ELSEVIER

British Mycological Society promoting fungal science

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

Plenary Paper

Progress in molecular and morphological taxon discovery in *Fungi* and options for formal classification of environmental sequences

David S. HIBBETT[a,*], *Anders OHMAN*[a], *Dylan GLOTZER*[a], *Mitchell NUHN*[a], *Paul KIRK*[b], *R. Henrik NILSSON*[c,d]

[a]*Biology Department, Clark University, Worcester, MA 01610, USA*
[b]*CABI UK, Bakeham Lane, Egham, Surrey TW20 9TY, UK*
[c]*Department of Plant and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden*
[d]*Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai St., 51005 Tartu, Estonia*

ARTICLE INFO

Article history:
Received 14 December 2010
Accepted 3 January 2011

Keywords:
Biodiversity
Classification
Environmental sequences
Molecular ecology
MOTU
Taxonomy

ABSTRACT

Fungal taxonomy seeks to discover, describe, and classify all species of *Fungi* and provide tools for their identification. About 100,000 fungal species have been described so far, but it has been estimated that there may be from 1.5 to 5.1 million extant fungal species. Over the last decade, about 1200 new species of *Fungi* have been described in each year. At that rate, it may take up to 4000 y to describe all species of *Fungi* using current specimen-based approaches. At the same time, the number of molecular operational taxonomic units (MOTUs) discovered in ecological surveys has been increasing dramatically. We analyzed ribosomal RNA internal transcribed spacer (ITS) sequences in the GenBank nucleotide database and classified them as "environmental" or "specimen-based". We obtained 91,225 sequences, of which 30,217 (33 %) were of environmental origin. Clustering at an average 93 % identity in extracted ITS1 and ITS2 sequences yielded 16,969 clusters, including 6230 (37 %) clusters with only environmental sequences, and 2223 (13 %) clusters with both environmental and specimen-based sequences. In 2008 and 2009, the number of purely environmental clusters deposited in GenBank exceeded the number of species described based on specimens, and this does not include the huge number of unnamed MOTUs discovered in pyrosequencing studies. To enable communication about fungal diversity, there is a pressing need to develop classification systems based on environmental sequences. Assigning Latin binomials to MOTUs would promote their integration with specimen-based taxonomic databases, whereas the use of numerical codes for MOTUs would perpetuate a disconnect with the taxonomic literature. MOTUs could be formally named under the existing *International Code of Botanical Nomenclature* if the concept of a nomenclatural type was expanded to include environmental samples or illustrations of sequence chromatograms (or alignments). Alternatively, a "candidate species" category could be created for *Fungi*, based on the *candidatus* taxon status employed by microbiologists.

* *Corresponding author*. Tel.: +1 508 793 7332; fax: +1 508 793 7174.
E-mail address: dhibbett@clarku.edu (D. S. Hibbett).
1749-4613/$ – see front matter
doi:10.1016/j.fbr.2011.01.001

1. Introduction

The goals of fungal taxonomy are to discover and describe all species of *Fungi*, to classify them according to their phylogenetic relationships, and to provide tools for their identification. The names produced by fungal taxonomists enable discourse in all of the human endeavors that concern *Fungi*, and the hypotheses of relationships embodied in taxonomic names communicate our understanding of the history of life. Here, we address recent progress in fungal taxonomy, with a focus on species-level classification, and we consider options for accelerating this important work.

2. Progress in morphological species description

Annual progress in fungal species description is recorded in *Index Fungorum*, which is the product of an ongoing survey of the world's mycological literature. From 1999 to 2009, an average of 1196 new species of *Fungi* have been described each year, with 1030 species described in 2009 (Fig. 1; a table showing the number of species described per year is available as Supplementary material). This steady progress attests to the dedication and perseverance of the fungal taxonomic community, which has so far cataloged about 97,861 species of *Fungi*, according to the most recent *Dictionary of the Fungi* (Kirk *et al.*, 2008). Of course, it is hard to say how close we are to completion of the global inventory, because the number of extant fungal species is not known (Mueller and Schmit, 2007). Hawksworth's (2001) estimate of 1.5 million species is the most frequently cited figure, but O'Brien *et al.* (2005) suggested that there may be as many as 3.5–5.1 million species, while Schmit and Mueller (2007) suggested a conservative minimum estimate of 712,000 species. Based on Hawksworth's estimate, at the current rate of species description it will take 1170 y to complete the global fungal inventory; O'Brien *et al*'s. estimates imply that 2840–4170 y will be needed.

To approach a complete catalog of fungal diversity within a reasonable time frame, it will be necessary to dramatically accelerate the pace of species description. Traditional morphology-based taxonomy will be critical to this effort, but the massive increase in the number of active taxonomists that would be needed to achieve the goal in the near term seems highly unlikely. In fact, the number of active taxonomists of traditional training may be declining (Lücking, 2008). Fortunately, a new methodology for taxon discovery has emerged in recent years through the work of fungal molecular ecologists. The remainder of this article considers the prospects for sequence-based taxon discovery and description through analysis of environmental sequences.

3. Progress in species discovery through environmental sequences

Molecular operational taxonomic units (MOTUs) potentially representing novel taxa are routinely discovered in fungal molecular ecology studies. It is difficult to track the discovery of such unnamed entities, however, because there is no centralized database that compiles these records in the manner that *Index Fungorum* includes details of new taxa named in a traditional way. Moreover, there are no universally accepted standard criteria for MOTU delimitation based on sequences. A cut-off of around 97 % identity in the internal transcribed spacer (ITS) region of the nuclear ribosomal RNA (rRNA) genes is frequently invoked for MOTU delimitation, but there is variation in the exact cut-off value employed, as well as in the inclusion or exclusion of sequences encoding 5.8S rRNA (and the flanking 18S and 25S regions) in similarity calculations (Hughes *et al.*, 2009; Nilsson *et al.*, 2009b). Moreover, there is general agreement that "species" should not be recognized solely on the basis of sequence similarity insofar as additional relevant characters are available or could be retrieved. Nonetheless, quantitative community comparisons require assignment of organisms into taxonomic units, so the majority of fungal molecular ecology studies employ some form of MOTU, with the tacit implication that these are roughly comparable to species (Peay *et al.*, 2008).

To assess the rates of ITS sequence deposition and discovery of MOTUs, we created a series of Perl scripts that we used to analyze complete ITS1-5.8S-ITS2 sequences in the GenBank (Benson *et al.*, 2009) nucleotide database (all scripts and data are available as Supplementary material and at http://www.clarku.edu/faculty/dhibbett/resources.html). We did not examine data from studies using pyrosequencing methods, which produce massive volumes of data and that, while transforming fungal molecular ecology, would skew the present effort. The analyses presented here, based solely on data from the Sanger sequencing method, provide a baseline comparison of the relative contributions of fungal molecular ecology and fungal taxonomy to the sequence database through 2009.

In February 2010, we downloaded all reasonably complete fungal ITS sequences, screened them according to length (400–1500 bp) and quality (sequences with no more than 3 % [IUPAC] DNA ambiguity symbols were allowed), and classified them as "specimen-based" or "environmental" based on a set of key terms in the Locus, Features, and Organism fields (e.g., ENV, unidentified, ectomycorrhiz(a/ae/al), endophyt(e/ic), environmental, root tip, unclassified, uncultur(ed/able), unidentified, and uncultured). Manual inspection of the parsed sequences indicated that most "environmental" sequences were obtained directly from soil, roots, or other mixed natural substrates by polymerase chain reaction (PCR) amplification, but some are from cultures. We thus obtained 91,225 sequences, of which 30,217 (33 %) were classified as environmental. The first environmental sequence was deposited in 1992 (*Epichloe* hybrid sp. e187, isolated from *Festuca arizonica* (An *et al.*, 1992)). Until 2003 on an average only 7.6 % of all fungal ITS sequences were from environmental sources in each year (Fig. 1). Since then, an average 37.5 % of the ITS sequences have come from the environment in each year (spreadsheets with all sequence statistics are available as Supplementary material). Our results parallel those of Ryberg *et al.* (2009), who surveyed ITS sequences deposited from 2000 to 2007 and divided them into "fully identified" and "insufficiently identified" categories, as well as Brock *et al.* (2009), who examined ITS sequences deposited up to April 2008.

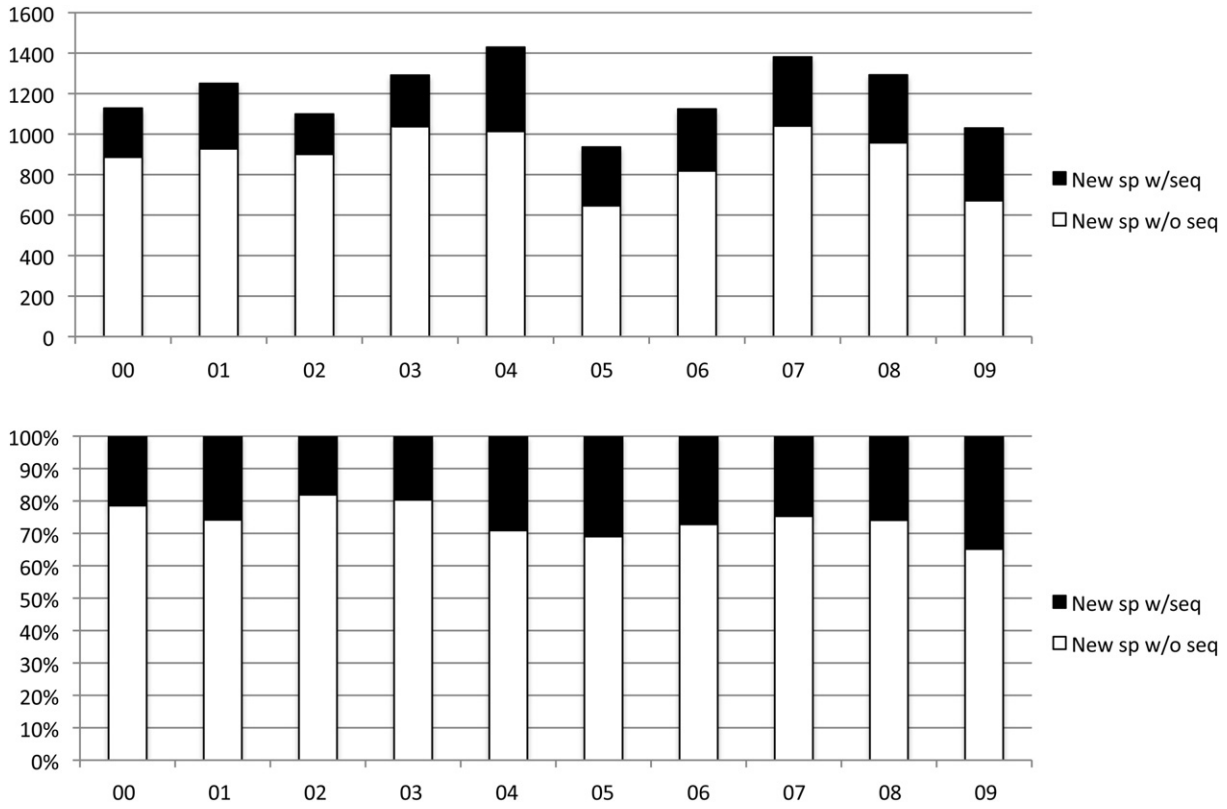

Fig. 1 – Rate of description of new species recorded in *Index Fungorum*, 2000–2009, including those with (black) and without (white) sequences of any locus now present in GenBank.

To assess the frequency of chimeric sequences among the environmental and specimen-based fractions, we used the ITS chimera checker of Nilsson *et al.* (2010), which identified 3089 (10 %) of the environmental sequences and 5178 (8 %) of the specimen-based sequences as potentially chimeric at the ordinal level. The chimera checker has been estimated to have a false-positive rate of about 80 %, but a negligible false negative rate. The values we obtained overestimate, therefore, the actual proportion of chimeras, but they nonetheless suggest that chimeras are more frequent in environmental sequences than in specimen-based sequences, as expected due to the greater template complexity of environmental samples.

To estimate the number of MOTUs represented in the ITS sequences, we performed hierarchial clustering. We first separated ITS1 and ITS2 sequences from the rRNA coding regions using the ITS extractor of Nilsson *et al.* (2010; 2009a). Next, to estimate an appropriate similarity threshold for MOTU delimitation, we examined 611 sets of sequences (16,194 sequences in total) that were deposited in GenBank with identical taxon names and that had at least 10 exemplars, and we obtained the minimum and average similarity scores from pairwise Clustal W 2.1 (Larkin *et al.*, 2007) alignments in both ITS1 and ITS2. We excluded 95 species in which the minimum pairwise similarity was less than 80 %, on the grounds that these probably include grossly misidentified sequences. Of the remaining 516 sets of sequences, the average minimum similarity in ITS1 and ITS2 was 93 %, which we used as a cut-off for subsequent MOTU delimitation. We do not wish to suggest that this value absolutely corresponds to species boundaries, but it does reflect the average minimum similarity in the ITS region among putatively conspecific sequences, based on names assigned to sequences in GenBank, excluding obvious misidentifications and entries of appreciably substandard read quality.

We next clustered all of the ITS sequences based on average pairwise similarity in ITS1 and ITS2 as established with Clustal W 2.1. In each clustering run, the first unprocessed sequence in the database was compared to all others, and a cluster was formed if the sequence hits one or more sequences at or above the threshold level. Any sequence that failed to hit any other sequence closely was left as a singleton. The next clustering round proceeded from the first remaining unprocessed sequence, and so on. To expedite calculations, we performed a first round of clustering using a threshold of 80 % similarity (which yielded 6149 clusters), and then subclustered within each 80 % cluster at a 93 % similarity threshold. This process yielded 16,969 clusters, with an average of 5.38 sequences each. The 9432 singletons (55.7 % of all clusters) included 4452 environmental sequences and 4980 specimen-based sequences; 582 (13 %) of the environmental singletons were identified as potential chimeras, compared to only 80 (1.6 %) of the specimen-based singletons. When potentially chimeric sequences were excluded, there were 16,141 clusters at the 93 % threshold level. These values correspond closely to the number of unique organism names

among the specimen-based sequences (16,856, which is inflated by orthographic variation and synonymy).

We next examined the distribution of environmental and specimen-based sequences among the clusters, as well as the date of deposition of the first sequence in each cluster. Overall, 8516 (50 %) of the clusters contained only specimen-based sequences, 6230 (37 %) contained only environmental sequences, and 2223 (13 %) contained both kinds of sequences (including putative chimeras) (Fig. 2). Our results contrast somewhat with those of Brock *et al.* (2009), who grouped ITS sequences from GenBank (using BLASTclust with a 97 % identity criterion) into 50,526 clusters, of which 72.7 % contained only "sufficiently identified" sequences, 25.9 % contained only "insufficiently identified" sequences, and 1.4 % contained both kinds of sequences. Nonetheless, both sets of results indicate that specimen-based studies and environmental surveys are sampling largely non-overlapping sets of taxa. The clusters containing only environmental sequences represent potentially novel taxa, while the clusters containing both kinds of sequences represent opportunities to expand understanding of the ecology and geographic distributions of described species. The large fraction of purely specimen-based clusters recalls the findings of Brock *et al.* (2009), who suggested, based on a sample of 509 specimens, that about 70 % of fungal taxa in herbaria are not yet represented in GenBank

The history of discovery of purely environmental clusters parallels that of the deposition of environmental sequences (Fig. 1). In 2007, 2008, and 2009, at least half of the clusters deposited in GenBank were derived from environmental sources, and in 2008 and 2009 the number of new environmental clusters exceeded the number of species described through traditional means as recorded in *Index Fungorum*. Once again, these figures are based only on Sanger sequencing data. To assess the potential impact of pyrosequencing methods, we surveyed 10 recent studies that use the new technology to estimate fungal diversity in various environments (Table 1) (Amend *et al.*, 2010; Buée *et al.*, 2009; Ghannoum *et al.*, 2010; Jumpponen and Jones, 2009; Jumpponen *et al.*, 2010; Lumini *et al.*, 2009; Öpik *et al.*, 2009; Rousk *et al.*, 2010; Tedersoo *et al.*, 2010; Wallander *et al.*, 2010). Very different approaches to filtering and analyzing the sequences were employed, which make it difficult to compare the results of the studies. Nevertheless, on average each study analyzed 54,000 sequences, which were grouped by various criteria into from 48 to 4473 MOTUs, of which from 32 % to 85 % were deemed to be insufficiently identified and could represent undescribed taxa (this information could not be extracted from all studies). In sum, it appears that fungal molecular ecologists have joined, or surpassed, fungal taxonomists at the forefront of species discovery.

Some of the unidentified MOTUs probably represent species that have been described but for which there are no reference sequences in GenBank (Brock *et al.*, 2009). However, the odds that an unidentified MOTU represents one of the approximately 77,000 described species for which there are no data in GenBank are about eighteen to one, based on the estimate of 1.5 million extant fungal species by Hawksworth

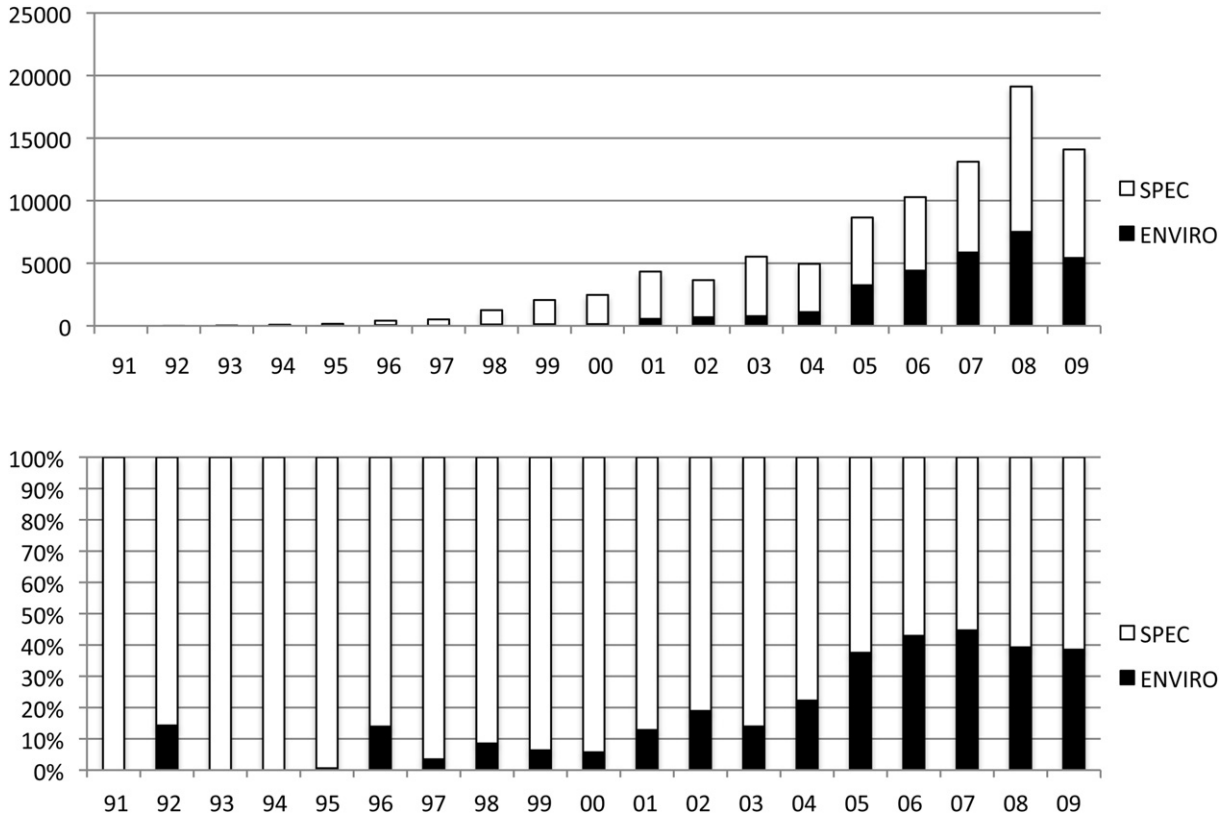

Fig. 2 – Rate of deposition of specimen-based (white) and environmental (black) ITS sequences in GenBank, 1991–2009.

Table 1 – Ten studies drawing from massively parallel pyrosequencing of *Fungi* in various environmental substrates.

Study	Number of sequences after quality filtering	Number/proportion of sequences regarded as fully identified	Number/proportion of sequences regarded as insufficiently identified	Number of fungal species/MOTUs	Habitat/host	Geographical region	Genetic marker	MOTU generation specifics
Buée et al. (2009)	166,000	76,000 (46 %) sequences	90,000 (54 %) sequences	1000	Soil samples in forest	France	ITS1	BLASTclust 97 %
Jumpponen and Jones (2009)	18,020	33 % of OTUs	66 % of OTUs	360 non-singletons	*Quercus* phyllosphere	US (Kansas)	ITS1	CAP3 95 %
Lumini *et al.* (2010)	4192	~25 % of OTUs assigned to Latin binomial	~75 % of OTUs not assigned to Latin binomial	~130 taxa of AMF species	Soil	Italy	18S	DOTUR 97 %
Öpik *et al.* (2009)	138,919	>91 % assignable to "virtual taxon" in MaarjAM; 27 % of the OTUs have full Latin names	<9 % not assignable to "virtual taxon", most of which were probably sequencing artifacts and errors; 73 % of OTUs lack full Latin names	48	Vascular plants sampled for AM *Fungi* in boreonemoral forest	Estonia	18S	BLAST/clustering
Amend *et al.* (2010)	97,557	90.5 % of OTUs identified to class	9.5 % of OTUs not identified to class	4473	Indoor dust	Global	ITS2	CD-HIT-EST 97 %
Ghannoum *et al.* (2010)	34,049	68 % of species fully identified	32 % of species not fully identified	101	*Homo sapiens* (oral cavity)	US (Ohio)	ITS1	CD-HIT [98 %?]
Jumpponen *et al.* (2010)	14,336	28 % of non-singletons matched closely with fully named taxon	72 % of non-singletons did not match closely with fully named taxon	1151 (50 % of which were singletons)	Tallgrass prairie soil	USA (Kansas)	ITS2	CAP3 95 %
Rousk *et al.* (2010)	4700	80 % of OTUs tentatively identified to species level	20 % of OTUs not identified to species level	945	Soil	UK	18S	CD-HIT 97 %
Tedersoo *et al.* (2010)	44,411	15 % of OTUs confidently assigned to Latin binomial	85 % of OTUs could not be confidently assigned to Latin binomial	214	Angiosperm root tips (primarily ECM fungi) in tropical rain forest	Cameroon	ITS1	Extracted ITS1, TGICL 97 %
Wallander *et al.* (2010)	18,000	33 % of OTUs identified to species level; another 21 % identified to genus level	67 % of the OTUs not identified to species level	248	Ingrowth bags in Norway spruce forests	Sweden	Full ITS region	Clustering through iterated BLAST searches at 98.5 %

(2001), or at least forty-four to one, based on the diversity estimates of O'Brien *et al.* (2005) (these odds are also based on an admittedly naïve assumption that all species are uniformly distributed in nature).

Whatever one thinks about the actual number of fungal species, the flood of unidentified MOTUs highlights the need to generate sequences based on described species for identification purposes (Brock *et al.*, 2009). It is therefore unfortunate that most new descriptions of fungal species do not include sequence data. Expanding on analyses that we reported previously (Hibbett *et al.*, 2009), we screened GenBank for sequences of any locus for the 11,960 new species from all groups of *Fungi* recorded in *Index Fungorum* in 1999–2009. We found that 8895 (74.4 %) of the newly described species have no sequence data in GenBank (including sequences deposited under homotypic synonyms) (Fig. 3).

4. The *status quo* and prospects for accelerating fungal taxonomy

Fungal taxonomy faces serious challenges. At current rates, it may take centuries or millennia to describe all the species of *Fungi* on Earth (assuming they can be found before they become extinct); molecular ecologists are now playing a significant, if not dominant, role in species discovery, which has traditionally been the sole province of taxonomists; and most new species are still being described without molecular data. To remedy this situation, fungal taxonomists should be prepared to re-evaluate current standard practices and consider adopting new modes of working.

To a limited extent, the rate of species description could be accelerated by modernizing the *International Code of Botanical Nomenclature* (McNeill *et al.*, 2006). In fact, proposals have been published to eliminate the requirement for Latin diagnoses (Figueiredo *et al.*, 2010) and to mandate deposition of new names in a publicly accessible database such as MycoBank or *Index Fungorum* (Hawksworth *et al.*, 2010). Polling of the delegates to the Ninth International Mycological Congress suggested that most mycologists support such changes (Norvell *et al.*, 2010), and we can only hope that they are approved at the next International Botanical Congress in 2011. Similarly, electronic publication of species descriptions could also improve the efficiency of the taxonomic enterprise, and a proposal to make this a reality under the *Code* will be considered.

The changes to nomenclatural practices described above would be helpful, but they would probably result in only a modest acceleration in the rate of taxon description, and they would not help close the gap between the described and sequenced dimensions of fungal diversity. To accomplish the latter, it will be necessary to encourage (and enable) authors of new taxa to deposit reference sequences, and to simultaneously generate sequences for the growing backlog of unsequenced fungal taxa, including epitypification of names for which satisfactory sequence data cannot be retrieved from the existing types (Hyde and Zhang, 2008). However, even if type or authentic materials of all the described species of *Fungi* could be sequenced (a virtual impossibility), the estimates of global diversity from

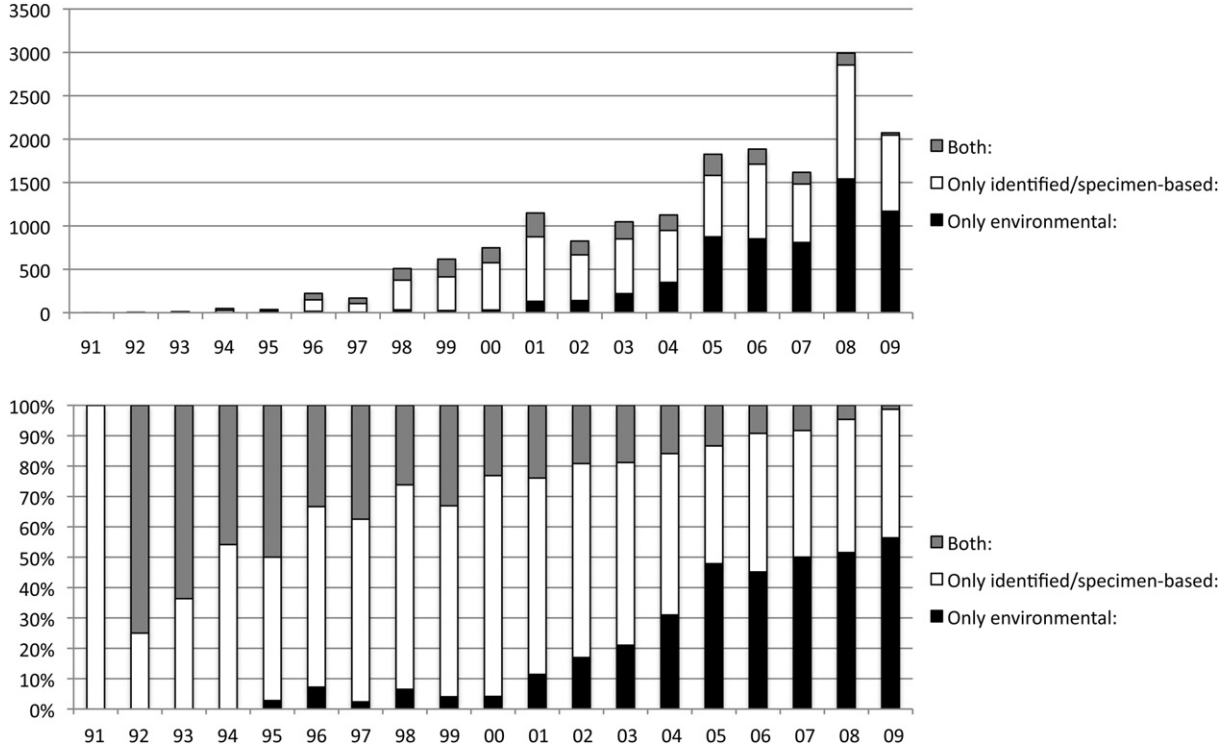

Fig. 3 – Rate of deposition of clusters of ITS sequences grouped with 93 % average similarity in ITS1 and ITS2 containing specimen-based (white), environmental (black), or both specimen-based and environmental (gray) sequences in GenBank, 1991–2009.

Hawksworth and others imply that fungal molecular ecologists will continue to detect large numbers of unidentified MOTUs representing undescribed taxa.

5. The desirability of naming MOTUs

Names enable communication. Phylogenetic and ecological studies routinely assign informal names or codes to MOTUs, but their application varies from one publication to another, which is certain to cause confusion. For example, an ITS-based MOTU represented by GenBank accessions AB244041 and DQ054545 was called "*Inocybe* sp. 3" when it was discovered by Nara (2006), but was later labeled "*Inocybe* sp. 2" by Ryberg *et al.* (2008). The *Code* promotes uniform application of specimen-based names, but there are no rules for MOTUs. To stabilize the nomenclature of MOTUs, special-purpose, sequence-based taxonomies are already being developed, including the MAARJam database for "Virtual Taxa" of *Glomeromycota*, which is based on 18S rRNA genes (Öpik *et al.*, 2010), a multilocus *Fusarium* database (Park *et al.*, 2010), and a forthcoming incremental MOTU registry in UNITE (Abarenkov *et al.*, 2010). Such resources facilitate communication among experts, but the MOTUs they contain are not translated into Latin binomials, and they are not integrated into species-based biodiversity databases, such as *Index Fungorum*, MycoBank, the Catalog of Life, the Global Biodiversity Information Facility, or the Encyclopedia of Life. This disconnect makes it difficult for information about MOTUs to reach non-specialists and to impact applied fields, such as conservation biology and plant pathology. The failure to include MOTUs in species counts for genera and higher taxa may also contribute to incomplete taxon sampling in phylogenetic studies and their downstream applications, such as historical biogeographic analyses and character evolution studies. To promote communication and awareness about fungal biodiversity, we suggest that well-supported MOTUs should be given names that are identical in form to those of specimen-based taxa, and that could be seamlessly integrated into existing taxonomic databases.

6. Strengths and limitations of environmental sequences for formal taxon description

Sequence data offer many advantages for taxon description. They are readily accessible through public databases and they facilitate formal testing of phylogenetic hypotheses. Importantly for high-throughput environmental studies, sequences are amenable to automated approaches to taxon identification and phylogenetic placement, such as those employed by UNITE (Abarenkov *et al.*, 2010), the naïve Bayesian classifier of the Ribosomal Database Project (Cole *et al.*, 2009; Wang *et al.*, 2007), or the *mor* automated phylogenetic taxonomy system (Hibbett *et al.*, 2005). Finally, sequences can be used to create fluorescent probes for visualization of organisms, and they can be extended using chromosome walking approaches (Porter *et al.*, 2008).

It must be accepted that sequence data, particularly those from single loci, have limitations for taxonomy and phylogenetic reconstruction. Chimeric sequences, which our results suggest are common among environmental singletons, are one potential source of error. The chimera checker of Nilsson *et al.* (2010) shows promise, and this and other programs, notably UCLUST (Edgar, in press), should only become more accurate as the reference database of ITS sequences expands. Gene tree/species tree incongruence is another vexing problem for systematics studies, particularly at low taxonomic levels, where incomplete lineage sorting and hybridization are the most likely. The current gold standard for assessing species limits using molecular data is the genealogical concordance approach, which uses multiple unlinked loci to assess the limits of reticulation (Taylor *et al.*, 2000), but this is not applicable to single-locus datasets, such as those generated in standard environmental surveys.

Perhaps the most worrisome source of error in sequence-based taxon description is intragenomic heterogeneity. Gene duplication and divergence can affect all loci, but intragenomic heterogeneity is of particular concern for the tandemly repeated nuclear rRNA genes. Major variation among ITS copies within organisms has been recognized in a few fungal taxa, such as *Fusarium* (O'Donnell and Cigelnik, 1997), and minor variation causing sequencing artifacts frequently necessitates cloning of PCR products in diverse groups. Nevertheless, the combined effect of concerted evolution and the swamping of rare variants in PCR amplicon pools has been thought to effectively homogenize copies of ITS, which remains a widely used marker for species-level taxonomy. However, a recent study in the polypore genus *Laetiporus* (Linder and Banik, in press) suggests that intragenomic heterogeneity in ITS may be more extensive than previously thought. Linder and Banik cloned and sequenced 1399 ITS amplicons from 21 individuals representing seven species, and they computed consensus ITS sequences for each individual. Phylogenetic analysis of consensus sequences yielded groupings consistent with recognized species limits, but analyses of individual clones revealed rare ITS variants that appear to have escaped concerted evolution and that did not cluster according to expected species groupings. Specifically, 45 sequences (i.e., 3.2 % of the clones) were less than 95 % identical to the consensus sequences of the strains from which they were derived, and therefore would be recognized as unique MOTUs using standard criteria. One species, *Laetiporus cincinnatus*, was particularly problematical; in addition to possessing highly divergent ITS copies, *L. cincinnatus* also yielded multiple clones that were very similar to those from *L. sulphureus*, possibly reflecting hybridization.

All scientific measurements have sources of error, but that is not a reason to cease data collection and analysis. Morphology-based species delimitation can also be misled by factors such as hybridization, cryptic speciation, convergent evolution, and pleomorphy, yet there are no formal barriers to the publication of new taxa based solely on morphology. In the case of MOTUs, we suggest that an emphasis should be placed on quantifying the impact of gene tree/species tree incongruence and intragenomic heterogeneity on taxon delimitation, based on data obtained from templates of varying degrees of complexity (e.g., soils, ectomycorrhizal root tips, pure cultures), and with different protocols (clone libraries *vs.* direct sequencing; pyrosequencing *vs.* Sanger sequencing) (Tedersoo *et al.*, 2010). A better understanding of the naturally occurring variation in ITS and other

markers, and their representation in empirical samples, will enable development of appropriate quality-control criteria, perhaps involving detection of accelerated rates of evolution, or loss of Watson–Crick base pairing in rRNA secondary structure (Schultz and Wolf, 2009).

7. Options for formal classification of environmental sequences

The absence of a type specimen is an obstacle to formal naming of MOTUs under the *Code*. At the same time, the absence of standardized names for MOTUs is a barrier to communication and understanding that must be overcome. One way to resolve this impasse would be to radically expand the concept of an acceptable nomenclatural type to include a portion of a sample from which an environmental DNA extract was made. This solution would go beyond Reynolds and Taylor's (1991) suggestion that a pure DNA sample of a single fungus, or an amplified segment of that DNA, could serve as type material. Alternatively, an image of the sequence trace file (or an alignment) could be designated as the type, since illustrations are already permitted as types in the current *Code* (Art. 37.5).

Even if the concept of a nomenclatural type could be broadened as described above, legitimate concerns regarding intragenomic heterogeneity and other sources of error could create resistance to formal naming of species based on environmental sequences. Another potential solution is available, based on the *candidatus* taxon concept that was proposed for prokaryotes in the mid-1990s (Murray and Schleifer, 1994; Murray and Stackebrandt, 1995). Like the botanical *Code*, the *International Code of Nomenclature of Bacteria* (Lapage *et al.*, 1992) stipulates that type material (a culture) is required to formally describe a new species of prokaryotes. Murray and colleagues proposed that taxa detected with sequence data, and for which some other information such as morphology was available, could be described on a provisional basis with *candidatus* status. A *candidate species* category for *Fungi* (omitting the requirement for non-molecular characters) would enable the naming of fungal MOTUs, while signaling that they are provisional taxa for which there are no specimens or cultures. Candidate genera could also be described, which would make it possible to assign binomials to MOTUs that are not nested within any clade currently classified at the generic level, but for the present we restrict our focus to candidate species.

If formal naming of MOTUs is to be adopted (whether as species or candidate species), it will be necessary for the mycological community to reach consensus about many issues, including the criteria for recognition. Minimum standards for naming MOTUs might include the following: (1) representation by at least two full-length sequences of an appropriate genetic marker, ideally consistent with the upcoming barcode standard for *Fungi* (Eberhardt, 2010) and derived from independent studies, with one sequence designated as the reference sequence; (2) a published phylogenetic analysis demonstrating monophyly, and considering all relevant publicly available sequences retrieved using BLAST or a similar method; (3) application of chimera checking software and other quality-control measures; (4) locality data, at least for the reference sequence; and, (5) registration in a taxonomic database. These criteria do not include a specific sequence similarity cut-off, and they do not mandate use of ITS. Rather, we feel that taxon delimitation should be based on the judgment of systematists, with explicit reference to phylogenetic trees created using the gene(s) that communities of experts have determined are appropriate for species delimitation in their clades of interest. Beyond these general suggestions, it would be premature (to say the least) to specify detailed protocols for MOTU-based taxonomy here. If the idea of formal classification of environmental sequences gains support, it will be necessary for the community of fungal taxonomists to work out detailed plans for implementation. In the meantime, we offer one hypothetical species description based on environmental sequences, with an environmental sample as the type material. This example demonstrates that taxon descriptions based solely on environmental sequences can be rich in information regarding ecology, biogeography, and phylogeny, while also providing excellent resources for identification. Additional hypothetical examples based on pure DNA extracts of a single fungus were presented by Reynolds and Taylor (1991).

Inocybe narai Hibbett & P.M. Kirk **sp. nov.**

MycoBank no.: MB XXXXXX

Etymology: The epithet honors Kazuhide Nara, who obtained the reference sequence.

Diagnosis: The least inclusive group containing organisms with nuclear rRNA ITS sequences with GenBank accessions AB244041 and DQ054545.

Reference phylogeny: M. Ryberg *et al. BMC Evolutionary Biology* **8**: 50, 2008 (additional file 1, fig. A).

Reference sequence: GenBank AB244041 (K. Nara. *New Phytologist* **171**: 187–198, 2006).

Other included sequences: GenBank DQ054545 (Wilson *et al.*, 2008). Sequence similarity: 99.09 % (ITS1), 98.92 % (ITS2).

Nomenclatural type: Soil sample and *Larix kaempferi* root tips, collected by Nara in October, 2010, preserved in the Kew fungarium K(M) nnnnnn.

Quality control: Chimera checker (Nilsson *et al.*, 2009b) results negative for both included sequences. Boundaries of 18S, 25S, and 5.8S rRNA coding regions identified with ITS extractor (Nilsson *et al.*, 2010). The sequence contains no DNA ambiguity symbols.

Reference sequence locality: Japan: Shizuoka, Gotenba, Mt. Fuji, 1450–1600 m asl.

Synonyms: *Inocybe* sp. 2 (Ryberg *et al.*, 2008); *Inocybe* sp. 3 (Nara, 2006).

Phylogenetic notes: Strongly supported as monophyletic (parsimony bootstrap = 100 %). Environmental sequence AY702727 was placed as the sister group (parsimony bootstrap = 95 %).

Ecological notes: The reference sequence was obtained from an ectomycorrhizal root tip of *Larix kaempferi* in the "volcanic desert" of Mt. Fuji, Japan. Nara considered this to be a later-stage species in succession. The other included sequence was obtained from soil at ca. 50 cm depth under beech and chestnut at ca. 1000 m asl on the extinct volcano, Monte Amiata, Tuscany, Italy. The closely related undescribed sequence AY702727 was obtained from ectomycorrhizal root tips of *Abies* sp. at 2600 m asl in the Sierra National Forest, California, USA (Izzo *et al.*, 2005).

8. Concluding thoughts

Formal naming of fungal environmental sequences would accelerate the rate of taxon description and enhance

awareness of the diversity, distribution, and ecological roles of *Fungi*. However, this endeavor faces serious technical and conceptual challenges, and we can imagine that some mycologists would prefer that taxonomy remains exclusively based on specimens. What will happen if fungal taxonomy fails to incorporate environmental sequences? It seems certain that the flood of environmental sequences will only grow in volume, and that this will intensify the need for sequence-based classification systems. Consequently, special-purpose databases of MOTUs, such as those that already exist for *Fusarium* and *Glomeromycota*, will proliferate, and their use of alphanumerical codes (or GUIDs/UUIDs), rather than Latin binomials, will exacerbate the disconnect with traditional taxonomy. As the MOTU databases expand, *Code*-compliant specimen-based taxonomy will come to represent less and less of the total knowledge of global fungal biodiversity, which could threaten its long-term support. To avert this scenario, we suggest that it will be necessary to integrate environmental sequences into the specimen-based taxonomic data stream.

Acknowledgements

DSH thanks Nick Read for the invitation to present the plenary lecture on which this article is based at the Ninth International Mycological Congress in Edinburgh. The authors also thank Dan Linder for sharing his unpublished manuscript on *Laetiporus*, and Anthony Amend, Valeria Bianciotto, Ari Jumpponen, Björn Lindahl, Marja Öpik, Johannes Rousk, and Leho Tedersoo for information regarding their recent pyrosequencing studies. The United States National Science Foundation supported this work through awards DEB-0732968, DEB-0933081, and DEB-1020421, including Research Experiences for Undergraduates Supplements for AO and DG.

Supplementary material

Supplementary data related to this article can be found online at doi:10.1016/j.fbr.2011.01.001.

REFERENCES

Abarenkov, K., Nilsson, R.-H., Larsson, K.-H., Alexander, I., Eberhardt, U., Erland, S., Høiland, K., Kjøller, R., Larsson, E., Pennanen, T., Sen, R., Taylor, A., Tedersoo, L., Ursing, B., Vrålstad, T., Liimatainen, K., Peintner, U., Kõljalg, U., 2010. The UNITE database for molecular identification of fungi – recent updates and future perspectives. New Phytologist 186, 281–285.

Amend, A.S., Seifert, K.A., Samson, R., Bruns, T.D., 2010. Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. Proceedings of the National Academy of Sciences of the United States of America 107, 13748–13753.

An, Z.-Q., Liu, J.-S., Siegel, M.R., Bunge, G., Schardl, C.L., 1992. Diversity and origins of endophytic fungal symbionts of the North American grass *Festuca arizonica*. Theoretical and Applied Genetics 85, 366–371.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2009. GenBank. Nucleic Acids Research 38, D46–D51.

Brock, P.M., Döring, H., Bidartondo, M.I., 2009. How to know unknown fungi: the role of a herbarium. New Phytologist 181, 719–724.

Buée, M., Reich, M., Murat, C., Morin, E., Nilsson, R.H., Uroz, S., Martin, F., 2009. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. New Phytologist 184, 449–456.

Cole, J., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R., Kulam-Syed-Mohideen, A., McGarrell, D., Marsh, T., Garrity, G., Tiedje, J., 2009. The ribosomal database project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research 37, D141–D145.

Eberhardt, U., 2010. A constructive step towards selecting a DNA barcode for fungi. New Phytologist 187, 265–268.

Edgar, R. Search and clustering orders of magnitude faster than BLAST. Bioinformatics, in press.

Figueiredo, E., Moore, G., Smith, G.F., 2010. Latin diagnosis: time to let go. Taxon 59, 617–620.

Ghannoum, M.A., Jurevic, R.J., Mukherjee, P.K., Cui, F., Sikaroodi, M., Naqvi, A., Gillevet, P.M., 2010. Characterization of the oral fungal microbiome (Mycobiome) in healthy individuals. PLoS Pathogens 6 e1000713.

Hawksworth, D.L., 2001. The magnitude of fungal diversity: the 1.5 million species estimate revisited. Mycological Research 105, 1422–1432.

Hawksworth, D.L., Cooper, J.A., Crous, P.W., Hyde, K.D., Iturriaga, T., Kirk, P.M., Lumbsch, H.T., May, T.W., Minter, D.W., Misra, J.K., Norvell, L., Redhead, S.A., Rossman, A.Y., Seifert, K.A., Stalpers, J.A., Taylor, J.W., Wingfield, M.J., 2010. (117–119) Proposals to make the pre-publication deposit of key nomenclatural information in a recognized repository a requirement for valid publication of organisms treated as fungi under the Code. Taxon 59, 656–666.

Hibbett, D., Nilsson, R., Snyder, M., Fonseca, M., Costanzo, J., Shonfeld, M., 2005. Automated phylogenetic taxonomy: an example in the Homobasidiomycetes (mushroom-forming fungi). Systematic Biology 54, 660–668.

Hibbett, D.S., Ohman, A., Kirk, P.M., 2009. Fungal ecology catches fire. New Phytologist 184, 279–282.

Hughes, K.W., Petersen, R.H., Lickey, E.B., 2009. Using heterozygosity to estimate a percentage DNA sequence similarity for environmental species' delimitation across basidiomycete fungi. New Phytologist 182, 792–794.

Hyde, K.D., Zhang, Y., 2008. Epitypification: should we epitypify? Journal of Zhejiang University Science B 9, 842–846.

Izzo, A., Agbowo, J., Bruns, T.D., 2005. Detection of plot-level changes in ectomycorrhizal communities across years in an old-growth mixed-conifer forest. New Phytologist 166, 619–630.

Jumpponen, A., Jones, K.L., 2009. Massively parallel 454 sequencing indicates hyperdiverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. New Phytologist 184, 438–448.

Jumpponen, A., Jones, K.L., Blair, J., 2010. Vertical distribution of fungal communities in tallgrass prairie soil. Mycologia 102.

Kirk, P., Cannon, P., Stalpers, J., 2008. Dictionary of the Fungi, 10th ed. CABI, Wallingford.

Lapage, S., Sneath, P., Lessel, E., Skerman, V., Seeloger, H., Clark, W., 1992. International Code of Nomenclature of Bacteria and Statutes of the International Committee on Systematic Bacteriology and Statutes of the Bacteriology and Applied Microbiology Section of the International Union of Microbiological Societies. Bacteriological Code (1990 Revision). American Society for Microbiology, Washington DC.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.

Linder, D., Banik, M. Intra-genomic variation in the ITS rDNA region obscures phylogenetic relationships and inflates estimates of operational taxonomic units in genus Laetiporus. Mycologia 107, in press.

Lücking, R., 2008. Taxonomy: a discipline on the brink of extinction. Are DNA barcode scanners the future of biodiversity research? Archives des Sciences 61, 75–88.

Lumini, E., Orgiazzi, A., Borriello, R., Bonfante, P., Bianciotto, V., 2010. Disclosing arbuscular mycorrhizal fungal biodiversity in soil through a land-use gradient using a pyrosequencing approach. Environmental Microbiology 12, 2165–2179.

McNeill, J., Barrie, F., Burdet, H., Demoulin, V., Hawksworth, D., Marhold, K., Nicolson, D., Prado, J., Silva, P., Skog, J., Wiersema, J., Turland, N., 2006. International Code of Botanical Nomenclature (Vienna Code) adopted by the Seventeenth International Botanical Congress Vienna, Austria, July 2005. Gantner Verlag, Liechtenstein.

Mueller, G.M., Schmit, J.P., 2007. Fungal biodiversity: what do we know? What can we predict? Biodiversity and Conservation 16, 1–5.

Murray, R., Schleifer, K., 1994. Taxonomic notes: a proposal for recording the properties of putative taxa of procaryotes. International Journal of Systematic Bacteriology 44, 174–176.

Murray, R., Stackebrandt, E., 1995. Taxonomic note: implementation of the provisional status candidatus for incompletely described procaryotes. International Journal of Systematic Bacteriology 45, 186–187.

Nara, K., 2006. Pioneer dwarf willow may facilitate tree succession by providing late colonizers with compatible ectomycorrhizal fungi in a primary successional volcanic desert. New Phytologist 171, 187–198.

Nilsson, R.H., Abarenkov, K., Veldre, V., Nylinder, S., De Wit, P., Brosché, S., Alfredsson, J.F., Ryberg, M., Kristiansson, E., 2010. An open source chimera checker for the fungal ITS region. Molecular Ecology Resources 10, 1076–1081.

Nilsson, R.H., Bok, G., Ryberg, M., Kristiansson, E., Hallenberg, N., 2009a. A software pipeline for processing and identification of fungal ITS sequences. Source Code for Biology and Medicine 4, 1.

Nilsson, R.H., Ryberg, M., Abarenkov, K., Sjökvist, E., Kristiansson, E., 2009b. The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. FEMS Microbiology Letters 296, 97–101.

Norvell, L.L., Hawksworth, D.L., Petersen, R.H., Redhead, S.A., 2010. IMC9 Edinburgh nomenclature sessions. IMA Fungus 1, 143–147.

O'Brien, H.E., Parrent, J.L., Jackson, J.A., Moncalvo, J.M., Vilgalys, R., 2005. Fungal community analysis by large-scale sequencing of environmental samples. Applied and Environmental Microbiology 71, 5544–5550.

O'Donnell, K., Cigelnik, E., 1997. Two divergent intragenomic rDNA ITS2 types within a monophyletic lineage of the fungus Fusarium are nonorthologous. Molecular Phylogenetics and Evolution 7, 103–116.

Öpik, M., Metsis, M., Daniell, T.J., Zobel, M., Moora, M., 2009. Large-scale parallel 454 sequencing reveals host ecological group specificity of arbuscular mycorrhizal fungi in a boreonemoral forest. New Phytologist 184, 424–437.

Öpik, M., Vanatoa, A., Vanatoa, E., Moora, M., Davison, J., Kalwij, J.M., Reier, Ü, Zobel, M., 2010. The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). New Phytologist 188, 223–241.

Park, B., Park, J., Cheong, K.-C., Choi, J., Jung, K., Kim, D., Lee, Y.-H., Ward, T.J., O'Donnell, K., Geiser, D.M., Kang, S., 2010. Cyber infrastructure for Fusarium: three integrated platforms supporting strain identification, phylogenetics, comparative genomics and knowledge sharing. Nucleic Acids Research 2010, 1–7.

Peay, K.G., Kennedy, P.G., Bruns, T.D., 2008. Fungal community ecology: a hybrid beast with a molecular master. Bioscience 58, 799–810.

Porter, T., Schadt, C., Rizvi, L., Martin, A., Schmidt, S., Scottdenton, L., Vilgalys, R., Moncalvo, J., 2008. Widespread occurrence and phylogenetic placement of a soil clone group adds a prominent new branch to the fungal tree of life. Molecular Phylogenetics and Evolution 46, 635–644.

Reynolds, D., Taylor, J., 1991. DNA specimens and the 'International code of botanical nomenclature'. Taxon 40, 311–351.

Rousk, J., Baath, E., Brookes, P.C., Lauber, C.L., Lozupone, C., Caporaso, J.G., Knight, R., Fierer, N., 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME Journal 4, 1340–1351.

Ryberg, M., Kristiansson, E., Sjökvist, E., Nilsson, R.H., 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, 471–477.

Ryberg, M., Nilsson, R.H., Kristiansson, E., Töpel, M., Jacobsson, S., Larsson, E., 2008. Mining metadata from unidentified ITS sequences in GenBank: a case study in Inocybe (Basidiomycota). BMC Evolutionary Biology 8, 50.

Schmit, J.P., Mueller, G.M., 2007. An estimate of the lower limit of global fungal diversity. Biodiversity and Conservation 16, 99–111.

Schultz, J., Wolf, M., 2009. ITS2 sequence–structure analysis in phylogenetics: a how-to manual for molecular systematics. Molecular Phylogenetics and Evolution 52, 520–523.

Taylor, J., Jacobson, D., Kroken, S., Kasuga, T., Geiser, D., Hibbett, D., Fisher, M., 2000. Phylogenetic species recognition and species concepts in fungi. Fungal Genetics and Biology 31, 21–32.

Tedersoo, L., Nilsson, R.H., Abarenkov, K., Jairus, T., Sadam, A., Saar, I., Bahram, M., Bechem, E., Chuyong, G., Kõljalg, U., 2010. 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. New Phytologist 188, 291–301.

Wallander, H., Johansson, U., Sterkenburg, E., Brandström Durling, M., Lindahl, B.D., 2010. Production of ectomycorrhizal mycelium peaks during canopy closure in Norway spruce forests. New Phytologist 187, 1124–1134.

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73, 5261–5267.

Wilson, M.J., Certini, G., Campbell, C.D., Anderson, I.C., Hillier, S., 2008. Does the preferential microbial colonisation of ferromagnesian minerals affect mineral weathering in soil? Naturwissenschaften 95, 851–858.