MycoKeys 12: 29–40 (2016)
doi: 10.3897/mycokeys.12.7553
http://mycokeys.pensoft.net

RESEARCH ARTICLE

A peer-reviewed open-access journal
MycoKeys
Launched to accelerate biodiversity research

Top 50 most wanted fungi

R. Henrik Nilsson[1], Christian Wurzbacher[1], Mohammad Bahram[2,3], Victor R. M. Coimbra[1,4], Ellen Larsson[1], Leho Tedersoo[3], Jonna Eriksson[1], Camila Duarte Ritter[1], Sten Svantesson[1], Marisol Sánchez-García[5], Martin Ryberg[2], Erik Kristiansson[6], Kessy Abarenkov[7]

1 *Department of Biological and Environmental Sciences, University of Gothenburg, Box 463, 405 30 Göteborg, Sweden* **2** *Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Sweden* **3** *Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia* **4** *Departamento de Micologia, Centro de Ciências Biológicas (CCB), Universidade Federal de Pernambuco (UFPE), Av. Prof. Nelson Chaves, s/n, 50760-901 Recife, Pernambuco, Brazil* **5** *Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, 37996-1610, USA* **6** *Department of Mathematical Sciences, Chalmers University of Technology/University of Gothenburg, 412 96, Göteborg, Sweden* **7** *Natural History Museum, University of Tartu, Vanemuise 46, Tartu 510 14, Estonia*

Corresponding author: *R. Henrik Nilsson* (henrik.nilsson@bioenv.gu.se)

Academic editor: *T. Lumbsch* | Received 18 December 2015 | Accepted 25 January 2016 | Published 18 March 2016

Citation: Nilsson RH, Wurzbacher C, Bahram M, Coimbra VRM, Larsson E, Tedersoo L, Eriksson J, Duarte Ritter C, Svantesson S, Sánchez-García M, Ryberg M, Kristiansson E, Abarenkov K (2016) Top 50 most wanted fungi. MycoKeys 12: 29–40. doi: 10.3897/mycokeys.12.7553

Abstract

Environmental sequencing regularly recovers fungi that cannot be classified to any meaningful taxonomic level beyond "Fungi". There are several examples where evidence of such lineages has been sitting in public sequence databases for up to ten years before receiving scientific attention and formal recognition. In order to highlight these unidentified lineages for taxonomic scrutiny, a search function is presented that produces updated lists of approximately genus-level clusters of fungal ITS sequences that remain unidentified at the phylum, class, and order levels, respectively. The search function (https://unite.ut.ee/top50.php) is implemented in the UNITE database for molecular identification of fungi, such that the underlying sequences and fungal lineages are open to third-party annotation. We invite researchers to examine these enigmatic fungal lineages in the hope that their taxonomic resolution will not have to wait another ten years or more.

Key words

Fungi, environmental sequencing, taxonomic orphans, metabarcoding, taxonomy feedback loop

Introduction

Fungi form a large and diverse kingdom of heterotrophic eukaryotes. Recent studies suggest that there may be more than 6 million extant species of fungi (Taylor et al. 2014), a number that contrasts sharply with the ca. 100,000 formally described species (Hibbett et al. 2011). Several factors contribute to the discrepancy between the estimated and the known number of fungal species. In particular, the subterranean or otherwise difficult-to-observe nature of much of fungal life sets mycology apart from the study of many other groups of multicellular eukaryotes (Blackwell 2011). Molecular (DNA sequence) data have revolutionized the scientific study of fungi, and DNA sequence data are now a routine source of information in fungal systematics, taxonomy, and ecology across the fungal tree of life (Stajich et al. 2009). Fungal environmental sequencing (molecular ecology) studies, where some particular environmental habitat or substrate is examined for fungal diversity, span these disciplines in seeking to detail what fungi are present and what their ecological and functional roles are in the system studied.

Molecular ecology studies regularly struggle to identify the recovered fungi to meaningful taxonomic levels. Lack of reference sequences, mis-annotated reference sequences, and reference sequences annotated only to, e.g., kingdom or phylum level combine to make taxonomic identification of newly recovered sequence data challenging (Nilsson et al. 2012). These issues are to some extent mitigated by initiatives such as the UNITE database for molecular identification of fungi (Kõljalg et al. 2013), but they remain significant challenges to any molecular ecology effort. In particular, environmental sequencing studies regularly recover fungal sequences that are difficult to assign to any fungal lineage at all, even at the phylum level. The discovery and subsequent description of the class *Archaeorhizomycetes* (Schadt et al. 2003; Rosling et al. 2011) and the phylum *Cryptomycota* (Lara et al. 2010; Jones et al. 2011) both involve environmental samples that initially could not be assigned to any resolved taxonomic level. Similarly, the global soil sequencing study of Tedersoo et al. (2014) recovered 16 large groups of fungal sequences that could not be classified to any meaningful taxonomic level beyond Fungi. Indeed, more or less all environmental sequencing studies feature a non-trivial proportion of sequences simply classified as "Unidentified fungi" (cf. Hardoim et al. 2015) due to the lack of more explicit taxonomic information. There is no taxonomic feedback loop in place to highlight the presence of these enigmatic lineages to the mycological community, and they often end up in sequence databases for years without attracting significant research interest.

In our work with environmental sequencing of fungi, we regularly run across these unidentified lineages. We typically encounter them through sequences of the internal transcribed spacer (ITS), the formal fungal barcode (Schoch et al. 2012) and the marker of choice in fungal molecular ecology studies (Lindahl et al. 2013; Tedersoo et al. 2015). A quick BLAST search in the International Nucleotide Sequence Database Collaboration (INSDC: GenBank, ENA, and DDBJ; Nakamura et al. 2013) or UNITE typically hints at the impossibility of coming up with any resolved taxonomic affiliation, and the matter is left at that. This situation is untenable in the long run. These lineages will give rise to identification problems for other research groups too, such that a limited number of taxonomic orphans will affect the scientific results of a large number of research efforts negatively. This is not in the best interest of mycology. In the present study we seek to bridge the gap between fungal taxonomy and molecular ecology by putting the spotlight on the 50 largest of these unidentified lineages at the phylum, class, and order levels. Our effort takes the form of an automatically updated search function targeting the largest taxonomic orphans in the UNITE database. The lists of the largest orphans and the constituent sequences are subject to third-party sequence annotation, such that anyone who has information on these species is invited to share it with the scientific community. The lists are updated monthly, and by highlighting these fungal lineages we hope to speed up their characterization and formal description.

Materials and methods

UNITE clusters all public fungal ITS sequences (~500,000 at the time of this writing) to approximately the genus/subgenus level (called a "compound cluster") using a clustering threshold of 80% sequence similarity. A second round of clustering inside each such compound cluster seeks to produce molecular operational taxonomic units (OTUs) at approximately the species level; these OTUs are called *species hypotheses* (SHs; Kõljalg et al. 2013). The species hypotheses are open for viewing and querying (http://unite.ut.ee/search.php) through uniform resource identifiers (URIs) such as https://plutof.ut.ee/#/datacite/10.15156/BIO/SH154595.07FU. Each SH has a unique digital object identifier (DOI, 10.15156/BIO/SH154595.07FU for the example above) to enable precise species-level taxonomic communication across publications and studies also in the absence of precise Latin names.

Although UNITE offers various search functions targeting the compound clusters and species hypotheses, none of the search functions were designed to find truly poorly known lineages. To remedy this, we devised a search function to retrieve fungal lineages for which little to no taxonomic information is available. The user is presented with two main choices: 1) the taxonomic level to be considered (phylum, class, or order), and 2) whether the list of compound clusters should be ordered by the number of constituent sequences or by the number of studies in which the sequences were found. In addition, the user can exercise control over how the output is shown through several other options.

Taxonomic scope (phylum, class, or order)

To enable exploration of different hierarchical levels in the classification system, the search function supports three different levels: phylum, class, and order. Thus, the search function will retrieve clusters of sequences where none of the sequences are identified at the phylum, class, or order level depending on the choice of the user.

Sorting of the list of taxa (sequence or study count)

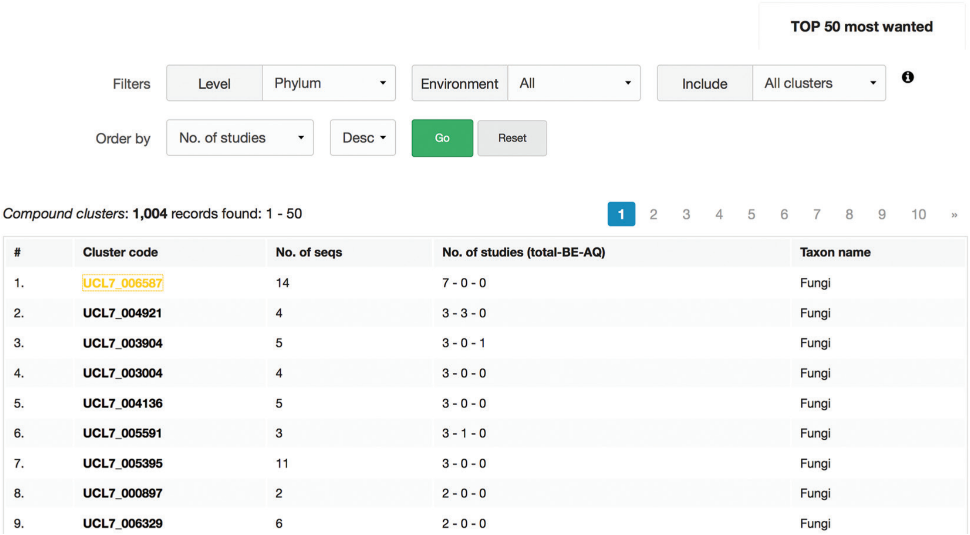
Multiple independent recoveries of some particular fungal sequence type would strengthen one's belief that the lineage indeed corresponds to a biological reality. In analogy, for sequence types found only in a single study, some sound skepticism is perhaps in place given the sequence quality-related issues involved in studies based on cloning as well as next-generation sequencing (Hyde et al. 2013; Lindahl et al. 2013; Hughes et al. 2015). However, there are examples to the contrary for both of these situations: sequence types found only in one particular study have proved to be authentic, and "species" found in several different studies have proved to be chimeras (Brown et al. 2015; Nilsson et al. 2015). This search parameter offers some degree of flexibility by allowing the user to specify whether the number of sequences or the number of studies should be used to order the list of compound clusters.

Each search will retrieve all clusters of sequences fulfilling the criteria. Thus, there are 3 (phylum, class, and order) * 2 (order by sequences or by studies) = 6 lists of "poorly known" fungal lineages. Some degree of overlap among these lists is likely; a compound cluster where all sequences are unidentified at the order level may also qualify as a cluster where all of the sequences are unidentified at the phylum level. No attempt was made to account for such redundancy.

A concern was that these sequences could be subject to quality issues. Alternatively they could be false positives in that they lacked explicit taxonomic annotation but nevertheless were easy to assign to a known taxonomic lineage. To minimize these concerns, we examined the 50 largest lineages at the phylum, class, and order levels (as ordered by the number of constituent studies) through BLAST searches in UNITE and the INSDC following Kang et al. (2010) and Nilsson et al. (2012). The full length of the sequences as they were deposited in INSDC/UNITE, as well as the ITS2 and 5.8S separately, were used for these searches. Many of the sequences were annotated to the barest minimum and lacked, for example, metadata on country and substrate of collection. In an attempt at restoring as much of these data as possible, we examined the underlying papers when specified in the corresponding INSDC entries.

Results

The phylum-level search returned 1,004 compound clusters, of which 830 (83%) were singletons. Out of the 1,364 class-level clusters, 1,056 (77%) were singletons; and out of the 1,738 order-level clusters, 1,290 (74%) were singletons. The results presented here focus on the 50 topmost entries in each of these lists. The largest of the phylum-level clusters comprised 30 sequences, and the average number of sequences in the 50 topmost clusters was 7.4 (standard deviation: 4.9). At the class level, the largest cluster comprised 60 sequences (average cluster size 8.5 sequences, standard deviation 9.7). At the order level, the largest cluster comprised 60 sequences (average cluster size 9.5 sequences, standard deviation 9.5). The cluster with the highest



TOP 50 most wanted

Filters: Level: Phylum; Environment: All; Include: All clusters

Order by: No. of studies; Desc

Compound clusters: **1,004** records found: 1 - 50

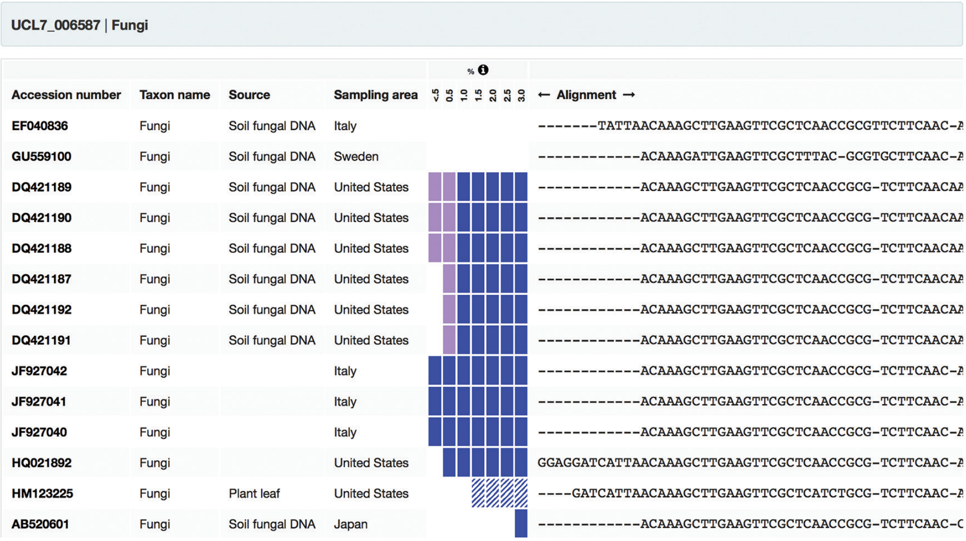
#	Cluster code	No. of seqs	No. of studies (total-BE-AQ)	Taxon name
1.	UCL7_006587	14	7 - 0 - 0	Fungi
2.	**UCL7_004921**	4	3 - 3 - 0	Fungi
3.	**UCL7_003904**	5	3 - 0 - 1	Fungi
4.	**UCL7_003004**	4	3 - 0 - 0	Fungi
5.	**UCL7_004136**	5	3 - 0 - 0	Fungi
6.	**UCL7_005591**	3	3 - 1 - 0	Fungi
7.	**UCL7_005395**	11	3 - 0 - 0	Fungi
8.	**UCL7_000897**	2	2 - 0 - 0	Fungi
9.	**UCL7_006329**	6	2 - 0 - 0	Fungi

Figure 1. A web-based screenshot of the upper part of the top 50 list of compound clusters where all sequences are unidentified at the phylum level. The clusters are ordered by the number of contributing studies in this screenshot.

number of independent recoveries had been found in 23 different studies and was unidentified at the order level.

The lists, with accompanying multiple sequence alignments and geo/ecological metadata, are available for viewing and third-party annotation at https://unite.ut.ee/top50.php (Figs 1–3 from September 2015). Our taxonomic examination of the lineages at the compound cluster level was unsuccessful – we could not assign any of the lineages to any known fungal lineage with confidence. For some lineages, there were hints or clues pointing to a tentative assignment of the sequences to phylum or class level, but the disparate or heterogeneous nature of the available reference sequences did not lend confidence to any robust assignment. In line with the UNITE policy, no speculative (non-robust) assignments were made in these lineages. In other cases, the publicly available reference sequences offered absolutely no guidance as to the taxonomic affiliation of the query sequences (e.g., “Uncultured eukaryote”). In 39 cases, we found the sequences to be associated with quality-related problems, mainly a chimeric nature (cf. Nilsson et al. 2012). We marked those sequences as substandard/chimeric and re-ran the search function to make sure that none of the top 50 clusters in the compound cluster list would be obvious cases of compromised sequence data as of the date of the preparation of this paper.

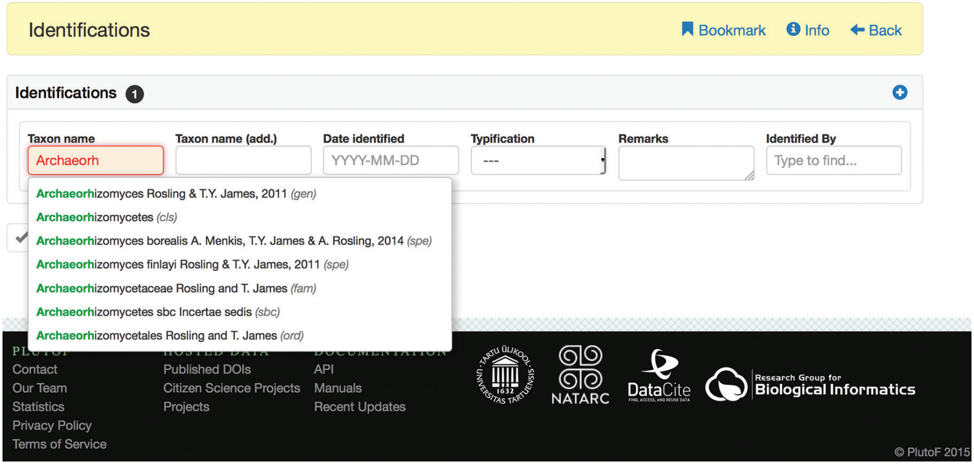
Our data assembly effort to restore data on the country and host of collection resulted in 60 sequences being tagged with a country of collection and 261 with a substrate of collection. Data on country and substrate of collection for the 50 largest compound clusters that were not identified at the phylum, class, and order level, respectively, are shown in Figs 4–5 (September 2015). Soil, living plants, and mycor-



UCL7_006587 | **Fungi**

				% ℹ							
Accession number	**Taxon name**	**Source**	**Sampling area**	**<.5**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**← Alignment →**
EF040836	Fungi	Soil fungal DNA	Italy								-------TATTAACAAAGCTTGAAGTTCGCTCAACCGCGTTCTTCAAC-A
GU559100	Fungi	Soil fungal DNA	Sweden								------------ACAAAGATTGAAGTTCGCTTTAC-GCGTGCTTCAAC-A
DQ421189	Fungi	Soil fungal DNA	United States	🟪	🟪	🟦	🟦	🟦	🟦	🟦	------------ACAAAGCTTGAAGTTCGCTCAACCGCG-TCTTCAACAA
DQ421190	Fungi	Soil fungal DNA	United States	🟪	🟪	🟦	🟦	🟦	🟦	🟦	------------ACAAAGCTTGAAGTTCGCTCAACCGCG-TCTTCAACAA
DQ421188	Fungi	Soil fungal DNA	United States	🟪	🟪	🟦	🟦	🟦	🟦	🟦	------------ACAAAGCTTGAAGTTCGCTCAACCGCG-TCTTCAACAA
DQ421187	Fungi	Soil fungal DNA	United States		🟪	🟦	🟦	🟦	🟦	🟦	------------ACAAAGCTTGAAGTTCGCTCAACCGCG-TCTTCAACAA
DQ421192	Fungi	Soil fungal DNA	United States		🟪	🟦	🟦	🟦	🟦	🟦	------------ACAAAGCTTGAAGTTCGCTCAACCGCG-TCTTCAACAA
DQ421191	Fungi	Soil fungal DNA	United States		🟪	🟦	🟦	🟦	🟦	🟦	------------ACAAAGCTTGAAGTTCGCTCAACCGCG-TCTTCAACAA
JF927042	Fungi		Italy	🟦	🟦	🟦	🟦	🟦	🟦	🟦	------------ACAAAGCTTGAAGTTCGCTCAACCGCG-TCTTCAAC-A
JF927041	Fungi		Italy	🟦	🟦	🟦	🟦	🟦	🟦	🟦	------------ACAAAGCTTGAAGTTCGCTCAACCGCG-TCTTCAAC-A
JF927040	Fungi		Italy	🟦	🟦	🟦	🟦	🟦	🟦	🟦	------------ACAAAGCTTGAAGTTCGCTCAACCGCG-TCTTCAAC-A
HQ021892	Fungi		United States		🟦	🟦	🟦	🟦	🟦	🟦	GGAGGATCATTAACAAAGCTTGAAGTTCGCTCAACCGCG-TCTTCAAC-A
HM123225	Fungi	Plant leaf	United States				▨	▨	▨	▨	----GATCATTAACAAAGCTTGAAGTTCGCTCATCTGCG-TCTTCAAC-A
AB520601	Fungi	Soil fungal DNA	Japan							🟦	------------ACAAAGCTTGAAGTTCGCTCAACCGCG-TCTTCAAC-C

Figure 2. A compound cluster displayed in the web browser of the user. The INSDC accession numbers and their taxonomic annotation are shown in columns 1 and 2. The DNA source and the country of collection are shown in columns 3 and 4. Column 5 shows the inclusiveness of the species hypotheses at the 97% similarity level (rightmost filled column), the 97.5% similarity level (second-to-rightmost filled column), and so on up to 100% similarity. The aligned sequence data are shown in column 6.



Identifications 1

Taxon name	Taxon name (add.)	Date identified	Typification	Remarks	Identified By
Archaeorh		YYYY-MM-DD	---		Type to find...

Archaeorhizomyces Rosling & T.Y. James, 2011 *(gen)*
Archaeorhizomycetes *(cls)*
Archaeorhizomyces borealis A. Menkis, T.Y. James & A. Rosling, 2014 *(spe)*
Archaeorhizomyces finlayi Rosling & T.Y. James, 2011 *(spe)*
Archaeorhizomycetaceae Rosling and T. James *(fam)*
Archaeorhizomycetes sbc Incertae sedis *(sbc)*
Archaeorhizomycetales Rosling and T. James *(ord)*

Figure 3. Web-based third-party taxonomic annotation of the sequences in a species hypothesis is demonstrated. Third-party annotation requires non-anonymous registration, and such annotations are subject to peer review. Annotations are tagged with the name of the annotator as well as the date. Multiple annotations for individual entries are supported.

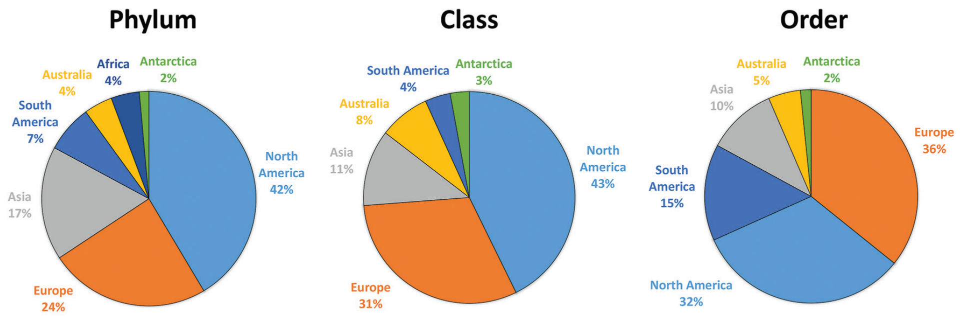


Figure 4. Geographical distribution of the top 50 most wanted fungi at the phylum, class, and order level. Each fungal sequence was assigned to country of origin according to its INSDC entry (or underlying publication as applicable) and then summarized based on the continents: Africa (dark blue), Antarctica (green), Asia (grey), Australia (yellow), Europe (orange), North America (light blue), and South America (blue).

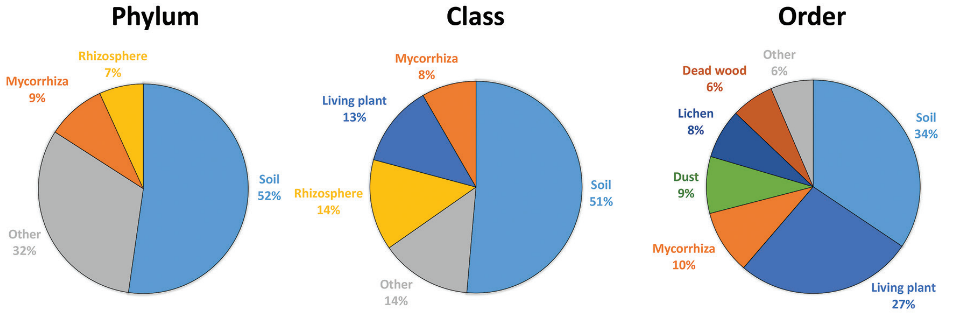


Figure 5. The most common substrates associated with the top 50 most wanted fungi at the phylum, class, and order level. Each fungal sequence was assigned to substrate according to its GenBank entry (or underlying publication as applicable). The major substrates included soil (light blue), living plants (blue), mycorrhiza (orange), dust (green), lichen (dark blue), dead wood (red), and other (grey). To improve readability, rare substrates (<3 occurrences) were merged into the 'other' category.

rhiza stand out as frequently sampled substrates. Europe and North America stand out as frequent targets for environmental sequencing studies. These are well-known biases towards the most commonly targeted molecular ecology substrates and the Western world, respectively (Ryberg et al. 2009; Tedersoo et al. 2011; Lindahl et al. 2013), and should not necessarily be taken to mean that fungal diversity is the highest in these substrates and geographical locations. Along the same line, it is pleasing to note that all seven continents are represented in Figure 4, hinting perhaps at the increasingly ambitious sampling efforts undertaken by the mycological and molecular ecology communities. Somewhat unexpectedly, perhaps, dust and lichens seem to be relatively rich sources of sequences and species hypotheses that cannot be identified at the order level.

Discussion

This paper presents a set of lists of fungi for which taxonomic assignment is very troublesome at present. These lists matter, because the underlying fungi are regularly recovered in environmental sequencing efforts, where they contribute to the proportion of unidentified sequences. Mycology is a comparatively small discipline that struggles for funding (cf. Pautasso 2013), and it would be beneficial for mycology to show that when researchers sequence fungi as a part of their scientific pursuits, they get clean, unequivocal results. That is not the case at present. Worse, the taxonomic discovery potential of environmental sequencing is not made full use of by the mycological community. History shows that evidence of unknown lineages of fungi may sit in sequence databases for upwards of 10 years before receiving scientific attention and formal recognition. Indeed, several of the present lineages feature sequence data that are at least that old. We hope that these lists – largely consisting of sequences from environmental sequencing efforts – will establish a feedback loop back to taxonomy. We furthermore hope that anyone who has information that sheds light on the taxonomic affiliation of these lineages would be willing to share this information with the research community through the third-party sequence annotation tools of UNITE (or otherwise). Even phylum-level annotations, as applicable, would help. UNITE serves as data provider for a range of sequence identification pipelines and databases (Bates et al. 2013; https://unite.ut.ee/repository.php), and any such contributed taxonomic information would be shared with all downstream resources.

We examined all sequence types from the 50 largest compound clusters for telltale signs of a technically compromised nature, such as chimeric insertions or low read quality (cf. Nilsson et al. 2012, 2015). In this process we found and excluded 39 substandard sequences, after which the search was re-run. We could not assert with confidence that any of the remaining lineages were technically compromised. However, such examinations should ideally be carried out in light of other sequences from closely related lineages, of which none or very few are available for these lineages. Our sequence quality control was, therefore, not carried out under optimal conditions. Even so, all sequences passed the quality measures we exercised. Importantly, none of the lineages examined were singletons – on the contrary, the largest one comprised 60 sequences, and most were recovered in two or more different studies (with 23 being the largest number of studies). Although independent recovery of some particular sequence type does not rule out, e.g., a chimeric nature, it does increase the likelihood that the sequence is genuine.

It is not immediately clear that all of these lineages indeed are fungi, although at least one fungus-specific primer seems to have been involved in the generation of many of them. Many studies have reported the occasional (even frequent) co-amplification of, e.g., plants and metazoans with fungus-specific primers (cf. Tedersoo et al. 2011; 2014). We are certainly open to the possibility that one or more of the present lineages will prove to be non-fungal organisms in the end. Since they evidently are prone to co-amplification with fungus-specific primers or otherwise are retrieved in research efforts targeting fungi, it would seem important to be able to tell them from fungi in the sequence identification step. Getting the naming of these sequences right, even if they are not fungal, would thus still appear to be of relevance to mycology.

Precise and robust taxonomic assignment of these ITS sequences is not possible at present due to the lack of similar reference sequences in the public sequence databases. Sequence data from the much more conserved, neighboring small and large subunit genes (18S/SSU and 28S/LSU, respectively) would presumably have alleviated this problem by allowing phylogenetic placement in the context of known SSU and LSU sequences. However, ITS sequences are typically sequenced and deposited without significant parts of the SSU and LSU, particularly in environmental sequencing efforts, rendering this approach difficult. Deeply sequenced metagenomes – as well as emerging sequencing technologies producing very long reads – offer a route by which to retrieve parts of the ITS region attached to either the SSU or LSU, or indeed span them both. Thus, the increasing popularity of metagenomics and genomics may solve many of these cases over time. However, also someone doing traditional systematics and taxonomy can contribute. Supplying, as a minimum, an ITS sequence with each new species description would offer structure to available sequence data and would significantly reduce interpretation difficulties of species names (Hyde et al. 2008). Similarly, GenBank is known to contain thousands of sequences from type material – sequences that are not annotated as stemming from type material at present. GenBank has recently implemented standards for marking and querying sequences from type material (Schoch et al. 2014), and we hope that the mycological community will be quick to embrace these standards for newly generated as well as already deposited sequences. Another helpful move would be to provide an ITS sequence with each new fungal genome. For technical reasons, ITS and other ribosomal sequences tend to be hard to assemble and are therefore left out from many genome sequencing efforts (Schoch et al. 2014).

We are working to add additional flexibility in the generation of these lists. Some researchers may, for example, be interested only in unknown fungi found in the built environment, or in a medical context, or from aquatic environments. We will seek to address these needs by compiling a set of keywords for each such research field. For the built environment, these keywords would include, e.g., "house", "dust", "building", and "gypsum". For the search function, we will then require that a compound cluster contains at least one sequence where at least one of these keywords occurs either in the title of the underlying scientific study or in the FEATURES field of the corresponding INSDC/UNITE entry. The search function would then retrieve compound clusters with at least one fungal sequence that has a relation to the built environment. We will similarly endeavor to add support for the genus and species levels in the search function.

We refer to this list as the "most wanted" fungi. That is not meant to suggest that these fungi are the ecologically or economically most important extant fungi. Indeed, we make no claim as to the importance of these fungi from whatever point of view. We do make a claim to their uniqueness though, because it is frustrating, in the year 2016, not to be able to assign a name to a fungal sequence even at the phylum level.

When it comes to taxonomic discovery potential, we argue that these lineages definitely should be counted among the most interesting candidates. Even if we assume that some proportion of the present lineages in fact are technical artifacts or represent non-fungal organisms, it is reasonable to assume that some proportion of them indeed represent new or previously unsequenced lineages of fungi. None of them are at least 80% similar to sequences with richer taxonomic annotations; many are much more distant from known reference sequences than that. Common rules of thumb for ITS sequence similarity thresholds (Schoch et al. 2012, 2014; Irinyi et al. 2014) suggest that these lineages each represent at least a new (or previously unsequenced) genus, and in some cases an order or potentially even higher. We hope that the present publication will serve to put the spotlight on these uncharted parts of the fungal tree of life, and we invite the reader to examine them through our online tools or otherwise. These lists of the most wanted fungi are recomputed automatically on a monthly basis. We hope that they will speed up the formal recognition of the underlying species, and we challenge users to try to identify these species – because we failed ourselves. Until formal scientific names are available for these species, UNITE provides DOIs to promote unambiguous communication, and data harvesting, across datasets and studies.

Acknowledgements

RHN acknowledges financial support from the Swedish Research Council of Environment, Agricultural Sciences, and Spatial Planning (FORMAS, 215-2011-498). Support from the Sloan Foundation is gratefully acknowledged. CW acknowledges a Marie Skłodowska-Curie post doc grant (660122, CRYPTRANS).

References

Bates ST, Ahrendt S, Bik HM et al. (2013) Meeting Report: Fungal ITS Workshop (October 2012). Standards in Genomics Sciences 8(1): 118–123. doi: 10.4056/sigs.3737409

Blackwell M (2011) The Fungi: 1, 2, 3 … 5.1 million species? American Journal of Botany 98(3): 426–438. doi: 10.3732/ajb.1000298

Brown SP, Veach AM, Rigdon-Huss AR, Grond K, Lickteig SK, Lothamer K, Oliver AK, Jumpponen A (2015) Scraping the bottom of the barrel: are rare high throughput sequences artifacts? Fungal Ecology 13: 221–225. doi: 10.1016/j.funeco.2014.08.006

Hardoim PR, van Overbeek LS, Berg G, Pirttilä AM, Compant S, Campisano A, Döring M, Sessitsch A (2015) The hidden world within plants: ecological and evolutionary considerations for defining functioning of microbial endophytes. Microbiology and Molecular Biology Reviews 79(3): 293–320. doi: 10.1128/MMBR.00050-14

Hibbett DS, Ohman A, Glotzer D, Nuhn M, Kirk P, Nilsson RH (2011) Progress in molecular and morphological taxon discovery in Fungi and options for formal classification of environmental sequences. Fungal Biology Reviews 25(1): 38–47. doi: 10.1016/j.fbr.2011.01.001

Hughes KW, Morris SD, Segovia AR (2015) Cloning of ribosomal ITS PCR products creates frequent, non-random chimeric sequences – a test involving heterozygotes between *Gymnopus dichrous* taxa I and II. MycoKeys 10: 45–56. doi: 10.3897/mycokeys.10.5126

Hyde KD, Zhang Y (2008) Epitypification: should we epitypify? Journal of Zhejiang University Science B 9: 842–846. doi: 10.1631/jzus.B0860004

Hyde KD, Udayanga D, Manamgoda DS et al. (2013) Incorporating molecular data in fungal systematics: a guide for aspiring researchers. Current Research in Environmental and Applied Mycology 3(1): 1–32. doi: 10.5943/cream/3/1/1

Irinyi L, Serena C, Garcia-Hermoso D et al. (2015) International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database—the quality controlled standard tool for routine identification of human and animal pathogenic fungi. Medical Mycology: myv008. doi: 10.1093/mmy/myv008

Jones MD, Forn I, Gadelha C, Egan MJ, Bass D, Massana R, Richards TA (2011) Discovery of novel intermediate forms redefines the fungal tree of life. Nature 474(7350): 200–203. doi: 10.1038/nature09984

Kang S, Mansfield MAM, Park B et al. (2010) The promise and pitfalls of sequence-based identification of plant pathogenic fungi and oomycetes. Phytopathology 100(8): 732–737. doi: 10.1094/PHYTO-100-8-0732

Kõljalg U, Nilsson RH, Abarenkov K et al. (2013) Towards a unified paradigm for sequence-based identification of Fungi. Molecular Ecology 22(21): 5271–5277. doi: 10.1111/mec.12481

Lara E, Moreira D, López-García P (2010) The environmental clade LKM11 and *Rozella* form the deepest branching clade of fungi. Protist 161(1): 116–121. doi: 10.1016/j.protis.2009.06.005

Lindahl BD, Nilsson RH, Tedersoo L et al. (2013) Fungal community analysis by high-throughput sequencing of amplified markers - a user's guide. New Phytologist 199(1): 288–299. doi: 10.1111/nph.12243

Nakamura Y, Cochrane G, Karsch-Mizrachi I (2013) The international nucleotide sequence database collaboration. Nucleic Acids Research 41(D1): D21–D24. doi: 10.1093/nar/gks1084

Nilsson RH, Tedersoo L, Abarenkov K et al. (2012) Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. MycoKeys 4: 37–63. doi: 10.3897/mycokeys.4.3606

Nilsson RH, Tedersoo L, Ryberg M et al. (2015) A comprehensive, automatically updated fungal ITS sequence dataset for reference-based chimera control in environmental sequencing efforts. Microbes and Environments 30(2): 145–150. doi: 10.1264/jsme2.ME14121

Pautasso M (2013) Fungal under-representation is (indeed) diminishing in the life sciences. Fungal Ecology 6(5): 460–463. doi: 10.1016/j.funeco.2013.03.001

Rosling A, Cox F, Cruz-Martinez K, Ihrmark K, Grelet G-A, Lindahl BD, Menkis A, James TY (2011) *Archaeorhizomycetes*: Unearthing an ancient class of ubiquitous soil fungi. Science 333(6044): 876–879. doi: 10.1126/science.1206958

Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH (2009) An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity. New Phytologist 181(2): 471–477. doi: 10.1111/j.1469-8137.2008.02667.x

Schadt CW, Martin AP, Lipson DA, Schmidt SK (2003) Seasonal dynamics of previously unknown fungal lineages in tundra soils. Science 301(5638): 1359–1361. doi: 10.1126/science.1086940

Schoch CL, Seifert KA, Huhndorf S et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences USA 109(16): 6241–6246. doi: 10.1073/pnas.1117018109

Schoch CL, Robbertse B, Robert V et al. (2014) Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. Database (Oxford). doi: 10.1093/database/bau061

Stajich JE, Berbee ML, Blackwell M, Hibbett DS, James TY, Spatafora JW, Taylor JW (2009) The fungi. Current Biology 19(18): R840–R845. doi: 10.1016/j.cub.2009.07.004

Taylor DL, Hollingsworth TN, McFarland JW, Lennon NJ, Nusbaum C, Ruess RW (2014) A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. Ecological Monographs 84(1): 3–20. doi: 10.1890/12-1693.1

Tedersoo L, Abarenkov K, Nilsson RH et al. (2011) Tidying up International Nucleotide Sequence Databases: ecological, geographical, and sequence quality annotation of ITS sequences of mycorrhizal fungi. PLoS ONE 6: e24940. doi: 10.1371/journal.pone.0024940

Tedersoo L, Bahram M, Põlme S et al. (2014) Global diversity and geography of soil fungi. Science 346: 6213. doi: 10.1126/science.1256688

Tedersoo L, Anslan S, Bahram M et al. (2015) Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. MycoKeys 10: 1–43. doi: 10.3897/mycokeys.10.4852