Received Date : 10-May-2013
Revised Date : 08-Jul-2013
Accepted Date : 17-Jul-2013
Article type : Opinion

Towards a unified paradigm for sequence-based identification of Fungi

Authors

Urmas Kõljalg[1, 2], R. Henrik Nilsson[3], Kessy Abarenkov[2], Leho Tedersoo[2], Andy FS Taylor[4, 5], Mohammad Bahram[1], Scott T. Bates[6], Thomas D. Bruns[7], Johan Bengtsson-Palme[8], Tony Martin Callaghan[9], Brian Douglas[9], Tiia Drenkhan[10], Ursula Eberhardt[11], Margarita Dueñas[12], Tine Grebenc[13], Gareth W. Griffith[9], Martin Hartmann[14, 15], Paul M. Kirk[16], Petr Kohout[1, 17], Ellen Larsson[3], Björn D. Lindahl[18], Robert Lücking[19], María P. Martín[12], P. Brandon Matheny[20], Nhu H. Nguyen[7], Tuula Niskanen[21], Jane Oja[1], Kabir G. Peay[22], Ursula Peintner[23], Marko Peterson[1], Kadri Põldmaa[1], Lauri Saag[1], Irja Saar[1], Arthur Schüßler[24], James A. Scott[25], Carolina Senés[24], Matthew E. Smith[26], Ave Suija[1], D. Lee Taylor[27], M. Teresa Telleria[12], Michael Weiß[28], Karl-Henrik Larsson[29].

1 Institute of Ecology and Earth Sciences, University of Tartu, Lai 40, 51005 Tartu, Estonia.

2 Natural History Museum, University of Tartu, Vanemuise 46, 51014 Tartu, Estonia.

3 Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, SE-40530 Göteborg, Sweden.

4 The James Hutton Institute, Craigiebuckler, Aberdeen, AB15 8QH, Scotland, UK.

5 Institute of Biological and Environmental Sciences, University of Aberdeen, Cruickshank Building, St Machar Drive, Aberdeen, AB24 3UU, UK.

6 Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309 Colorado, USA.

7 Department of Plant and Microbial Biology, University of California, Berkeley California, 94720, USA.

8 Department of Neuroscience and Physiology, The Sahlgrenska Academy, University of Gothenburg, Box 434, SE-40530 Göteborg, Sweden.

9 Institute of Biological, Environmental and Rural Sciences, Cledwyn Building, Aberystwyth University, SY23 3DD Aberystwyth, Wales, UK.

10 Institute of Forestry and Rural Engineering, Estonian University of Life Sciences, Fr. R. Kreutzwaldi 5, 51014 Tartu, Estonia.

11 Staatliches Museum f. Naturkunde Stuttgart, Abt. Botanik, Rosenstein 1, D-70191 Stuttgart, Germany.

12 Departmento de Micología, Real Jardín Botánico (RJB-CSIC), Plaza de Murillo 1, 28014 Madrid, Spain.

13 Department of Forest Physiology and Genetics, Slovenian Forestry Institute, Vecna pot 2, SI-1000 Ljubljana, Slovenia.

14 Forest Soils and Biogeochemistry, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland.

15 Molecular Ecology, Agroscope Reckenholz-Tänikon Research Station ART, Zurich, Switzerland.

16 Mycology Section, Jodrell Laboratory, Royal Botanic Gardens Kew, Surrey TW9 3DS, UK.

17 Institute of Botany, Academy of Science of the Czech Republic, CZ-252 43 Průhonice, Czech Republic.

18 Swedish University of Agricultural Sciences, Department of Forest Mycology and Plant Pathology, Box 7026, SE-75007 Uppsala, Sweden.

19 Department of Botany, The Field Museum, 1400 South Lake Shore Drive, Chicago, IL 60605-2496, USA.

20 Department of Ecology and Evolutionary Biology, University of Tennessee, Hesler Biology Building 332, Knoxville, TN 37996-1610, USA.

21 Plant Biology, Department of Biosciences, University of Helsinki, P.O. Box 56, FI-00014 Helsinki, Finland.

22 Department of Biology, Stanford University, CA 94305 Stanford, USA.

23 Institute of Microbiology, University Innsbruck, Technikerstr. 25, 6020 Innsbruck, Austria.

24 Genetics, Department Biology, Ludwig-Maximilians-University, Munich, Grosshaderner Str. 4, 82152 Martinsried, Germany.

25 Dalla Lana School of Public Health, University of Toronto, 223 College Street, Toronto ON, M5T 1R4 Canada.

26 Department of Plant Pathology, University of Florida, Gainesville, FL 32611-0680, USA.

27 Department of Biology, MSC03 2020, University of New Mexico, Albuquerque, NM 87131-0001, USA.

28 Department of Biology, University of Tübingen, Auf der Morgenstelle 5, D-72076 Tübingen, Germany.

29 Natural History Museum, University of Oslo, P.O. Box 1172 Blindern, 0318 Oslo, Norway.

Keywords: microbial diversity, DNA barcoding, ecological genomics, bioinformatics, fungi.

Corresponding author: Urmas Kõljalg, Institute of Ecology and Earth Sciences, University of Tartu, Lai 40, 51005 Tartu, Estonia. E-mail: urmas.koljalg@ut.ee

Running title: Unified paradigm for identification of Fungi

Abstract

The nuclear ribosomal internal transcribed spacer (ITS) region is the formal fungal barcode and in most cases the marker of choice for exploration of fungal diversity in environmental samples. Two problems are particularly acute in the pursuit of satisfactory taxonomic assignment of newly generated ITS sequences: (i) the lack of an inclusive, reliable public reference dataset, and (ii) the lack of means to refer to fungal species, for which no Latin name is available in a standardized stable way. Here we report on progress in these regards through further development of the UNITE database (http://unite.ut.ee) for molecular identification of fungi. All fungal species represented by at least two ITS sequences in the international nucleotide sequence databases are now given a unique, stable name of the accession number type (e.g., *Hymenoscyphus pseudoalbidus*|GU586904|SH133781.05FU), and their taxonomic and ecological annotations were corrected as far as possible through a distributed, third-party annotation effort. We introduce the term "species hypothesis" (SH) for the taxa discovered in clustering on different similarity tresholds (97-99%). An automatically or manually designated sequence is chosen to represent each such species hypothesis. These reference sequences are released (http://unite.ut.ee/repository.php) for use by the scientific community in, e.g., local sequence similarity searches and in the QIIME pipeline. The system and the data will be updated automatically as the number of public fungal ITS sequences grows. We invite everybody in the position to improve the annotation or metadata associated with their particular

fungal lineages of expertise to do so through the new web-based sequence management system in UNITE.

Introduction

The nuclear ribosomal internal transcribed spacer (ITS) region has a long history of use as a molecular marker for species-level identification in ecological and taxonomic studies of Fungi (Hibbett et al. 2011). It offers several advantages over other species-level markers in terms of high information content and ease of amplification, and it was recently designated the official barcode for fungi (Schoch et al. 2012). The publicly available fungal ITS sequences vary significantly in reliability and technical quality, however, and third party annotation is not currently allowed (Bidartondo et al. 2008). To facilitate ITS-based molecular identification of fungi for the scientific community, the first fungal ITS annotation workshop was held on the premises of the University of Tartu, Estonia, on January 29-30, 2013. The 28 physical and online participants were chiefly fungal taxonomists whose expertise covered various lineages of Ascomycota, Basidiomycota, Glomeromycota, and Neocallimastigomycota. The researchers also comprised bioinformaticians and molecular ecologists with experience in sequence quality assessment. The workshop centered on the annotation of fungal ITS sequences in the extended UNITE database (http://unite.ut.ee; Abarenkov et al. 2010a) through the web-based sequence management workbench PlutoF (Abarenkov et al. 2010b; see also Fig. 1). Because UNITE mirrors the fungal ITS sequences in the International Nucleotide Sequence Databases (INSD: GenBank, EMBL, and DDBJ), the full set of ca. 300,000 fungal ITS entries generated by the scientific community as of December 2012 served as the target dataset.

The first version of the UNITE database was released in 2003 with a focus on ITS sequences of ectomycorrhizal fungi in northern Europe (Kõljalg et al. 2005). The database has been under continuous development since then and has become a full-blown sequence management environment with analysis and storage modules. At present, UNITE targets all fungi and geographical regions, but the founding principle – to provide reliable reference sequences for molecular identification – remains the same. Hereafter, UNITE refers not only to the original database of annotated ectomycorrhizal sequences, but also encompasses all fungal ITS sequences in the INSD database that are not of poor quality. The demand for high-quality reference sequences has risen rapidly due to the increasing use of high-throughput sequencing technologies (such as 454 pyrosequencing, Illumina, and Ion Torrent; Glenn 2011; Shokralla et al. 2012; Bates et al. 2013). These approaches generate vast amounts of sequence data – hundreds of thousands to billions of reads within a few hours or days – such that various automated approaches to analysis represent the only viable option of handling the data. Several software pipelines are available for overseeing more or less the entire analysis procedure, from data cleaning to sequence clustering and taxonomic assignment (e.g., QIIME: Caporaso et al. 2010; MOTHUR: Schloss et al. 2009; Lindahl et al. 2013). However, satisfactory taxonomic identification remains problematic in the kingdom Fungi due to the vast, largely unexplored diversity and the lack of reliable and richly annotated reference sequences.

The ~300,000 public fungal ITS sequences constitute a poor candidate for the basis of taxonomic annotation of newly generated sequences, especially when used in conjunction with fully automated pipelines. Only about half of these sequences are annotated to the level of species (Schoch et al. 2012). This half represents roughly 20,000 different species (Latin

binomials), which corresponds to 0.2-4.5% of the estimated 0.5-10 million extant fungal species (Blackwell 2011; Bass & Richards 2011). More than 10% of the public, fully identified fungal ITS sequences have been shown to be incorrectly annotated at the species level, making uncritical use of this dataset problematic (Nilsson et al. 2006). Among the 50% of entries not annotated to species level, many correspond to species that are not yet formally described. There is no unified way to refer to such species, and different researchers adopt different, ad hoc naming systems to such taxa compromising comparability over studies and time (Ryberg et al. 2008). Many of the entries furthermore suffer from quality issues such as low read quality or chimeric unions. Thus, both data structuring and filtering are needed to make the dataset a useful tool for annotation of new sequences.

To generate a concise set of reference sequences, UNITE applies a two-tier clustering process, first clustering all sequences to roughly the subgenus/genus level, and then to approximately the species level (Fig. S1). Both levels represent operational taxonomic units (OTUs) as defined in Sokal & Sneath (1963) and Blaxter et al. (2005), but here we introduce the term "species hypotheses" (SH) for the taxa arising from the second round of clustering. A SH is normally composed of two or more sequences to avoid excessive inflation of SHs due to singleton sequences of substandard technical quality, but users can sanction individual singleton sequences to serve as SHs. A representative sequence for each species hypotheses is chosen automatically by computing the consensus sequence of the SH and then finding the best matching sequence of the SH (Fig. S1). Taxonomic experts may override the choice of representative sequence by designating a reference sequence based on type status, source of isolation, and sequence quality (Fig. S2). Thus, all species hypotheses have either an

automatically chosen representative sequence or a manually designated reference sequence. These representative and reference sequences are released (http://unite.ut.ee/repository.php) as a reference dataset for local sequence similarity searches as well as high-throughput sequencing bioinformatics platforms including the QIIME pipeline (Fig. S3). An annotation-aware FASTA file with all UNITE/INSD fungal ITS sequences not known to be of poor quality is also maintained at the same URL.

The species hypotheses can be viewed and edited in a web browser through the PlutoF workbench (Fig. 1, Fig. S4, and Fig. S5). Viewing sequence data by eye in the form of a multiple sequence alignment is a powerful means both to spot meaningful patterns in the data and to detect sequences of substandard quality or insufficient/incorrect annotation. Implementing changes in response to such observations in PlutoF involves only a few mouse clicks (Fig. S2). The user also has the opportunity to re-designate a representative sequence for any species hypothesis.

During the workshop we targeted four aspects of sequence reliability and annotation: 1) selection of reference sequences; 2) improving/adding taxonomic annotations; 3) improving/adding taxonomic and ecological metadata; and 4) tagging (and thus excluding) sequences of compromised technical quality.

1. Selection of representative and reference sequences

The automated choice of representative sequences in UNITE is based on nucleotide frequency and hence the sequence most similar to the consensus becomes representative. Although this approach is intuitively appealing and logical in most situations, there are some potential

drawbacks. For example, a single specimen may have been sequenced several times (including cloned samples), or some particular study may have exhausted a limited geographical region for records of a single species. The special authoritative standing of type specimens in systematics similarly gives rise to the desire to re-designate representative sequences on a regular basis (cf. Hyde and Zhang 2008). Not all sequences from type specimens (hereinafter "type sequences") form ideal reference sequences though. From a bioinformatics point of view, an ideal representative sequence should cover the full ITS region and should ideally not feature many IUPAC DNA ambiguity symbols (Cornish-Bowden 1985) or manifest signs of a potentially compromised technical/read quality-related nature (cf. Nilsson et al. 2012). Type specimens, in contrast, might be tens to hundreds years old, making it difficult to obtain long, high-quality DNA sequences (Larsson and Jacobsson 2004).

For these reasons we re-examined the representative sequences for species hypotheses for which we have taxonomic expertise and manually re-designated a reference sequence whenever relevant (see Fig. 1). In the absence of (high-quality) type sequences, we sought to designate a sequence that originated from the same country or geographical region as the type specimen. Sequences from vouchered fruiting bodies and living cultures were preferred over uncloned sequences from other sources (e.g., root tips and sclerotia) that in turn were given priority over cloned sequences from various complex environmental substrates where vouchering typically proves impossible. We sought to make sure that the automatically chosen representative had the most accurate taxonomic annotation possible. For example, when the automatic procedure had selected a sequence annotated as "uncultured fungus" for a species for which the name of lower taxonomic levels (genus to phylum) was available, we made the appropriate re-annotation. We

also re-annotated sequences by providing a more conservative name if the species name given by the original sequence authors did not accurately reflect recent results and findings (e.g., a misidentified *Hymenoscyphus albidus* would be annotated as *Hymenoscyphus sp*., Helotiales, or Ascomycota depending on the severity of the misannotation). In recognition of the fact that no single sequence similarity threshold value - such as 97% - will demarcate intraspecific from interspecific variability in all fungi, reference sequences were set at the level they made taxonomic sense based on the results of previous studies. Many *Cortinarius* species hypotheses were, accordingly, specified at the 99% similarity level; many lichenized fungi, in contrast, were set at the 97% similarity level.

2. Improving/adding taxonomic annotation

UNITE follows the Index Fungorum (http://www.indexfungorum.org) nomenclature of fungi. Approximately 84% of the sequences in UNITE are assigned at least to ordinal level, but sequences annotated as, e.g., “uncultured fungus” are assigned only at the kingdom level. If the user assigns such a sequence at a lower taxonomic level such as genus, the sequence will adopt the full hierarchical classification leading up to that genus, typically phylum, order, and family. When examining the species hypotheses, we adjusted the taxonomic annotation of the reference and representative sequences. A genus or order name was added to most sequences originally named, e.g.,“cf. *Athelia*” or “uncultured fungus”; this was only done for taxa with which we were sufficiently familiar.

3. Improving/adding metadata

Concurrent with the process of taxonomic annotation of sequences, we added relevant metadata such as type status, voucher specimen/culture, country of origin, and host/substrate of collection. In most cases this involved manual extraction of data from publications and sometimes contacting the original authors of the sequences.

4. Excluding sequences of compromised technical quality

Based on the PlutoF multiple sequence alignments, we checked the sequences for substandard quality in terms of chimeric nature and read reliability following Tedersoo et al. (2011) and Nilsson et al. (2012). During the workshop we also made an effort to find additional chimeras using UCHIME v. 6.0.307 (Edgar et al. 2011). As a reference dataset we used all representative/reference sequences from the UNITE species hypotheses. We ran the full UNITE sequence set through UCHIME using its reference mode and then subjected sequences that exceeded the default threshold at which UCHIME considers a sequence chimeric to further scrutiny through BLAST and occasionally also through multiple sequence alignment. Sequences that were clearly unreliable or overly short were marked as such in UNITE. While all sequences marked as substandard remain searchable in the database, they are removed from BLAST searches in UNITE, the UNITE global key, and the releases of representative/reference sequences.

Results and Discussion

Our efforts resulted in approximately 5,300 manual changes to the corpus of public fungal ITS sequences in UNITE (Fig. 2). A full 1,860 of these represented re-designations of representative

sequences into reference sequences (317 of which into type sequences). This means that 3.5% of the 52,481 species hypotheses at the 98% similarity level now have a manually designated reference sequence. We implemented over 2,578 taxonomic annotations and re-annotations at the species and higher taxonomic levels. 248 sequences were excluded for being chimeric or of low quality in other regards. Finally, we added 654 items of metadata to the sequence data. It is clear that this is only the tip of the iceberg, though, and much remains to be done in all fungal phyla and the lineages covered by the present set of authors. In addition, new sequences are generated and being deposited in INSD and UNITE at an exponential rate, such that annotation efforts will always lag behind. The UNITE/PlutoF system offers third-party annotation capacities to all its registered users (Abarenkov et al. 2010b). Thus, we invite all fungal biologists to participate. In particular, we hope that all fungal taxonomists and ecologists will examine their lineages of expertise in UNITE and make sure that relevant sequences are chosen to represent species hypotheses, and that the sequences are annotated to a satisfactory level in terms of taxonomy and ecology.

The issue of naming DNA-based taxa in ecological and taxonomical studies has been debated for a long time (Hibbett & Taylor 2013). Studies that identify unknown DNA from biological samples typically apply their own ad hoc naming system (e.g., "*Tulasnella sp*. 14"; see Ryberg et al. 2008), which is certain to be different from that adopted by other researchers. This makes comparison among studies complicated if not impossible. Therefore we implemented an automated, all-inclusive naming system for SHs found at various sequence similarity threshold values. The name of the SH is based on the reference or representative sequence and compiled automatically from three data fields. First is the taxonomic name of the sequence, viz. species,

genus, family, or higher-level name. The next field is the INSD or UNITE accession code of the sequence, and the third field is the SH accession code. Thus the name of the SH causing ash-dieback shown in Fig.1 is "*Hymenoscyphus pseudoalbidus*|GU586904|SH133781.05FU" and its sister SH "*Hymenoscyphus albidus*|GU586876|SH114093.05FU". In contrast to names of the "*Tulasnella* sp. 14" type, this allows for exact communication across scientific studies and time. Names in this format allow anybody to visit the same SH years later and if feasible to reproduce identification analyses based on new versions of the key. It is also easy to hyperlink those names in publication to the SH and associated information (see Fig. S3). Unique SH accession codes are generated automatically for all species hypotheses at all similarity cut-off levels. The accession code begins with SH (acronym for the Species Hypothesis) and a unique six-digit number followed by period, a two-digit version number (version number of the key), and FU (acronym for Fungi). The version number allows to place the SHs in time, and the two-letter acronym of the taxon enables quick placement of the SH in the full eukaryote classification. This would be highly useful feature if the same platform will be used for other kingdoms too.

We hope that the present effort will lead to improved taxonomic accuracy and resolution of species hypotheses for biologists using the UNITE database, the standalone FASTA files of UNITE, and the QIIME pipeline. Taxonomic precision and availability of rich metadata are clearly among the most important goals from an ecological perspective. After all, a growing number of non-mycologists now study fungi as a part of their scientific pursuit (Pautasso 2013), and it is imperative that we provide them with state-of-the-art data since they may not always be in a position to discriminate good data from bad data. For example, fully annotated ITS sequences facilitate global-scale meta-studies on phylogeny, evolutionary ecology, and

biogeography (Bonito et al. 2010; Veldre et al. in press). Taxonomic precision facilitates distinguishing of emerging pathogens such as *Hymenoscyphus pseudoalbidus* from their non-pathogenic close relatives (Fig. 1). Rapid and precise identification of pathogenic organisms forms a basis for efficient countermeasure, which is particularly relevant for forest, agricultural and human diseases. Arriving at the best and richest possible set of reference sequences is, however, not a question of bioinformatics or computational power but rather one of taxonomic and ecological expertise.

Acknowledgements

We thank Estonian research infrastructure roadmap NATARC for the hosting UNITE database. The North European Forest Mycologists network is acknowledged for support. Urmas Kõljalg and Kessy Abarenkov are supported by the Estonian Research Council grant no 8235.

References

Abarenkov K, Nilsson RH, Larsson K-H et al. 2010a. The UNITE database for molecular identification of fungi - recent updates and future perspectives. New Phytologist 186: 281-285.

Abarenkov K, Tedersoo L, Nilsson RH et al. 2010b. PlutoF - a web-based workbench for ecological and taxonomical research, with an online implementation for fungal ITS sequences. Evolutionary Bioinformatics 6: 189-196.

Bates S.T., S. Ahrendt, H. Bik, T. Bruns, J.G. Caporaso, J. Cole, M. Dwan, N. Fierer, D. Gu, S. Houston, R. Knight, J. Leff, C. Lewis, D. McDonald, H. Nilsson, A. Porras-Alfaro, V. Robert, C.

Schoch, J. Scott, L. Taylor, L. Wegener-Parfrey, and J.E. Stajich. 2013. Meeting report: Fungal ITS workshop (October 2012). Standards in Genomic Sciences 8: 1 (doi:10.4056/sigs.3737409).

Bidartondo M, Bruns TD, Blackwell M et al. 2008. Preserving accuracy in GenBank. Science 319: 1616.

Bass D, Richards TA. 2011. Three reasons to re-evaluate fungal diversity "on Earth and in the ocean". Fungal Biology Reviews 25: 159-164.

Blackwell M. 2011. The Fungi: 1, 2, 3 … 5.1 million species? American Journal of Botany 98: 936-948.

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E. 2005. Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society of London Series B – Biological Sciences 360: 1935–1943.

Bonito GM, Gryganskyi AP, Trappe JM, Vilgalys R. 2010. A global meta-analysis of *Tuber* ITS rDNA sequences: species diversity, host associations and long-distance dispersal. Molecular Ecology 19: 4994-5008.

Caporaso JG, Kuczynski J, Stombaugh J et al. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7: 335-336.

Cornish-Bowden A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Research 13: 3021–3030.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27: 2194-2200.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11: 759–769.

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: 38–47.

Hibbett DS, Taylor JW. 2013. Fungal systematics: is a new age of enlightenment at hand? Nature Reviews Microbiology. doi:10.1038/nrmicro2942

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

Kõljalg U, Larsson K-H, Abarenkov K, Nilsson RH, Alexander IJ, Eberhardt U, Erland S, Hoiland K, Kjoller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Vralstad T, Ursing BM. 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytologist 166: 1063–1068.

Larsson E, Jacobsson S. 2004. Controversy over *Hygrophorus cossus* settled using ITS sequence data from 200 year-old type material. Mycological Research 108: 781–786.

Lindahl, B.D., Nilsson, R.H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjøller, R., Koljalg, U., Pennanen, T., Rosendahl, S., Stenlid, J. & Kauserud, H. 2013. Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. New Phytologist: doi: 10.1111/nph.12243.

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

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.

Queloz V, Grünig CR, Berndt R, Kowalski T, Sieber TN, Holdenrieder O. 2011. Cryptic speciation in *Hymenoscyphus albidus*. Forest Pathology 41: 133-142.

Pautasso M. 2013. Fungal under-representation is (slowly) diminishing in the life sciences. Fungal Ecology 6: 129-135.

Ryberg M, Nilsson RH, 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.

Schoch CL, Seifert KA, Huhndorf A 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: 6241–6246.

Schloss PD, Westcott SL, Ryabin T 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.

Shokralla S, Spall JL, Gibson JF, Hajibabaei M. 2012. Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21: 1794–1805.

Sokal RR, Sneath PHA. 1963. Principles of numerical taxonomy. San Francisco, W. H. Freeman & Co.

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.

Veldre V, Abarenkov K, Bahram M, Martos F, Selosse M-A, Tamm H, Kõljalg U, Tedersoo L. 2013. Evolution of nutritional modes of Ceratobasidiaceae (Cantharellales, Basidiomycota) as revealed from publicly available ITS sequences. Fungal Ecology, in press.

Figures

Fig.1. Screenshot of the UNITE global key workbench depicting one of the 7470 genus/subgenus-level clusters. This cluster contains five species hypotheses covering the well-known *Hymenoscyphus pseudoalbidus*, the causal agent of ash-dieback disease, its non-pathogenic sister species *H. albidus*, and other closely related taxa. The workbench enables the users to annotate individual sequences with taxonomic and ecological metadata and to determine a reference sequence for each species hypothesis at different sequence similarity cutoff levels that represent hierarchical structures among these sequences and taxa. A reference sequences provides a proxy for the species hypothesis at user-defined cut-off levels. The coloured squares in the column DSH are for the visualization of inclusiveness of species hypotheses (SHs) at five different cut-off levels (from left to right 99%; 98.5%; 98%; 97.5%, and 97% similarity). Reference sequences of species hypotheses chosen by an expert are indicated by circles. In this example, *H. pseudoalbidus* (green squares) and *H. albidus* (grey squares) fall into a single SH at 97.5% and lower sequence similarity. The reference sequence of *H. albidus* is used for the naming of SHs in these levels, because it has nomenclatural priority over *H. pseudoalbidus* that was described later (Queloz et al. 2011). Therefore, all sequences of these two SHs are indicated in grey at 97.5 and 97% cut off values. It is up to the researcher to decide which cut-off values are used for identification in ecological studies. Names of SHs in publications can be hyperlinked to the cluster of sequences supplemented with metadata. The system enables saving identification results of ecological studies in a standardized and reproducible manner. The name of the SH is based on the reference or representative sequence and is compiled automatically from three data fields, viz. the taxonomic name of the sequence, the INSD or UNITE accession number of the sequence, and the SH accession code. For the full description of the workbench

and annotation guidelines, see Supplementary Materials. In this figure, 115 sequences of this cluster were removed for better visualization. The full cluster is illustrated in Fig. S4.

Fig.2. The statistics of the UNITE global key. Table (a) shows the number of SHs based on a 98% threshold value, the number of ITS sequences in the current version of UNITE which passed through the quality filters, and other associated statistics. The high number of unspecified sequences and SHs that lack information on locality (over 40%) illustrate the need for annotation effort. Circle graphs (b) illustrate the geographic distribution of those SHs that occur on more than one continent. North America, Europe, and Asia are more similar to each other compared to other continents. The comparatively high number of shared SHs between Southern and Northern hemisphere continents mark potential invasions that call for fine scale ecological studies (Antarctica has too few ITS sequences to make any sensible comparison). Table (f) provides the number of SHs for five different sequence similarity threshold values. It demonstrates how the selection of a threshold value may influence the results of studies. The new version of UNITE makes studies that employ different threshold values comparable and reproducible. Table (e) shows the number of SHs and reference sequences per fungal phylum. The Basidiomycota and Glomeromycota are the most annotated phyla, reflecting the current composition of experts. Four phyla that have the smallest number of SHs are probably underrepresented in INSD databases because of difficulties to culture those fungi or find tangible reproductive/somatic structures. The graph of the subfigure (d) shows that the numbers of fungal ITS sequences in INSD and UNITE are growing much faster than the number of SHs. This is probably biased because most sequences are still coming from North America, Europe, and Asia. Potentially species-rich regions in the Southern hemisphere are much less well represented (see also (a)). To investigate

the fungal sequencing effort at the global scale, we generated rarefied curves demonstrating the number of SHs detected versus the number of sequences at three similarity threshold levels, viz. 97, 98, and 99% (c).

Abbreviations: SH – species hypothesis; RefSeq – reference sequence

Fig.S1. Generation of global key: technical description.

Fig.S2. Guidelines for annotating and choosing reference sequences.

Fig.S3. Format of the UNITE reference sequences FASTA file available for download at unite.ut.ee and used by QIIME

Fig.S4. Screenshot of the UNITE global key workbench depicting the cluster UCL5_005639.

Fig.S5. Screenshot of the UNITE global key workbench depicting the species hypothesis SH155686.05FU. This workbench enables the selection of reference sequences.

Data accessibility

FASTA files of the annotated UNITE + INSD datasets are available:

1. for download at http://unite.ut.ee/repository.php;
2. integrated into QIIME software package for comparison and analysis of fungal communities (qiime.org).

Author contributions

Idea and design: U Kõljalg, H Nilsson, K Abarenkov, K-H Larsson

Software development and analyses: K Abarenkov, H Nilsson, M Bahram

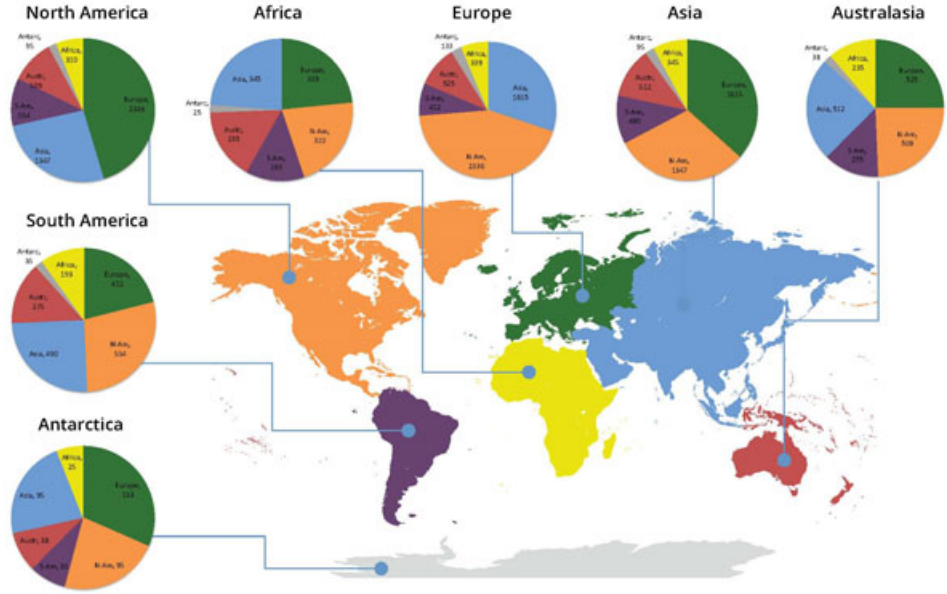
Performed research: all authors

Writing the paper: H Nilsson, U Kõljalg, L Tedersoo, K Abarenkov, K-H Larsson

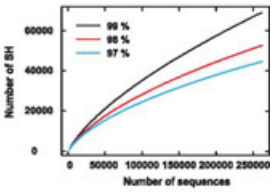
(a)

Statistics based on 98% threshold value	Total	Europe	Asia	N-America	S-America	Australasia	Antarctica	Africa	Unspecified
No of SHs*	52481	13779	8819	12941	2929	3417	313	2425	20294
No of ITS sequences	261521	50963	31560	53931	9635	8211	930	6183	100108
SH shared between continents		2832	1956	2633	686	745	130	507 N/A	
Percent of unique SHs		79.5%	78%	79.6%	76.6%	78.2%	58.5%	79.1%	
No of singletons (out. clusters)	4130	694	569	742	208	266	13	182	1456
No of singletons (inside clusters)	28395	5854	3730	5825	1183	1587	97	1006	9113
* Including all singletons									

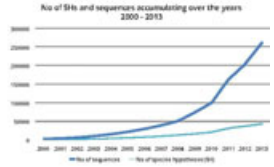
(b)



(c)



(d)



(e)

Phylum	No of SHs*	No of RefSeq
Basidiomycota	20834	1476
Ascomycota	20754	287
Chytridiomycota	285	0
Glomeromycota	1658	77
Zygomycota	602	0
Neocallimastigomycota	85	2
Blastocladiomycota	31	0
Incertae sedis	9	0
Unspecified	11387	0
* Based on 98% threshold value		

(f)

Threshold value	Number of SHs
97%	44537
97.5%	48007
98%	52481
98.5%	58594
99%	68938