PHYMYCO-DB: A Curated Database for Analyses of Fungal Diversity and Evolution

Stéphane Mahé, Marie Duhamel, Thomas Le Calvez, Laëtitia Guillot, Ludmila Sarbu, Anthony Bretaudeau, Olivier Collin, Alexis Dufresne, E. Toby Kiers, Philippe Vandenkoornhuyse

▶ **To cite this version:**

Stéphane Mahé, Marie Duhamel, Thomas Le Calvez, Laëtitia Guillot, Ludmila Sarbu, et al.. PHYMYCO-DB: A Curated Database for Analyses of Fungal Diversity and Evolution. PLoS ONE, 2012, 7 (9), pp.e43117. 10.1371/journal.pone.0043117 . hal-00756250

HAL Id: hal-00756250
https://hal.science/hal-00756250v1

Submitted on 22 Nov 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

OPEN ACCESS Freely available online

PLOS ONE

PHYMYCO-DB: A Curated Database for Analyses of Fungal Diversity and Evolution

Stéphane Mahé[1,2], Marie Duhamel[1,2,3], Thomas Le Calvez[1,2,5], Laetitia Guillot[2,4], Ludmila Sarbu[2,4], Anthony Bretaudeau[2,4], Olivier Collin[2,4], Alexis Dufresne[1,2], E. Toby Kiers[3], Philippe Vandenkoornhuyse[1,2]*

1 Université de Rennes I, CNRS, UMR 6553 ECOBIO, Campus de Beaulieu, Rennes, France, **2** Université Européenne de Bretagne, Rennes, France, **3** Department of Ecological Science, Vrije Universiteit, Amsterdam, The Netherlands, **4** Université de Rennes I, CNRS, UMR 6074 IRISA, Campus de Beaulieu, Rennes, France, **5** Centre Scientifique et Technique du Bâtiment, AQUASIM, Nantes, France

Abstract

Background: In environmental sequencing studies, fungi can be identified based on nucleic acid sequences, using either highly variable sequences as species barcodes or conserved sequences containing a high-quality phylogenetic signal. For the latter, identification relies on phylogenetic analyses and the adoption of the phylogenetic species concept. Such analysis requires that the reference sequences are well identified and deposited in public-access databases. However, many entries in the public sequence databases are problematic in terms of quality and reliability and these data require screening to ensure correct phylogenetic interpretation.

Methods and Principal Findings: To facilitate phylogenetic inferences and phylogenetic assignment, we introduce a fungal sequence database. The database PHYMYCO-DB comprises fungal sequences from GenBank that have been filtered to satisfy stringent sequence quality criteria. For the first release, two widely used molecular taxonomic markers were chosen: the nuclear SSU rRNA and EF1-α gene sequences. Following the automatic extraction and filtration, a manual curation is performed to remove problematic sequences while preserving relevant sequences useful for phylogenetic studies. As a result of curation, $\sim$20% of the automatically filtered sequences have been removed from the database. To demonstrate how PHYMYCO-DB can be employed, we test a set of environmental Chytridiomycota sequences obtained from deep sea samples.

Conclusion: PHYMYCO-DB offers the tools necessary to: (i) extract high quality fungal sequences for each of the 5 fungal phyla, at all taxonomic levels, (ii) extract already performed alignments, to act as 'reference alignments', (iii) launch alignments of personal sequences along with stored data. A total of 9120 SSU rRNA and 672 EF1-α high-quality fungal sequences are now available. The PHYMYCO-DB is accessible through the URL http://phymycodb.genouest.org/.

Citation: Mahé S, Duhamel M, Le Calvez T, Guillot L, Sarbu L, et al. (2012) PHYMYCO-DB: A Curated Database for Analyses of Fungal Diversity and Evolution. PLoS ONE 7(9): e43117. doi:10.1371/journal.pone.0043117

Editor: Dirk Steinke, Biodiversity Insitute of Ontario - University of Guelph, Canada

Received March 15, 2012; **Accepted** July 16, 2012; **Published** September 13, 2012

Copyright: © 2012 Mahé et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by a grant from the ''Total Corporate Foundation for biodiversity and the sea'' and a grant from the French National Agency for Research within the Systerra call (ANR-10-STRA-002). MD has a Ph.D. grant from the French ministry of research. ETK is supported by an Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) ''vidi'' and ''meervoud'' grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: philippe.vandenkoornhuyse@univ-rennes1.fr

Introduction

In recent years there has been an exponential increase in the number of gene sequences available in public-access databases. This is the result of new developments in molecular techniques and new generation sequencers that allow the collection of data at great speed. The use of molecular taxonomic markers associated with phylogenetic analyses has revealed considerable genetic diversity in fungi, especially those that are cryptic, unculturable or not easily distinguishable by morphological characters (e.g. [1]). As the species concept is employed for diversity measurements, systematics and evolutionary analyses [2], an efficient means of identifying boundaries, and thus number of species, is required. Molecular methods and the implicit adoption of the phylogenetic species concept [3] offer a standardized approach to delimit groups of organisms (e.g. [4–6]). Thanks to progress in sequencing technologies and bioinformatic methods, the detection of orthologous sequences using databases is relatively efficient. This approach can also be successfully applied to organisms that are not available in culture, increasing our ability to identify new diversity in various habitats [7,8]. Of course, this approach requires choosing a relevant molecular marker which: (i) targets a nucleic acid sequence with a limited proportion of homoplasy (i.e. correspondence between parts arising from evolutionary convergence), (ii) contains high phylogenetic information which is not sensitive to paralogy (i.e. single copy genes or highly conserved genes). This allows for accurate characterization of evolutionary affinities.

In this context, the nuclear gene coding for the small subunit of the ribosomal RNA (SSU rRNA) is often seen as the 'ultimate' molecular marker [9], (for review [10]). The SSU rRNA gene is present in all living organisms. Its sequence is highly conserved between taxa, reflecting strong functional constraints on the translational machinery. Indeed, most mutations in the SSU rRNA gene sequence reduce the stability of the secondary structure of the SSU rRNA molecule and thus the efficiency of protein synthesis. Furthermore, this gene, like other informational genes, appears to be less subject to horizontal gene transfers and is believed to provide better inferences of 'true' phylogenies [11]. Although the SSU rRNA gene can have a multicopy status within a single fungal genome, sequence variations have been shown to be extremely low or null. For example, from available complete annotated genomes (http://www.genomesonline.org/cgi-bin/GOLD/index.cgi), *Saccharomyces cerevisiae* has two SSU rRNA copies both on its chromosome XII. *Encephalitozoon cuniculi*, a Microsporidia, has two SSU rRNA genes copies one on its chromosome I, the other on chromosome IV. In these two cases, the copies display 100% identity. This is not surprising since the SSU rRNA gene is highly conserved. Thus this gene is less sensitive to paralogy compared to LSUrRNA gene and ITS where variations among copies have been clearly shown (e.g. [12–14]).

A second advantage of using the SSU rRNA gene sequence is its huge representation in international public databases - GenBank [15], EMBL/ENA [16], DDBJ [17] – which facilitates comparisons between a wide variety of organisms (for review [18]). One disadvantage is that because the SSU rRNA gene is highly conserved, the resolution of the phylogenetic analyses is poor for youngest fungal groups within Ascomycota. Other genes, such as those encoding for the elongation factor EF1-α (*tef1*), for β-tubulin (*tub1*, *tub2*), actin (*act1*), or for RNA polymerase II subunits (*rpb1* and *rpb2*), can be used as alternative markers. Among these ones, EF1-α sequence data are the most abundant but only represent a small fraction of the amount of SSU rRNA yet available (i.e. less than 7% of the total number of sequences contained in PHYMYCO-DB). Generally present as a single copy gene, the EF1-α gene is involved in protein synthesis and displays a higher mutation rate than SSU rRNA gene. Because of these attributes, EF1-α protein sequences have been used to resolve phylogenetic affinities between eukaryotic organisms [19–21], and particularly the sister clade relationship of animals and fungi [22]. The gene sequences also have the potential to help resolve phylogenetic relationships between closely related fungi [21,23–24], but they contain a higher proportion of homoplasious positions compared to SSU rRNA gene sequences. Studies of both SSU rRNA genes and EF1-α genes could greatly improve the resolution of fungal phylogenetic affinities. An online database incorporating data from both these sequences is a key step to achieving improved phylogenetic resolution for fungi.

Pollution of public sequence database and the aim of PHYMYCO-DB

One major obstacle for international public databases is constant pollution by non-negligible proportions of compromised sequences (GenBank/EMBL/DDBJ). This problem, discussed in several articles and journal forums (e.g. [25–32]), is becoming more and more obvious, but solutions remain elusive. Problematic data can arise from many different origins, including: (i) erroneous specimen identification [27], (ii) the use of separate names for different sexual stages [30], (iii) differences in taxonomy among specialists [27] and/or advances in knowledge since the time the sequence was deposited leading to wrong designations [29], (iv) the lack of precision in the description of the deposited sequences making their interpretation difficult [33], (v) sequences resulting from artefactual origin (i.e. chimeric sequences), and (vi) sequences of poor quality with undefined positions. Even more problematic is the erroneous annotated sequences that propagate within open access databases because of phylogenetic misinterpretation. Additionally, more and more sequence assignments are based solely on identity searches using heuristic local alignment (i.e. BLASTn searches). All these mistakes have the potential to jeopardize interpretations. Therefore, assessing the reliability of sequences is an increasingly important prerequisite to analyses.

Many of these errors can be limited via expert curation. Expert curation is critical for the continued advancement of the field because it allows for the production of sequence databases, containing accurate and reliable sequences. To date, most curated databases specialize in particular taxonomic groups (e.g. [34]), collect data associated to each nucleic acid sequence, and work with specimens validated by experts and deposited in public reference collections (e.g. [33]). Several important tools, such as the Ribosomal Database Project [35], SILVA [36], Greengenes database [37] exist online for the analysis of SSU rRNA gene sequences. Apart from SILVA, these databases use automated filters to remove part of the polluting sequences. However, manual curation is an essential component of these projects and should aim to be even more stringent.

Based on lessons learned from other curated databases, our aims at PHYMCO-DB are to: (i) develop an easy-to-use fungal-dedicated database with stored sequences of high quality, (ii) use selected molecular markers that are widely acknowledged, namely SSU rRNA and EF1-α, (iii) produce a tool, based on anchor sequences covering the fungal tree, that can be automatically updated, along with an expert curation of the new sequences, (iv) produce high quality multiple alignments for use in testing environmental sequences or evolutionary hypotheses.

Database Structure : Design and Implementation

The sequences constituting PHYMYCO-DB version 1 (Fig. 1) were retrieved in October 2011 from the release 185 of GenBank (NCBI). The nuclear SSU rRNA and EF1-α genes sequences are extracted from the GenBank database, using the following queries: "[organism] and (ssu|SSUrRNA|SSU rRNA|18SrRNA|18S|) not (16S|mitoch*|28S|5.8S|ITS|Internal Transcribed Spacer|internal transcribed spacer|)" and "[Organism] and (EF1 alpha|EF-1 alpha|EF1-alpha|EF-1alpha|EF-1-alpha|EF1alpha|EF1a|)". After this extraction step, automatic quality filter parameters are applied. For SSU rRNA, nucleic acid sequences that are shorter than 1000 nucleotides and longer than 2500 nucleotides are rejected. Likewise for EF1-α genes, sequences shorter than 700 nucleotides and longer than 2500 nucleotides are discarded. Also sequences containing more than 10 consecutive undetermined nucleotides are excluded. According to the automatic quality criteria, all accepted sequences are then stored in a MySQL 5 relational database. The MySQL table structure is presented as a figure available in supplementary online information (Fig. S1). PHYMYCO-DB is automatically updated 4 times a year and is managed by administrators using the web interfaces developed with PHP version 4 programming language.

Following automatic filtering, datasets are then cross-checked by expert curators (hereafter 'expert curation'). Multiple alignments are performed using Clustal X 2.1 [38] on small sequence groups (<400 sequences), which are closely related to obtain a high-quality alignment and to make the expert curation as accurate as possible. Sequences are deleted from the alignment and from the database in a manual cleaning process if they contain: errors of

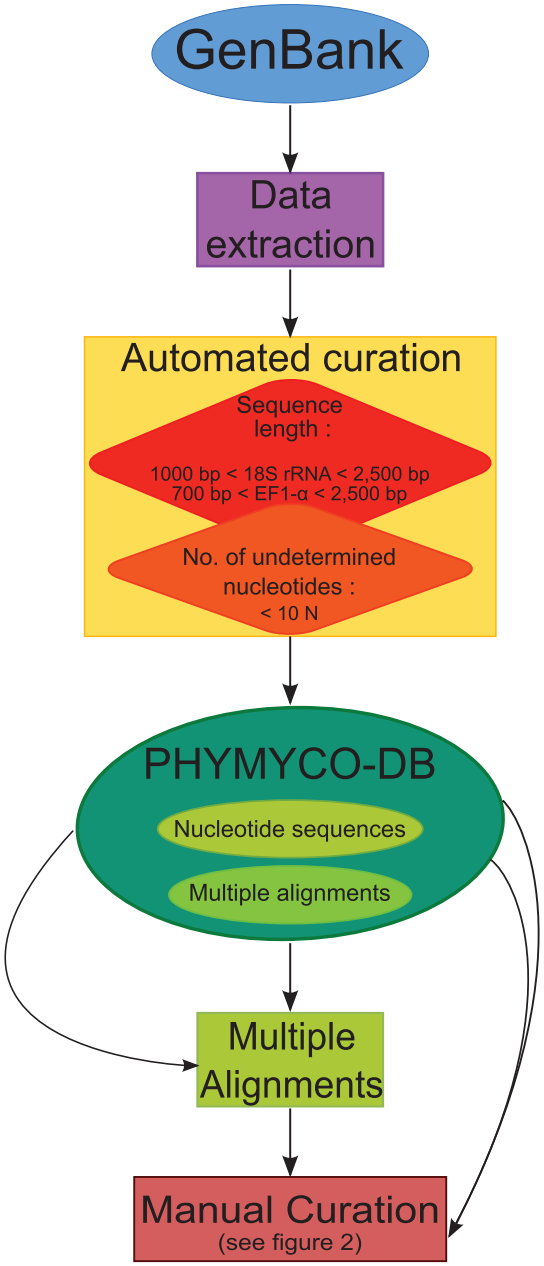


Figure 1. Flowchart of the data in the PHYMYCO-DB. The arrows indicate the flow of gene sequences extracted from the GenBank database, through the automated and manual curation steps. All the sequences made available to users has passed the 2 curation processes. After each upgrade of the database (i.e. 4 times per year), expert manual curation is performed.
doi:10.1371/journal.pone.0043117.g001

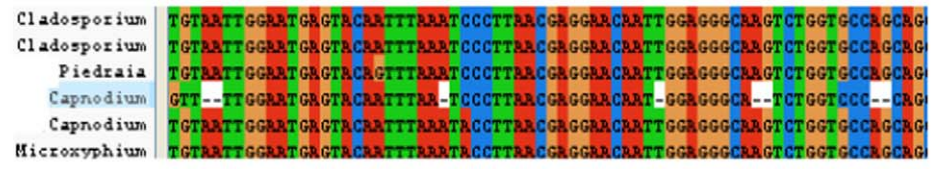
sequencing (i.e. containing several substitutions that are not found anywhere else, Fig. 2), errors in the annotation (i.e. a sequence with a naming inside a different group, Fig. 2), homopolymers insertions (Fig. 2), many undetermined nucleotides (Fig. 2), erroneous alignment or reverse complementary sequences (Fig. 2). This expert curation is time consuming but essential to obtain reliable sequences and high-quality alignments. By adopting strict rules of expert curation, subjectivity and mistakes become minimal. Following expert curation, species redundancy (i.e. identical sequences) are retained in the database to keep sequences arising from different origin and ecological settings. The detection of dubious sequences from the alignments does not result in correction of the sequence in international databases. They are, however, all removed from PHYMYCO-DB. When corrections are made for a given sequence, a new registration number is provided by GenBank for example. In this case, the corrected sequence will be automatically extracted (i.e. 4 updates per year) and will be examined by one of the expert curators.

During our development process, it became clear that our automatic filters were not stringent enough to retrieve only trustworthy sequences. For example, SSU rRNA can present intron-like regions which could also be chimeric insertions. Introns are abundant in particular lineages of fungi, especially within lichen-forming fungi (Ascomycota). These fungi can display up to eight introns in the SSU rRNA gene, as for example found in the taxon *Physconia* [39]. At the expert curation stage, we noticed that the position of introns was not consistently given in the deposited sequence description, and they were detectable after the alignment only. When a sequence containing non-positioned introns was the only sequence of a particular genus, this sequence was kept, Otherwise the sequence was discarded from PHYMYCO-DB. Employing our curation principles, we discarded 2090 additional unreliable sequences, i.e. 18% of the sequences extracted from GenBank.

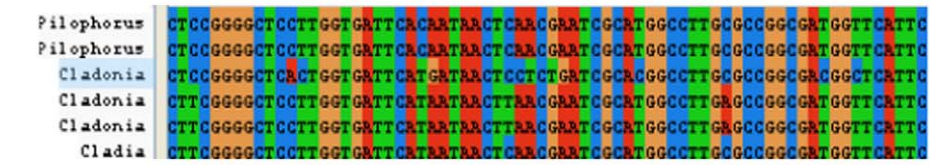
Following the curation steps, 8757 SSU rRNA gene sequences have been stored in PHYMYCO-DB (5088 Ascomycota, 2088 Basidiomycota, 366 Chytridiomycota, 1046 Glomeromycota, and 532 Zygomycota). PHYMYCO-DB also contains 648 EF1-α gene sequences (294 Ascomycota, 189 Basidiomycota, 10 Chytridiomycota, 25 Glomeromycota, and 154 Zygomycota). Our database contains less fungal sequences than SYLVA because of the level of curation stringency. All fungal genera has at a minimum one representative sequence within PHYMYCO-DB. Because of the heterogeneity among the number of sequences per taxonomic rank, and because we wanted a limited number of sequences for each alignment, the taxonomic level within these alignments is variable (family to phylum level). We therefore produced a total of about 50 'reference' alignment files. These online alignments contain mainly full-length sequences, even if rare, very long sequences were cut at the same length as the others. This was done to keep maximum information available. This is especially useful for designing primers, and to give a greater freedom for manipulation by online users.

Tools within PHYMYCO-DB

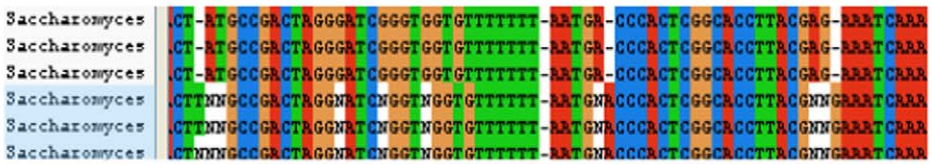
We designed PHYMYCO-DB with specific tools to facilitate online use. Firstly, users can easily select sequences by browsing our interface through hierarchical taxonomic lineages presented in an arborescent structure (GenBank taxonomy), and then download them in a FASTA format file. The number of sequences stored in the database for each taxonomic level is given in brackets. Secondly, users can download an alignment file using a filter to find an alignment with the gene and the taxonomic rank requested. Special attention must be paid to the fact that some sequence characteristics in PHYMYCO-DB format are inherited from the extraction of GenBank sequences. For example, in some cases (e.g. Agaromycotina, a subphylum of Ascomycota), information on sequences taxonomy was associated to a 'no rank' tag in GenBank. To avoid the problem that these sequences are mistakenly placed in another taxonomic group, they were qualified as 'undefined' at the subphylum rank in PHYMYCO-DB. For the next lower taxonomic rank, no known tag problem exists. Environmental sequences have, by definition, no clear taxonomic



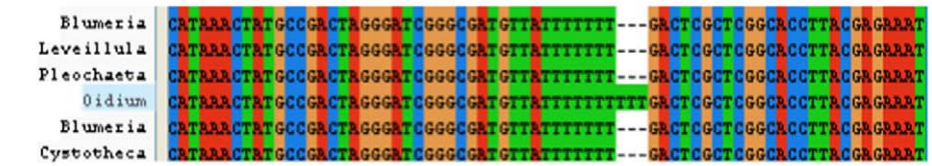
(a) Errors of sequencings



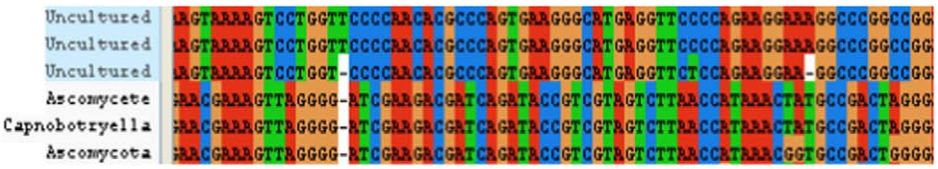
(b) Errors in the annotation



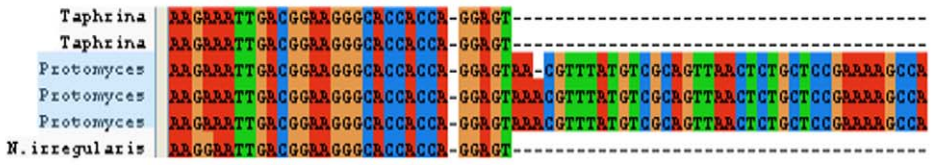
(c) Many undetermined nucleotides



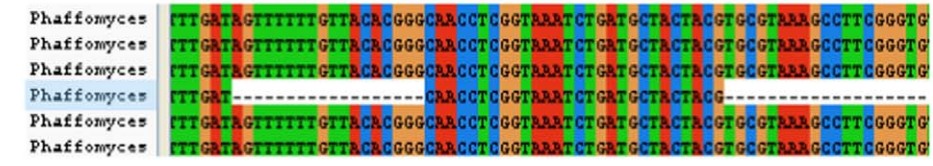
(d) Homopolymers



(e) Reverse complementary sequences



(f) Long insertions and introns



(g) Deletions

Figure 2. Visualisation of sequences deleted by the manual curation after alignment (ClustalX 2.1). The sequences highlighted in blue illustrate examples of sequences removed from PHYMYCO-DB. The compromised nature can stem from erroneous sequencing (e.g. repeated gaps), wrong annotation (e.g. sequence corresponding to another clade), high numbers of undetermined nucleotides, homopolymers insertions, erroneous alignment or reverse complementary sequences and presence of long insertions and introns or presence of deletions.
doi:10.1371/journal.pone.0043117.g002

ranking. Therefore, they were also qualified as 'undefined', but only until the lowest taxonomic rank. These are important features to take into account when using the PHYMYCO-DB.

Thirdly, users can launch a ClustalW 2.0 alignment on our back-end computer clusters by uploading their own personal sequences in a FASTA or ALN format file. A future PHYMYCO-DB release will offer the possibility to select the multiple alignment tool (i.e. ClustalW, MUSCLE, and MAFFT). Currently, users can choose to append an outgroup or sequences from a particular PHYMYCO-DB taxonomic group. We anticipate that this tool will be very efficient when combined with phylogenetic analyses for investigating the sequence diversity of fungal amplicons from an environmental sample and even to identify new fungal lineages.

PHYMYCO-DB as a Tool for Phylogenetic Identifications and Inferences

Based on a well-developed theoretical corpus, phylogenies can be computed using several different approaches (e.g. [40]). From a mathematical point of view, the maximum likelihood phylogenetic reconstruction provides the best possible tree for a given explicit sequence evolution model. The model that best fits the aligned sequence data can be selected, after using the popular Modeltest [41]. Achieving a good alignment is therefore of tremendous importance for good interpretation. Alignments should be refined using an 'influence function' that allows the removal of outlier columns from the matrix (i.e. nucleotide position where the phylogenetic signal differs from the general phylogenetic information recorded in the dataset) [42]. This approach allows for a 'blind detection' of outliers using measures of each site in a context of a ML phylogenetic reconstruction. It must be emphasized that the sequence-based identification using SSU rRNA gene could be at the species level or at higher taxonomic levels depending on the fungal affiliation.

Following the above strategy, we provide an analysis of chytrid diversity as a proof of concept. Sequencing of the SSU rRNA gene was achieved by targeting chytrids from deep marine hydrothermal samples (ciPCR). First, the alignment of SSU rRNA gene sequences of the Chytridiomycota from PHYMYCO-DB were used to design specific primers manually. Two sets of designed primers covered the V3 and V4 variable regions and were suitable for pyrosequencing: C130 (5′TACCTTACTACTTGGATAACCG3′) with SR8R (5′TCAAAGTAAAAGTCCTGGATC3′) modified from Vilgalys lab webpage (http://www.biology.duke.edu/fungi/mycolab/primers.htm), and MH2 (5′TTCGATGGTAGGATAGAGG3′) [43] with SR8R. Another set of primers, expected to be universal for fungi and to produce longer amplicons, were also tested: MH2 with NS7R (5′ATCACAGACCTGTTATTGCC3′) modified from [44]. Primers specificity was checked with a sample from a hydrothermal site from which several sequences of chytrids were retrieved [45]. The resulting sequences (GenBank accession numbers JN986721 to JN986723) were analyzed using the corresponding 'reference' alignment in PHYMYCO-DB and the sequences having the highest similarity score in BLASTn. The computed phylogeny highlights the presence of a new group within the Chytridiomycota phylum (Fig. 3). The three OTUs present high identity level (>98%) with environmental sequences ,and form a monophyletic group whose closest described relative is a sequence from the genus *Maunachytrium*. These OTUs constitute a new clade in the Lobulomycetaceae family [46]. BLASTn searches of these environmental sequences return the *Maunachytrium* sequence as the best hit, with a maximal identity of 96%. The widely used BLAST-based annotation for environmental sequences, would end with an assignation to *Maunachytrium keaense* or *Maunachytrium* sp. However, by choosing a phylogenetic approach, the analysis goes into greater depth. The initial positioning of these sequences suggests that they form a new clade within the Lobulomycetaceae family, outside the *Maunachytrium*, *Lobulomyces* (maximal identity 93%) and *Clydaea* (maximal identity 92%) genera.

This exercise thus highlights important differences between phylogeneticaly based annotation and BLASTn annotation. More and more identifications rely solely on BLAST searches which allow for faster analyses of the rapidly increasing numbers of environmental sequences. Indeed many analyses and tools developed for mass sequencing are based on BLAST searches (e.g. MEGAN). We would argue that this approach is less conservative and more prone to mistakes. The use of phylogenetic approaches, when it is possible should be favoured, to avoid increasing the presence of polluting sequences in international sequences databases.

Discussion

The release of PHYMYCO-DB is expected to provide comprehensive access to fungal sequences for two phylogenetic markers (SSU rRNA and EF1-α genes) obtained from cultivated isolates, as well as environmental samples. As a result of deep sequence cleaning, the aligned sequences available in PHYMYCO-DB are of high quality (Fig. 1). To our knowledge, this curation strategy provides a novel approach to the problem of database pollution. As such, we anticipate that it will complement other existing databases such as the "Assembling the Fungal Tree Of Life" project (AFTOL) [47], UNITE [33,48] and MaarjAM [34] which are restricted to fungal sequences.

Curation and annotation of ITS is made possible through the web-based-workbench of PlutoF [49]. Initially, the UNITE system contained ITS and nLSU/28S rRNA gene sequences from Basidiomycota and Ascomycota. Based on recent work, the ITS region is now being suggested as a possible universal DNA barcode marker for fungi [50]. It is accepted that the ITS region is valuable at species level and so, more taxonomically informative than SSU rRNA gene sequences for analysing groups of organisms that have emerged 'recently' and are closely related [51], e.g. Ascomycota and Basidiomycota. The ITS region is also often used to resolve phylogenetic relationships at the species level or at the infraspecific level [52]. However, as the ITS region displays high sequence variability, even within a given organism as in Glomeromycota (i.e. [13]), obtaining reliable alignments with this marker can be difficult [53] and potentially precludes multiple alignments. This is because accurate comparisons are hindered by the accumulated homoplasy and the high frequency of insertion/deletion events. The use of the SSU rRNA sequences is interesting since new groups, within all the fungal phyla including Ascomycota and Basidiomycota, can be detected (i.e. [1,54]). The MaarjAM database has focused on SSU rRNA gene of arbuscular

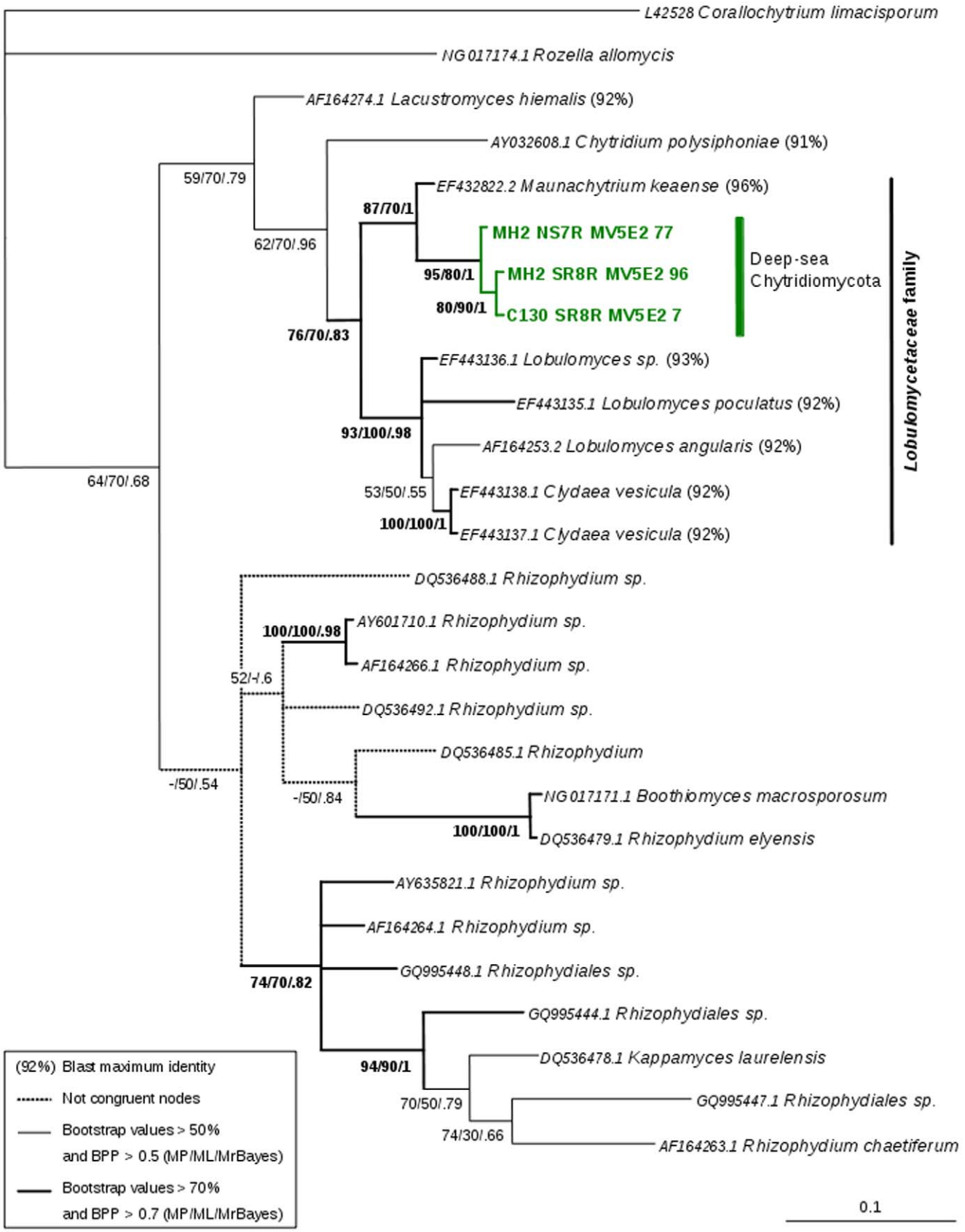


Figure 3. SSU rRNA phylogenetic positions of deep-sea Chytridiomycota (colored terminals) along with the closest known related SSU rRNA fungal sequences. Topology was built using MrBayes v.3.1.2 (Scale bar: 0.1 estimated substitutions per site, 3000000 generations sampled every 100 generations and an average standart deviation of split frequencies of 0.004140) from a ClustalW 2.1 alignment. The model

GTR+I+G was designated by jModelTest 0.1. Node support values are given in the following order: Maximum Parsimony/Maximum Likelihood (both calculated with PAUP 4.0β10 version, 500 bootstraps)/MrBayes. *Corallochytrium limacisporum* (L42528), a putative choanoflagellate, was used as outgroup. *Maunachytrium keaense* (it is not part of PHYMYCO-DB) was also used to help build the tree. All sequences are listed with their GenBank accession numbers. The topologies were congruent apart from doted lines indicated in the figure. Thin lines show bootstrap values >50% and BPP >0.5 (MP/ML/MrBayes) and thick lines: bootstrap values >70% and BPP >0.7 (MP/ML/MrBayes). The sequences belonging to the *Lobulomycetaceae* family are indicated with their BLASTn percentage of maximum identity compared to the three deep-sea Chytridiomycota OTUs.
doi:10.1371/journal.pone.0043117.g003

mycorrhizal fungi (Glomeromycota), with associated metadata. The existence of this database and the potential emergence of others should be encouraged. It enables the community to have access to reliable sequences.

For fungal sequence annotations and phylogenetic interpretations of fungal environmental sequences, one of the main advantages of PHYMYCO-DB is to facilitate the primer design and subsequent phylogenetic analyses of amplicons as shown in the example above (Fig. 3). The use of PHYMYCO-DB to perform expert analyses appears to be complementary to BLASTn, the latter allowing a quick look of the query sequence proximity compared to the available sequences. From the phylogenetic analyses performed one arising interpretation is that different apparent polyphyletic groups may be a consequence of wrong annotations. We anticipate that the use of PHYMYCO-DB will help to limit incorrect SSU rRNA and EF1-α genes fungal annotation propagation in sequence databases.

Availability and Future Directions

The PHYMYCO-DB is available via a web-based interface at http://phymycodb.genouest.org/ on the GenOuest bioinformatics platform web site. The web interface is divided into 2 parts. The first part, entitled "*DB admin*", is restricted to the administrators for use in cleaning and optimising the database. The second part, entitled "*DB explore*", is publicly accessible to all users. The next set of PHYMYCO-DB releases will include (i) the provision of alignment files in which outlier nucleotides identified from influence functions [42] will be highlighted, so that users can then delete these sites (ii) taxonomic modifications within Chytridiomycota and Zygomycota after Hibbett et al. (2007) [55] and after Jones et al. (2011) [5]. PHYMYCO-DB will continue to expand with new genes. We are currently investigating β-tubulin (*tub1*, *tub2*), actin (*act1*), and RNA polymerase II subunits (*rpb1* and *rpb2*) as potential interesting targets. PHYMYCO-DB will also be improved by incorporating all the finished fungal genomes available, and increasing the diversity of tools to perform multiple alignments.

Supporting Information

Figure S1 MySQL table structure of PHYMYCO-DB.
(TIF)

Acknowledgments

We thank the GenOuest Bioinformatics core facility and anonymous referees for valuable comments on the manuscript.

Author Contributions

Conceived and designed the experiments: PV AD OC. Performed the experiments: SM MD. Analyzed the data: SM MD TLC. Contributed reagents/materials/analysis tools: AD OC LG LS AB ETK. Wrote the paper: PV SM MD ETK. Creation and design of the database: OC LG LS AB. Extraction of data from Genbank: OC LG LS AB.

References

1. Vandenkoornhuyse P, Baldauf SL, Leyval C, Straczek J, Young JPW (2002a). Extensive fungal diversity in plant roots. Science 295: 2051.
2. Purvis A, Hector A (2000) Getting the measure of biodiversity. Nature 405: 212–219.
3. Taylor JW, Jacobson DJ, Kroken S, Kasuga T, Geiser DM (2000) Phylogenetic species recognition and species concept in fungi. Fung Genet Biol 31: 21–32.
4. Vandenkoornhuyse P, Husband R, Daniell TJ, Watson IJ, Duck JM, et al. (2002b). Arbuscular mycorrhizal community composition associated with two plant species in a grassland ecosystem. Mol Ecol 11: 1555–1564.
5. Jones MDM, Forn I, Gadelha C, Egan MJ, Bass D, et al. (2011) Discovery of novel intermediate forms redefines the fungal tree of life. Nature 474, 200–203
6. Powell JR, Monaghan MT, Öpik M, Rillig MC (2011) Evolutionary criteria outperform operational approaches in producing ecologically relevant fungal species inventories. Molecular Ecology 20: 655–666.
7. Hawksworth DL, Rossman AY (1997) Where are all the undescribed fungi? Phytopath 87: 888–891.
8. Blackwell M (2011) The Fungi: 1, 2, 3 … 5.1 million species? Am J Bot 98: 426–438
9. Woese CR (2000) Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA 97: 8392–8396.
10. Pace NR (2009) Mapping the Tree of Life: Progress and Prospects. Microbiology and Mol Biol Rev 73: 565–576.
11. Choi IG, Kim SH (2007) Global extent of horizontal gene transfer. Proc Natl Acad Sci USA 104: 4489–4494
12. Boon E, Zimmerman E, Lang BF, Hijri M (2010) Intra-isolate genome variation in arbuscular mycorrhizal fungi persists in the transcriptome. J Evol Biol 23:1519–1527
13. Sanders IR, Alt M, Groppe K, Boller T, Wiemken A (1995) Identification of ribosomal DNA polymorphisms among and within spores of the Glomales - Applications to study on the genetic diversity of Arbuscular Mycorrhizal fungal communities. New Phytol 130: 419–427
14. Lim YW, Sturrock R, Leal I, Pellow K, Yamaguchi T, et al. (2008) Distinguishing homokaryons and heterokaryons in Phellinus sulphurascens using pairing tests and ITS polymorphisms. Antonie van Leeuwenhoek 93: 99–110.
15. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. Nucl Acids Res 39: D32–D37
16. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, et al. (2007) The EMBL nucleotide sequence database in 2006. Nucl Acids Res 35: D16–D20.
17. Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, et al. (2011) DDBJ progress report. Nucl Acids Res 39: D22–D27.
18. Avise JC (2004) Molecular Markers, Natural History, and Evolution. Sunderland, Massachusetts: Sinauer Associates.
19. Baldauf SL (1999) A search for the origins of Animals and Fungi: comparing and combining molecular data. Am Nat 154: S178–S188
20. Baldauf SL, Roger AJ, Wenk-Siefert I, Doolitle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290: 972–977
21. Helgason T, Watson IJ, Young JPW (2003) Phylogeny of the Glomerales and Diversisporales (Fungi: Glomeromycota) from actin and elongation factor 1-alpha sequences. FEMS Microb Let 229: 127–132.
22. Baldauf SL, Palmer JD (1993) Animals and fungi are each others closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci USA 90: 11558–11562
23. Moon CD, Miles CO, Jarlfors U, Schardl CL (2002) The evolutionary origins of three new Neotyphodium endophyte species from grasses indigenous to the Southern Hemisphere. Mycologia 94: 694–711.
24. Tanabe Y, Saikawa M, Watanabe MM, Sugiyama J (2004). Molecular phylogeny of Zygomycota based on EF-1α and RBP1 sequences: limitations and utility of alternative markers to rDNA. Mol Phyl Evol 30: 438–449.
25. Bidartondo MI (2008) Preserving accuracy in GenBank. Science 319: 1616
26. Bridge PD, Roberts PJ, Spooner BM, Panchal G (2003) On the unreliability of published DNA sequences. New Phytol 160: 43–48
27. Vilgalys R (2003) Taxonomic misidentification in public DNA databases. New Phytol 160: 4–5.
28. Bridge PD, Spooner BM, Roberts PJ (2004) Reliability and use of published sequence data. New Phytol 161: 15–17

29. Hawksworth DL (2004) 'Misidentification' in fungal DNA sequence databanks. New Phytol 161: 13–15.
30. Hawksworth DL (2009) Separate name for fungus's sexual stage may cause confusion. Nature 458: 29.
31. Holst-Jensen A, Vrålstad T, Schumacher T (2004) On reliability. New Phytol 161: 11–13.
32. Nilsson RH, Ryberg M, Kristiansson E, Abarenkov K, Larsson KH, et al. (2006) Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. PloS One 1: e59.
33. Kõljalg U, Larsson KH, Abarenkov K, Nilsson RH, Alexander IJ, et al. (2005) UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytol 166: 1063–1068.
34. Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, et al. (2010) The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). New Phytol 188: 223–241.
35. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al. (2009) The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. Nucl Acids Res 37: D141–D145
36. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucl Acids Res 35: 7188–7196.
37. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Env Microb 72: 5069–5072
38. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.
39. Bhattacharya D, Friedl T, Helms G (2002) Vertical evolution and intragenic spread of lichen-fungal group I intron. J Mol Evol 55: 74–84
40. Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.
41. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817–818.
42. Bar-Hen A, Mariadassou M, Poursat MA, Vandenkoornhuyse P (2008) Influence function for robust phylogenetic reconstructions. Mol Biol Evol 25: 869–873
43. Vandenkoornhuyse P, Leyval C (1998) SSU rDNA sequencing and PCR-fingerprinting reveal genetic variation within Glomus mossae. Mycologia 90: 791–797.
44. White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ editors. PCR protocols, a guide to methods and applications. San Diego, California Academic Press pp. 315–322.
45. Le Calvez T, Burgaud G, Mahé S, Barbier G, Vandenkoornhuyse P (2009) Fungal diversity in deep-sea hydrothermal ecosystems. Appl Env Microb 75: 6415–6421.
46. Simmons DR, James TY, Meyer AF, Longcore JE (2009) Lobulomycetales, a new order in the Chytridiomycota. Mycological Res 113: 450–460.
47. Lutzoni F, Kauff F, Cox CJ, McLaughlin D, Celio G, et al. (2004) Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits. American J Bot 91: 1446–1480.
48. Abarenkov K, Nilsson RH, Larsson KH, Alexander IJ, Eberhardt U, et al. (2010a). The UNITE database for molecular identification of fungi - recent updates and future perspectives. New Phytol 186: 281–285
49. Abarenkov K, Tedersoo L, Nilsson RH, Vellak K, Saar I, et al. (2010b) PlutoF-a Web Based Workbench for Ecological and Taxonomic Research, with an Online Implementation for Fungal ITS Sequences. Evol Bioinformatics 6: 189–196
50. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barecode marker for fungi. Proc Natl Acad Sci USA 109: 6241–6246
51. Anderson IC, Parkin PI (2007). Detection of active soil fungi by RT-PCR amplification of precursor rRNA molecules. J Microb Meth 68: 248–253
52. Xu P, Han Y, Wu J, Lv H, Qiu L, et al. (2007) Phylogenetic Analysis of the Sequences of rDNA Internal Transcribed Spacer (ITS) of Phytophthora sojae. J Genet Genom 34: 180–188.
53. D'Auria G, Pushker R, Rodriguez-Valera F (2006) IWoCS analyzing ribosomal intergenic transcribed spacers configuration and taxonomic relationships. Bioinformatics 22: 527–531
54. Bass D, Howe A, Brown N, Barton H, Demidova M, et al. (2007) Yeast forms dominate fungal diversity in the deep oceans. Proc Royal Soc B 274: 3069–3077
55. Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, et al. (2007) A higher-level phylogenetic classification of the Fungi. Mycol Res 111: 509–547.