., Ree, R.H., Rees, J.A., Soltis, D.E., Williams, T., Cranston, K.A., 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc. Natl. Acad. Sci. U. S. A. 112, 12764–12769.
Horner-Devine, M.C., Bohannan, B.J.M., 2006. Phylogenetic clustering and overdispersion in bacterial communities. Ecology 87, S100–S108.
James, T.Y., Kauff, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J., Celio, G., Gueidan, C., Fraker, E., Miadlikowska, J., Lumbsch, H.T., Rauhut, A., Reeb, V., Arnold, A.E., Amtoft, A., Stajich, J.E., Hosaka, K., Sung, G.H., Johnson, D., O'Rourke, B., Crockett, M., Binder, M., Curtis, J.M., Slot, J.C., Wang, Z., Wilson, A.W., Schäler, A., Longcore, J.E., O'Donnell, K., Mozley-Standridge, S., Porter, D., Letcher, P.M., Powell, M.J., Taylor, J.W., White, M.M., Griffith, G.W., Davies, D.R., Humber, R.A., Morton, J.B., Sugiyama, J., Rossman, A.Y., Rogers, J.D., Pfister, D.H., Hewitt, D., Hansen, K., Hambleton, S., Shoemaker, R.A., Kohlmeyer, J., Volkmann-Kohlmeyer, B., Spotts, R.A., Serdani, M., Crous, P.W., Hughes, K.W., Matsuura, K., Langer, E., Langer, G., Untereiner, W.A., Lucking, R., Budel, B., Geiser, D.M., Aptroot, A., Diederich, P., Schmitt, I., Schultz, M., Yahr, R., Hibbett, D.S., Lutzoni, F., McLaughlin, D.J., Spatafora, J.W., Vilgalys, R., 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature 443, 818–822.
Jones, E.B.G., 2011. Are there more marine fungi to be described? Bot. Mar. 54, 343–354.
Kurtzman, C.P., Fell, J.W., Boekhout, T., 2011. The Yeasts, a Taxonomic Study. 3.
Lara, E., Moreira, D., Lopez-Garcia, P., 2010. The environmental clade LKM11 and Rozella form the deepest branching clade of fungi. Protist 161, 116–121.
Lozupone, C., Knight, R., 2005. UniFrac: a new phylogenetic method for comparing microbial communities. AEM 71, 8228–8235.
Ludwig, W., Klenk, H.-P., 2001. In: Boone, D.R., Castenholz, R.W., Garrity, G.M. (Eds.), Overview: A Phylogenetic Backbone and Taxonomic Framework for Procaryotic Systematics. Springer, New York.
Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner A., Lai, T., Steppi, S., Jobb, G., Forster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., Konig, A., Liss, T., Lussmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.H., 2004. ARB: a software environment for sequence data. Nucleic Acids Res. 32, 1363–1371.
Ludwig, W., Glöckner, F.O., Yilmaz, P., 2011. 16 - the use of rRNA gene sequence data in the classification and identification of prokaryotes. In: Rainey, F., Oren, A. (Eds.), Taxonomy of Prokaryotes. Academic Press Book 38.

Mahe, S., Duhamel, M., Le Calvez, T., Guillot, L., Sarbu, L., Bretaudeau, A., Collin, O., Dufresne, A., Kiers, E.T., Vandenkoornhuyse, P., 2012. PHYMYCO-DB: a curated database for analyses of fungal diversity and evolution. PLoS ONE 7.

Matsen, F.A., Kodner, R.B., Armbrust, E.V., 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinform. 11.

Nilsson, R.H., Wurzbacher, C., Bahram, M., Coimbra, V.R.M., Larsson, E., Tedersoo, L., Eriksson, J., Ritter, C.D., Svantesson, S., Sanchez-Garcia, M., Ryberg, M., Kristiansson, E., Abarenkov, K., 2016. Top 50 most wanted fungi. Mycokeys 29–40.

Panzer, K., Yilmaz, P., Weiß, M., Reich, L., Richter, M., Wiese, J., Schmaljohann, R., Labes, A., Imhoff, J.F., Glöckner, F.O., Reich, M., 2015a. Data from: identification of habitat-specific biomes of aquatic fungal communities using a comprehensive nearly full-length 18S rRNA dataset enriched with contextual data. Dryad Digital Repository.

Panzer, K., Yilmaz, P., Weiß, M., Reich, L., Richter, M., Wiese, J., Schmaljohann, R., Labes, A., Imhoff, J.F., Glöckner, F.O., Reich, M., 2015b. Identification of habitat-specific biomes of aquatic fungal communities using a comprehensive nearly full-length 18S rRNA dataset enriched with contextual data. PLoS One 10, e0134377.

Peplies, J., Kottmann, R., Ludwig, W., Glöckner, F.O., 2008. A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes. Syst. Appl. Microbiol. 31, 251–257.

Porter, T.M., Schadt, C.W., Rizvi, L., Martin, A.P., Schmidt, S.K., Scott-Denton, L., Vilgalys, R., Moncalvo, J.M., 2008. Widespread occurrence and phylogenetic placement of a soil clone group adds a prominent new branch to the fungal tree of life. Mol. Phylogenet. Evol. 46, 635–644.

Powell, J.R., Parrent, J.L., Hart, M.M., Klironomos, J.N., Rillig, M.C., Maherali, H., 2009. Phylogenetic trait conservatism and the evolution of functional trade-offs in arbuscular mycorrhizal fungi. P R Soc. B 276, 4237–4245.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glockner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596.

Rayner, A.D.M., Coates, D., 1987. Regulation of mycelial organization and responses. In: Rayner, A.D.M., Brasier, C.M., Moore, D. (Eds.), Evolutionary Biology of the Fungi. Cambridge Univ Press, Cambridge, UK.

Reblova, M., Miller, A., Rossman, A.Y., Seifert, K.A., Crous, P.W., Hawksworth, D.L., Abdel-Wahab, M.A., Cannon, P.F., Daranagama, D.A., De Beer, Z.W., Huang, S.K., Hyde, K.D., Jayawardena, R., Jaklitsch, W.M., Jones, E.B.G., Ju, Y.M., Judith, C., Maharachchikumbura, S.S., Pang, K.L., Petrini, L.E., Raja, H.A., Romero, A.I., Shearer, C., Senanayake, I.C., Voglmayr, H., Weir, B.S., Wijayawardene, N.N., 2016. Recommendations for competing sexual-asexually typified generic names in *Sordariomycetes* (except *Diaporthales*, *Hypocreales*, and *Magnaporthales*). IMA Fungus 7, 131–153.

Reich, M., Labes, A., 2017. How to boost marine fungal research: a first step towards a multidisciplinary approach by combining molecular fungal ecology and natural products chemistry. Mar. Genomics (in press).

Reich, M., Wichels, A., Panzer, K., Krause, E., Gimenez, L., Gerdts, G., 2017. Impacts of a reduction in seawater pH mimicking ocean acidification on the structure and diversity of mycoplankton communities. Aquat. Microb. Ecol. http://dx.doi.org/10.3354/ame01831 (in press).

Rosling, A., Cox, F., Cruz-Martinez, K., Ihrmark, K., Grelet, G.A., Lindahl, B.D., Menkis, A., James, T.Y., 2011. Archaeorhizomycetes: unearthing an ancient class of ubiquitous soil fungi. Science 333, 876–879.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Bolchacova, E., Voigt, K., Crous, P.W., Miller, A.N., Wingfield, M.J., Aime, M.C., An, K.D., Bai, F.Y., Barreto, R.W., Begerow, D., Bergeron, M.J., Blackwell, M., Boekhout, T., Bogale, M., Boonyuen, N., Burgaz, A.R., Buyck, B., Cai, L., Cai, Q., Cardinali, G., Chaverri, P., Coppins, B.J., Crespo, A., Cubas, P., Cummings, C., Damm, U., de Beer, Z.W., de Hoog, G.S., Del-Prado, R., B, D., Dieguez-Uribeondo, J., Divakar, P.K., Douglas, B., Duenas, M., Duong, T.A., Eberhardt, U., Edwards, J.E., Elshahed, M.S., Fliegerova, K., Furtado, M., Garcia, M.A., Ge, Z.W., Griffith, G.W., Griffiths, K., Groenewald, J.Z., Groenewald, M., Grube, M., Gryzenhout, M., Guo, L.D., Hagen, F., Hambleton, S., Hamelin, R.C., Hansen, K., Harrold, P., Heller, G., Herrera, G., Hirayama, K., Hirooka, Y., Ho, H.M., Hoffmann, K., Hofstetter, V., Hognabba, F., Hollingsworth, P.M., Hong, S.B., Hosaka, K., Houbraken, J., Hughes, K., Huhtinen, S., Hyde, K.D., James, T., Johnson, E.M., Johnson, J.E., Johnston, P.R., Jones, E.B., Kelly, L.J., Kirk, P.M., Knapp, D.G., Koljalg, U., GM, K., Kurtzman, C.P., Landvik, S., Leavitt, S.D., Liggenstoffer, A.S., Liimatainen, K., Lombard, L., Luangsa-Ard, J.J., Lumbsch, H.T., Maganti, H., Maharachchikumbura, S.S., Martin, M.P., May, T.W., McTaggart, A.R., Methven, A.S., Meyer, W., Moncalvo, J.M., Mongkolsamrit, S., Nagy, L.G., Nilsson, R.H., Niskanen, T., Nyilasi, I., Okada, G., Okane, I., Olariaga, I., Otte, J., Papp, T., Park, D., Petkovits, T., Pino-Bodas, R., Quaedvlieg, W., Raja, H.A., Redecker, D., T, R., Ruibal, C., Sarmiento-Ramirez, J.M., Schmitt, I., Schussler, A., Shearer, C., Sotome, K., Stefani, F.O., Stenroos, S., Stielow, B., Stockinger, H., Suetrong, S., Suh, S.O., Sung, G.H., Suzuki, M., Tanaka, K., Tedersoo, L., Telleria, M.T., Tretter, E., Untereiner, W.A., Urbina, H., Vagvolgyi, C., Vialle, A., Vu, T.D., Walther, G., Wang, Q.M., Wang, Y., Weir, B.S., Weiss, M., White, M.M., Xu, J., Yahr, R., Yang, Z.L., Yurkov, A., Zamora, J.C., Zhang, N., Zhuang, W.Y., Schindel, D., Consortium, F.B., 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. Proc. Natl. Acad. Sci. U. S. A. 109, 6241–6246.

Slepecky, R.A., Starmer, W.T., 2009. Phenotypic plasticity in fungi: a review with observations on *Aureobasidium pullulans*. Mycologia 101, 823–832.

Stajich, J.E., Berbee, M.L., Blackwell, M., Hibbett, D.S., James, T.Y., Spatafora, J.W., Taylor, J.W., 2009. The fungi. Curr. Biol. 19, R840–R845.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.

Teeling, E.C., Springer, M.S., Madsen, O., Bates, P., O'Brien, S.J., Murphy, W.J., 2005. A molecular phylogeny for bats illuminates biogeography and the fossil record. Science 307, 580–584.

Wang, H., Xu, Z., Gao, L., Hao, B.L., 2009. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evol. Biol. 9.

Webb, C.O., Ackerly, D.D., Kembel, S.W., 2008. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics 24, 2098–2100.

Wijayawardene, N.N., Crous, P.W., Kirk, P.M., Hawksworth, D.L., Boonmee, S., Braun, U., Dai, D.Q., D'souza, M.J., Diederich, P., Dissanayake, A., Doilom, M., Hongsanan, S., Jones, E.B.G., Groenewald, J.Z., Jayawardena, R., Lawrey, J.D., Liu, J.K., Lucking, R., Madrid, H., Manamgoda, D.S., Muggia, L., Nelsen, M.P., Phookamsak, R., Suetrong, S., Tanaka, K., Thambugala, K.M., Wanasinghe, D.N., Wikee, S., Zhang, Y., Aptroot, A., Ariyawansa, H.A., Bahkali, A.H., Bhat, D.J., Gueidan, C., Chomnunti, P., De Hoog, G.S., Knudsen, K., Li, W.J., McKenzie, E.H.C., Miller, A.N., Phillips, A.J.L., Piatek, M., Raja, H.A., Shivas, R.S., Slippers, B., Taylor, J.E., Tian, Q., Wang, Y., Woudenberg, J.H.C., Cai, L., Jaklitsch, W.M., Hyde, K.D., 2014. Naming and outline of *Dothideomycetes*-2014 including proposals for the protection or suppression of generic names. Fungal Divers. 69, 1–55.

Wu, Q.X., Mueller, G.M., Lutzoni, F.M., Huang, Y.Q., Guo, S.Y., 2000. Phylogenetic and biogeographic relationships of eastern Asian and eastern north American disjunct *Suillus* species (Fungi) as inferred from nuclear ribosomal RNA ITS sequences. Mol. Phylogenet. Evol. 17, 37–47.

Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzeby, J., Amann, R., Rossello-Mora, R., 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. 12, 635–645.

Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., Glöckner, F.O., 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 42, D643–D648.

Yilmaz, P., Yarza, P., Rapp, J.Z., Glöckner, F.O., 2015. Expanding the world of marine bacterial and archaeal clades. Front. Microbiol. 6, 1524.