Where is the unseen fungal diversity hidden? A study of *Mortierella* reveals a large contribution of reference collections to the identification of fungal environmental sequences

László G. Nagy[1], Tamás Petkovits[1], Gábor M. Kovács[2,3], Kerstin Voigt[4,5], Csaba Vágvölgyi[1] and Tamás Papp[1]

[1]Faculty of Science and Informatics, Department of Microbiology, University of Szeged, Közép fasor 52, H-6726 Szeged, Hungary; [2]Institute of Biology, Department of Plant Anatomy, Eötvös Loránd University, Pázmány Péter sétány 1/C, H-1117 Budapest, Hungary; [3]Plant Protection Institute of the Hungarian Academy of Sciences, H-1525 Budapest, PO Box 102, Hungary; [4]School of Biology and Pharmacy, Institute of Microbiology, Department of Microbiology and Molecular Biology, Jena Microbial Resource Collection, University of Jena, Neugasse 25, D-07743 Jena, Germany; [5]Leibnitz-Institute for Natural Product Research and Infection Biology (HKI), Department of Molecular and Applied Microbiology, Beutenbergstrasse 11a, D-07745 Jena, Germany

Author of correspondence:
László G. Nagy
Tel: +36 62 544005
Email: cortinarius2000@yahoo.co.uk

Summary

• Estimation of the proportion of undescribed fungal taxa is an issue that has remained unresolved for many decades. Several very different estimates have been published, and the relative contributions of traditional taxonomic and next-generation sequencing (NGS) techniques to species discovery have also been called into question recently.

• Here, we addressed the question of what proportion of hitherto unidentifiable molecular operational taxonomic units (MOTUs) have already been described but not sequenced, and how many of them represent truly undescribed lineages. We accomplished this by modeling the effects of increasing type strain sequencing effort on the number of identifiable MOTUs of the widespread soil fungus *Mortierella*.

• We found a nearly linear relationship between the number of type strains sequenced and the number of identifiable MOTUs. Using this relationship, we made predictions about the total number of *Mortierella* species and found that it was very close to the number of described species in *Mortierella*.

• These results suggest that the unusually high number of unidentifiable MOTUs in environmental sequencing projects can be, at least in some fungal groups, ascribed to a lag in type strain and specimen sequencing rather than to a high number of undescribed species.

Introduction

Estimation of the number of fungal species has always been a challenge for mycologists, and mutually exclusive estimates have been published over the last few decades. Recently, Hibbett *et al.* (2009, 2011) have hypothesized that the vast majority of 'unidentifiable' molecular operational taxonomic units (MOTUs) recovered by environmental sequencing projects represent undescribed species and/or lineages and asked 'whether ecology or taxonomy is currently leading the way in the discovery of new taxa'. It has also been suggested that environmental sequence-based formal species descrip-

tions should be given legitimacy. Their conclusions were illustrated by 10 different studies of fungal communities (Buée *et al.*, 2009; Jumpponen & Jones, 2009; Öpik *et al.*, 2009; Amend *et al.*, 2010; Ghannoum *et al.*, 2010; Jumpponen *et al.*, 2010; Lumini *et al.*, 2010; Rousk *et al.*, 2010; Tedersoo *et al.*, 2010; Wallander *et al.*, 2010) and their failure to find significant matches for the majority of their MOTUs when compared with public sequence databases, such as GenBank and UNITE (Kõljalg *et al.*, 2005). As a result of the inventory of studies listed above, Hibbett *et al.* (2009) reported that at least 1130 potentially novel taxa were recovered in these studies, which exceeded the

number of species (1009) formally described by taxonomists in three main fungal groups (Ascomycota, Basidiomycota and Glomeromycota) in 2008. This raised the question of how many of these taxa had already been described but not sequenced, and how many represented truly novel lineages (Hibbett *et al.*, 2009, 2011). Comparisons involving the numbers of described and estimated fungal taxa led to the conclusion that it is unlikely that many of the unidentified OTUs have already been described and the vast majority of them probably represent truly undescribed lineages. Indeed, there are great differences between the numbers of described (*c.* 100 000) and estimated (1.5–3.5 million) fungal taxa (Hawksworth, 2001; O'Brien *et al.*, 2005; Mueller & Schmit, 2007). Because of this discrepancy, in our opinion, any inferences or estimates based on them (either 100 000 or 1.5–3.5 million) should be treated with caution. In view of the taxonomic coverage of public databases and the reliability of published DNA sequences (Nilsson *et al.*, 2006; Bidartondo *et al.*, 2008; Brock *et al.*, 2008; Ryberg *et al.*, 2008), however, other scientists adopt a different standpoint; that is, that the high number of sequences that cannot be robustly assigned to species on the basis of BLAST searches mostly belong to species that have already been described, but have not been sequenced or are not present in the queried database (Brock *et al.*, 2008; Bidartondo *et al.*, 2009). Nevertheless, there is a consensus about the need for the large-scale sequencing of authentic strains and type specimens lying in a herbarium or culture collection.

The estimations by Hibbett *et al.* (2009) prompted us to look into this problem and to try to obtain a figure reflecting the proportion of MOTUs that could be identified to species level if all type materials in herbaria and other public collections were sequenced. Using the zygomycete genus *Mortierella* (Mortierellales), in the present study we modeled the effects of increasing the number of sequenced authentic strains on the identifiability of environmental sequences. The genus *Mortierella* is a good study group for this purpose, as it constitutes a ubiquitous group very frequently isolated from soils (e.g. Linnemann, 1941; Zycha *et al.*, 1969; Gams, 1977; Buée *et al.*, 2009) and the number of environmental and unidentified sequences in GenBank is high. Furthermore, *Mortierella*, as one of the 10 most frequently recovered fungal genera in environmental sequencing projects (http://www.emerencia.org/genuslist.html), and is thus a very suitable genus in which to address the question of unidentifiable sequences. Species of *Mortierella* are widespread in the temperate zone, where they are almost cosmopolitan with regard to several ecological factors, occurring in a wide range of habitats. Members of the genus *Umbelopsis*, now placed in the Mucorales, were also included in the analyses, because they had earlier been treated as a subgenus of *Mortierella*, which resulted in a large number of sequences deposited as *Mortierella* in public databases.

Materials and Methods

Taxon sampling and laboratory work

For this study, we assembled two data sets, one containing all of the mortierellalean GenBank sequences as of 10 August 2010, and a reference data set, which contained sequences of authentic and type strains of *Mortierella* and the closely related genus *Umbelopsis*. Taxa selected for inclusion in the reference data set were either type strains or other authentic strains for which the species identity is unambiguous and has been verified by us. Although extensive synonymy of some of the included species is widely accepted in the literature, we included all type strains under the name they have been described, as the databases to which we compared our estimates (see the 'Results' section) also contain these synonyms. In other words, the extent to which our reference data set is loaded with synonyms is presumably the same for the number of described species (obtained from Index Fungorum and MycoBank, see the following section, 'Alignment, clustering and the effects of taxon sampling depth'). An attempt has been made to locate as many type strains as possible. Therefore, we obtained type cultures from the Centraalbureau voor Schimmelcultures (CBS) (Utrecht, the Netherlands), NRRL (Agricultural Research Service Culture Collection, Peoria, IL, USA), FSU (Friedrich Schiller Universität, Jena Microbial Resource Collection, University of Jena, Jena, Germany) and the Szeged Microbiological Collection (SzMC) (University of Szeged, Szeged, Hungary). Despite repeated attempts to obtain them, type materials for many of the species proved to be unavailable.

For DNA extraction, strains were grown in liquid malt-extract medium (5% malt extract and 1% glucose) at 20–37°C for 7–12 d, depending on the requirements of the fungus. Genomic DNA was prepared from *c.* 10 mg of mycelium ground to a fine powder in liquid nitrogen and purified using the MasterPure Yeast DNA purification kit (Epicentre Biotechnologies, Madison, WI, USA) according to the instructions of the manufacturer. PCR amplification and sequencing of the ITS1-5.8S-ITS2 region was carried out with the ITS1-ITS4 primer pair, using standard protocols (White *et al.*, 1990). Cycle sequencing of both strands was performed by LGC Genomics (Berlin, Germany). Individual readings were assembled into contigs using the PREGAP and GAP4 programs of the Staden Package (Staden *et al.*, 2008). All sequences have been deposited in GenBank (Supporting Information Table S1).

Alignment, clustering and the effects of taxon sampling depth

In order to obtain all mortierellalean GenBank sequences, we performed sequential BLAST searches against GenBank and

queried the Emerencia database (Ryberg *et al.*, 2008) for the complete ITS region (ITS1-5.8S-ITS2) of *Mortierella* species. From GenBank, we downloaded both identified and unidentified sequences, as the likelihood of misidentification of database sequences is high in taxonomically difficult groups, such as the genus *Mortierella*. We subjected to BLAST analysis (Altschul *et al.*, 1990) a set of *Mortierella* and *Umbelopsis* sequences which we sorted out from an order-level multigene tree of the Mortierellales (T. Petkovits *et al.*, unpublished). BLAST was launched with default parameters. The first 100 best hits were saved from each round of BLAST search and the process was repeated until no new sequences could be obtained (each step contained a filter for duplicate items). To exploit all information in the ITS region and ensure complete overlap with the newly generated contigs, we discarded all sequences shorter than 400 bp. Thus, our data set excluded 454-based environmental sequence data.

Preliminary, crude alignments were computed using MUSCLE 3.7 (Edgar *et al.*, 2005), and neighbor-joining trees were built in PAUP 4.0b10 (Swofford, 2003) to identify non-mortierellalean sequences. After the removal of mucoralean and other taxa, the data set consisted of 832 unique ITS sequences. This was merged with 102 sequences of authentic strains, sequenced for the present study (alignment available at TreeBase No. 11260, www.treebase.org). The combined data set, containing 934 ITS sequences of *Mortierella* and *Umbelopsis*, was aligned by means of the PROBALIGN algorithm (Roshan, 2006), which uses a model for biologically realistic gap placement. We omitted formal model testing because of the computational complexity of the problem and because the GTR + G model has been suggested for highly variable, information-rich alignments (Stamatakis, 2006). After the exclusion of ambiguously aligned regions, we inferred maximum likelihood (ML) trees under the GTR + G model of evolution in PHYML 3.0 (Guindon & Gascuel, 2003). The gamma distribution was discretized into 10 rate categories, and the branch swapping algorithm was set to SPR. An ITS sequence of *Mucor racemosus* (DQ273434) was used as the outgroup. The resulting ML tree is presented as Fig. S1 and has been deposited in TreeBase (Study No. 11260).

We arranged the sequences into MOTUs by agglomerative clustering of ML genetic distances computed by pairwise comparisons under the HKY model of evolution in PAUP 4.0b10 (Swofford, 2003). Average linkage clustering was performed in MOTHUR (Schloss *et al.*, 2009). Following the general consensus for the ITS region, we set the similarity threshold to 97% (Nilsson *et al.*, 2008).

After MOTU designation, we made predictions about the number of undescribed species in *Mortierella*, by modeling the effects of gradually increasing efforts to sequence authentic strains for the unambiguous identification of environmental samples. In doing this, we randomly removed 10, 20, 30, 40, 50, 60, 70, 80 and 90% of the

authentic strains from our data set and counted the number of MOTUs that 'lost their names'. Random removal of authentic strains was performed by assigning a number to each authentic strain and generating random numbers covering 10, 20, 30 and 90% of them using a standard random number generator. Then we plotted the number of identified MOTUs (i.e. MOTUs that contain sequences of authentic strains) against the number of authentic strains and species included in the data set using Microsoft Excel (Fig. 1a,b). A correction for taxonomic synonyms both in the reference data set and in the taxonomic database used for comparison (e.g. Index Fungorum) has been omitted, as the sampling of *Mortierella* strains was random (no attempt was made to omit or collect taxonomic synonyms), and thus the distribution of taxonomic synonyms is expected to be the same as in the specific database.

Results

The BLAST searches resulted in 924 unique ITS1-5.8S-ITS2 sequences. After the removal of overly short (< 400 bp) and nonmortierellalean sequences, the data set contained 832 unique ITS sequences. A search for the complete ITS region in Emerencia (Ryberg *et al.*, 2008) resulted in 927 hits, of which 653 were longer than 400 bp. These showed complete overlap with the data set downloaded from GenBank. This was merged with the reference data set containing 102 newly obtained sequences of 78 described
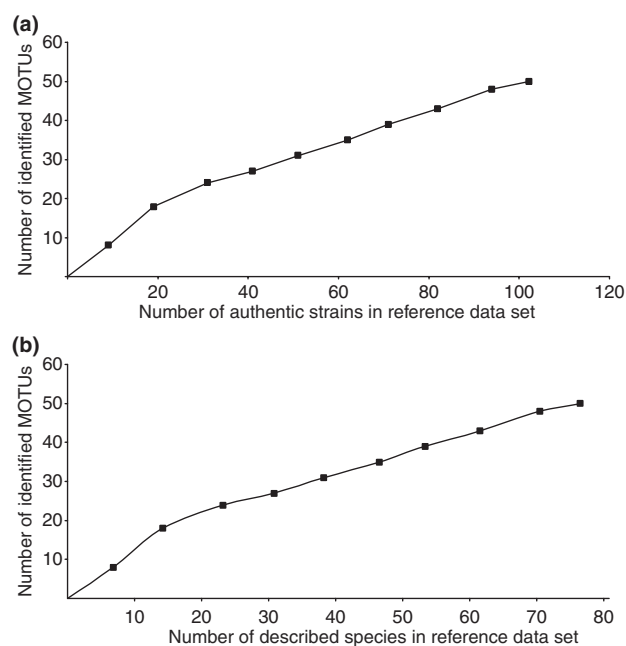

Fig. 1 A model of the effects of increasing type and authentic strain sequencing effort. The curves show the number of identified molecular operational taxonomic units (MOTUs) of mortierellalean GenBank sequences as a function of the number of authentic strains (a) and species (b) included in our reference data set.

species (Table S1). The resulting 934 sequences grouped into 92 MOTUs at the 97% similarity level, of which 52 contained one or more authentic strains sequenced by us. Of the 92 MOTUs, there were 17 singletons (for a detailed discussion on singletons, see Tedersoo *et al.*, 2010). Because of the excessive divergence within this data set, the multiple alignment consisted of 3675 characters, of which 909 were parsimony informative. Sequential removal of the reference sequences yielded an approximately linear increase in the number of unidentified MOTUs, as shown in Fig 1.

Discussion

In this study, we modeled the effects of increasing the number of type strains in the process of identification of environmental sequences, with the aim of estimating the number of undescribed *Mortierella* species detected by environmental sequencing projects. A plot of the number of identified MOTUs as a function of increasing number of authentic strains included in the data set yielded a nearly linear relationship (Fig. 1). As a result of the presence of synonymous names, saturation of the curve is expected as the number of identified MOTUs approaches 100%. As we know that there are taxonomic synonyms within *Mortierella*, and there is no sign of saturation on our curve, we hypothesize that there are several species not included in our reference data set. However, if a linear fit is assumed, the slope of the curve ($y = 0.540x$) implies that 100% identification of the MOTUs is achieved when the number of authentic strains approaches 170 (Fig. 1a). Because our reference data set is *c.* 25% redundant in that it contains more than one strain per species, the total number of species expected from these figures ($y = 0.721x$) turns out to be 126 (Fig. 1b). This implies that 100% identification of mortierellalean GenBank sequences would necessitate 126 species or 170 authentic strains, respectively, which is 49 species or 68 strains, respectively, more than the numbers that we have sequenced in this study. Because of the nature of our linear approximation, the estimate of 49 missing species certainly includes both described and undescribed species. Despite repeated attempts, however, it has proved impossible to obtain or trace type materials of many species, especially those described by Linnemann (1941).

These figures (126/170) correspond well to the number of species described in *Mortierella*, which suggests that intensifying the sequencing of type materials and authentic strains of already described species can help in the identification of as yet unidentifiable environmental sequences more than previous estimates suggested (Hibbett *et al.*, 2009, 2011). Unfortunately, several described *Mortierella* species lack type materials, or the type is unavailable. As is typical in fungi, an approximation of the number of nonredundantly described species in *Mortierella* is difficult. The most recent monographic studies reported *c.* 90–100 species

in *Mortierella* and 8–13 in *Umbelopsis* (Hawksworth *et al.*, 1995; Kirk *et al.*, 2008; Benny, 2009). In contrast, Index Fungorum (published online by Commonwealth Agricultural Bureaux International (UK) (CABI); http://www.indexfungorum.org) and MycoBank (published online by CBS; http://www.mycobank.org) report 241 and 196 species, respectively. When nomenclatural synonyms are not counted, these numbers decrease to *c.* 157–158 taxa of *Mortierella* and 6–7 taxa of *Umbelopsis*. These figures correspond well to the estimates obtained with our reference data set, suggesting that the vast majority of *Mortierella* and *Umbelopsis* species have already been described. At the same time, our results imply that most 'unidentifiable' environmental sequences in fact represent species already described by taxonomists, but have not been sequenced to date.

These results contrast with the prediction that most of the unidentifiable environmental sequences represent undescribed taxa (Hibbett *et al.*, 2009, 2011). Using *Mortierella*, a group frequently isolated from soil, we inferred that the lack of sequence information from type materials contributes more to the high number of unidentified environmental sequences than do undescribed species. Using our reference data set, for instance, we were able to identify all three mortierellalean MOTUs reported by Buée *et al.* (2009) among the 26 most abundant MOTUs in their forest soil samples, as follows: 'Uncultured Mortierellaceae' (FJ475737; 5253 reads) and 'Uncultured soil fungus sp. 5' (EU807054; 989 reads) were identified as *Mortierella zonata*, and 'Uncultured soil fungus sp. 4' (FJ554362; 2087 reads) was identified as *Mortierella verticillata*. These results have important implications for our views on the number of undescribed fungal species. If the patterns in *Mortierella* observed here could be considered general among fungi, the most widely accepted and cited figure for the number of fungal taxa (1.5 million, implying that > 90% of species have not yet been described; Hawksworth, 2001) would be an overestimation. In this case, the already described species (*c.* 100 000) would make up the majority of extant diversity. Because of the difficulties in objectively comparing the extent of taxonomic knowledge among various groups of fungi, establishing how the figures we inferred for *Mortierella* agree with or depart from the general trend in fungi is problematic. Species of *Mortierella* are easy to isolate and sustain in pure culture, and thus it is easier to obtain DNA from old type materials of *Mortierella*, as compared with, for example, obligate biotrophic fungi, which may make generalization untenable. However, there are only a handful of papers on the taxonomy of *Mortierella* (e.g. Linnemann, 1941; Zycha *et al.*, 1969; Gams, 1977), which has not been completely revised on the basis of molecular phylogenetic methods. Taking these factors together, we suggest that *Mortierella* may be used as an example of general trends in fungi, but the generalizability

of the findings should be tested by performing similar analyses in other groups; for example, those that have been more extensively studied taxonomically and/or unculturable fungi. Nevertheless, our results indicate a potential overestimation of the number of undescribed species and demonstrate the need for the sequencing of reference collections and types.

Bias in the identification procedure can also inflate the number of 'unidentifiable' sequences. In the case of very common but unidentified soil fungi, for instance, the number of unidentified sequences in public databases may be so high that the first match with an exact species name can fall out of the range of BLAST results checked by the researchers. In *Mortierella*, *Mortierella alpina* and *Mortierella verticillata* may be two examples of such a scenario. Both are represented by > 100 ITS sequences in our data set and it can very easily happen that no sequence with a complete species name falls within the first 100 best BLAST hits of a MOTU identification procedure (although it is now possible to exclude environmental sequences from the BLAST results). To avoid this, phylogeny-based MOTU identification involving many reference sequences from type specimens and strains is advisable. Alternatively, third-party annotations of GenBank sequences (as allowed now in UNITE; see Abarenkov *et al.*, 2010) may help to resolve these problems by providing species names for unidentified database sequences.

Of course, we in no way intend to deny the presence of novel lineages or the ability of Sanger-based and next-generation sequencing techniques used in environmental studies to recover them. Two examples are given, among others, by Blaszkowski *et al.* (2009) and Jacobsson & Larsson (2009), where the species were first detected in environmental samples and their sequences were then matched to sequences from voucher specimens after they had been described formally. Certainly, there are thousands of new species awaiting description, but our results suggest that the importance of herbarium voucher specimens should also be emphasized. The importance and contribution of traditional taxonomic practices are beyond question, and they should be incorporated in large-scale environmental sequencing projects, ideally by performing them simultaneously. We expect that the huge increase in the popularity of next-generation sequencing-based environmental sequencing projects will generate (if it has not already done so) an unprecedented demand for the sequencing of robustly identified voucher specimens, or barcoding. Preferably, these specimens should be selected and identified with the assistance of taxonomic experts in a particular group or habitat. We believe that at least the same amount of effort should be put into the generation of sequence data for old type specimens and type strains (or the collection, identification and preservation of new specimens when types are not available) preserved in various culture collections and herbaria as into the deciphering of patterns of unseen fungal diversity in soils, leaf-litter or other habitats. In addition to the hundreds or, more probably, thousands of undescribed species out there, it seems, there are thousands of described species in herbaria and collections awaiting recognition of their importance.

Acknowledgements

References

Abarenkov K, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T *et al.* 2010. The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytologist* 186: 281–285.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. "Basic local alignment search tool". *Journal of Molecular Biology* 215: 403–410.

Amend AS, Seifert KA, Samson R, Bruns TD. 2010. Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. *Proceedings of the National Academy of Sciences, USA* 107: 13748e13753.

Benny GL. 2009. *Zygomycetes*. published on the Internet at [WWW document]. URL http://www.zygomycetes.org [accessed on 17 July 2010].

Bidartondo MI, Ameri G, Döring H. 2009. Closing the mycorrhizal DNA sequence gap. *Mycological Research* 113: 1025–1026.

Bidartondo MI, Bruns TD, Blackwell M, Edwards I, Taylor AFS, Horton T, Zhang N, Kõljalg U, May G, Kuyper TW *et al.* 2008. Preserving accuracy in GenBank. *Science* 319: 1616.

Blaszkowski J, Kovács GM, Balázs T. 2009. *Glomus perpusillum*, a new arbuscular mycorrhizal fungus. *Mycologia* 101: 247–255.

Brock PM, Döring H, Bidartondo MI. 2008. How to know unknown fungi: the role of a herbarium. *New Phytologist* 181: 719–724.

Buée M, Reich M, Murat C, Morin E, Nilsson RH, Uroz S, Martin F. 2009. 454 pyrosequencing analyses of forest soils reveal an unexpected high fungal diversity. *New Phytologist* 184: 449–456.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

Gams W. 1977. A key to the species of *Mortierella*. *Persoonia* 9: 381–391.

Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, Naqvi A, Gillevet PM. 2010. Characterization of the oral fungal microbiome (Mycobiome) in healthy individuals. *PLoS Pathogens* 6: e1000713.

Guindon S, Gascuel O. 2003. PhyML: a simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52: 696–704.

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

Hawksworth DL, Kirk PM, Sutton BC, Pegler DN. 1995. *Ainsworth and Bisby's dictionary of fungi, 8th edn*. Wallingford, UK: CABI.

Hibbett DS, Ohman A, Glotzler D, Nuhn M, Kirk PM, Nilsson RH 2011. Progress in molecular and morphological taxon discovery in *Fungi* and options for formal classification of environmental sequences. *Fungal Biology Reviews* 25: 38–47.

Hibbett DS, Ohman A, Kirk PM. 2009. Fungal ecology catches fire. *New Phytologist* **184**: 279–282.

Jacobsson S, Larsson E. 2009. *Inocybe spuria*, a new species in section *Rimosae* from boreal coniferous forests. *Mycotaxon* **109**: 201–207.

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

Jumpponen A, Jones KL, Blair J. 2010. Vertical distribution of fungal communities in tallgrass prairie soil. *Mycologia* **102**: 1027.

Kirk PM, Cannon PF, Minter DW, Stalpers JA. 2008. *Dictionary of the fungi, 10th edn.* Wallingford, UK: CABI.

Kõljalg U, Larsson K-H, Abarenkov K, Nilsson RH, Alexander IJ, Eberhardt U, Erland S, Hoiland K, Kjoller R, Larsson E et al. 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist* **166**: 1063–1068.

Linnemann G. 1941. Die Mucorineen-Gattung *Mortierella* Coemans. *Pflanzenforschung* **23**: 1–64.

Lumini E, Orgiazzi A, Borriello R, Bonfante P, Bianciotto V. 2009. Disclosing arbuscular mycorrhizal fungal biodiversity in soil through a land-use gradient using a pyrosequencing approach. *Environmental Microbiology* **12**: 2165e2179.

Mueller GM, Schmit JP. 2007. Fungal biodiversity: what do we know? What can we predict? *Biodiversity and Conservation* **16**: 1–5.

Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson KH. 2008. Intraspecific ITS variability in the Kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary Bioinformatics* **4**: 193–201.

Nilsson RH, Ryberg M, Kristiansson E, Abarenkov K, Larsson KH. 2006. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS ONE* **1**: e59.

O'Brien HE, Parrent JL, Jackson JA, Moncalvo JM, Vilgalys R. 2005. Fungal community analysis by large-scale sequencing of environmental samples. *Applied and Environmental Microbiology* **71**: 5544–5550.

Öpik M, Metsis M, Daniell TJ, 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.

Rousk J, Baath E, Brookes PC, Lauber CL, Lozupone C, Caporaso JG, Knight R, Fierer N. 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME Journal* **4**: 1340e1351.

Roshan U, Livesay DR. 2006. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* **22**: 2715–2721.

Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH. 2008. 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.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**: 7537–7541.

Staden R, Beal KF, Bonfield JK. 2000. The Staden Package. *Methods in Molecular Biology* **132**: 115–130.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Swofford DL. 2003. *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.* Sunderland, MA, USA: Sinauer Associates.

Tedersoo L, Nilsson RH, 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, Brandstrom DM, Lindahl BD. 2010. Production of ectomycorrhizal mycelium peaks during canopy closure in Norway spruce forests. *New Phytologist* **187**: 1124–1134.

White TJ, Bruns T, Lee S, Taylor JW. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ, eds. *PCR protocols: a guide to methods and applications.* New York, NY, USA: Academic Press, 315–322.

Zycha H, Siepmann R, Linnemann G. 1969. *Mucorales. Eine Beschreibung aller Gattungen und Arten dieser Pilzgruppe.* Lehre, Germany: J Cramer.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Maximum likelihood phylogeny inferred from the 934-sequence data set.

Table S1 Name, strain origin, identifier and GenBank accession numbers of the sequences used in the study

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.