## *Finding needles in haystacks*

Schoch, Conrad L.; Robbertse, Barbara; Robert, Vincent; Vu, Duong; Cardinali, Gianluigi; Irinyi, Laszlo; Meyer, Wieland; Nilsson, R. Henrik; Hughes, Karen W.; Miller, Andrew N.; Kirk, Paul M.; Abarenkov, Kessy; Aime, M. Catherine; Ariyawansa, Hiran A.; Bidartondo, Martin.; Boekhout, Ten; Buyck, B; Cai, Qing; Chen, Jie; Crespo, Ana

*Published in:*
Database: The Journal of Biological Databases and Curation

*DOI:*
10.1093/database/bau061

*Publication date:*
2014

*Citation for published version (APA):*
Schoch, C. L., Robbertse, B., Robert, V., Vu, D., Cardinali, G., Irinyi, L., Meyer, W., Nilsson, R. H., Hughes, K. W., Miller, A. N., Kirk, P. M., Abarenkov, K., Aime, M. C., Ariyawansa, H. A., Bidartondo, M., Boekhout, T., Buyck, B., Cai, Q., Chen, J., ... Federhen, S. (2014). Finding needles in haystacks: Linking scientific names, reference specimens and molecular data for Fungi. *Database: The Journal of Biological Databases and Curation*, (2014), 1-21. https://doi.org/10.1093/database/bau061

DATABASE
The Journal of Biological Databases and Curation

*Database*, 2014, 1–21
doi: 10.1093/database/bau061
Original article

Original article

# Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi

**Conrad L. Schoch[1,*,†], Barbara Robbertse[1,†], Vincent Robert[2], Duong Vu[2], Gianluigi Cardinali[3], Laszlo Irinyi[4], Wieland Meyer[4], R. Henrik Nilsson[5], Karen Hughes[6], Andrew N. Miller[7], Paul M. Kirk[8], Kessy Abarenkov[9], M. Catherine Aime[10], Hiran A. Ariyawansa[11], Martin Bidartondo[12], Teun Boekhout[2], Bart Buyck[13], Qing Cai[14], Jie Chen[11], Ana Crespo[15], Pedro W. Crous[2], Ulrike Damm[16], Z. Wilhelm De Beer[17], Bryn T. M. Dentinger[8], Pradeep K. Divakar[15], Margarita Dueñas[18], Nicolas Feau[19], Katerina Fliegerova[20], Miguel A. García[21], Zai-Wei Ge[14], Gareth W. Griffith[22], Johannes Z. Groenewald[2], Marizeth Groenewald[2], Martin Grube[23], Marieka Gryzenhout[24], Cécile Gueidan[25], Liangdong Guo[26], Sarah Hambleton[27], Richard Hamelin[19], Karen Hansen[28], Valérie Hofstetter[29], Seung-Beom Hong[30], Jos Houbraken[2], Kevin D. Hyde[11], Patrik Inderbitzin[31], Peter R. Johnston[32], Samantha C. Karunarathna[11], Urmas Kõljalg[9], Gábor M. Kovács[33,34], Ekaphan Kraichak[35], Krisztina Krizsan[36], Cletus P. Kurtzman[37], Karl-Henrik Larsson[38], Steven Leavitt[35], Peter M. Letcher[39], Kare Liimatainen[40], Jian-Kui Liu[11], D. Jean Lodge[41], Janet Jennifer Luangsa-ard[42], H. Thorsten Lumbsch[35], Sajeewa S.N. Maharachchikumbura[11], Dimuthu Manamgoda[11], María P. Martín[18], Andrew M. Minnis[43], Jean-Marc Moncalvo[44], Giuseppina Mulè[45], Karen K. Nakasone[46], Tuula Niskanen[40], Ibai Olariaga[28], Tamás Papp[36], Tamás Petkovits[36], Raquel Pino-Bodas[47], Martha J. Powell[39], Huzefa A. Raja[48], Dirk Redecker[49], J. M. Sarmiento-Ramirez[18], Keith A. Seifert[27], Bhushan Shrestha[50], Soili Stenroos[47], Benjamin Stielow[2], Sung-Oui Suh[51], Kazuaki Tanaka[52], Leho Tedersoo[9], M. Teresa Telleria[18], Dhanushka Udayanga[11], Wendy A. Untereiner[53], Javier Diéguez Uribeondo[18], Krishna V. Subbarao[31], Csaba Vágvölgyi[36], Cobus Visagie[2], Kerstin Voigt[54], Donald M. Walker[55], Bevan S. Weir[32], Michael Weiß[56], Nalin N. Wijayawardene[11], Michael J. Wingfield[17], J. P. Xu[57], Zhu L. Yang[14], Ning Zhang[58], Wen-Ying Zhuang[26] and Scott Federhen[1]**

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA, [2]CBS-KNAW Fungal Biodiversity Centre, P.O. Box 85167, 3508 AD Utrecht, The Netherlands, [3]Department of Pharmaceutical Sciences – Microbiology, Università degli Studi di Perugia, Perugia, Italy, [4]Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Marie Bashir Institute for Infectious Diseases and Biosecurity, Sydney Medical School-Westmead Hospital, The University of Sydney, Westmead Millennium Institute, Westmead, Australia, [5]Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden, [6]Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37920, USA, [7]Illinois Natural History Survey, University of Illinois, 1816 South Oak Street, Champaign, IL 61820, USA, [8]Mycology Section, Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3DS, UK, [9]Natural History Museum, University of Tartu, 46 Vanemuise, 51014 Tartu, Estonia, [10]Purdue University, Department of Botany and Plant Pathology, 915 W. State Street, West Lafayette, IN 47907, USA, [11]Institute of Excellence in Fungal Research, and School of Science, Mae Fah Luang University, Chiang Rai 57100, Thailand, [12]Imperial College London, Royal Botanic Gardens, Kew TW9 3DS, England, UK, [13]Muséum National d'Histoire Naturelle, Dépt. Systématique et Evolution CP39, UMR7205, 12 Rue Buffon, F-75005 Paris, France, [14]Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, Yunnan, P. R. China, [15]Departamento de Biología Vegetal II, Facultad de Farmacia, Universidad Complutense de Madrid, Madrid 28040, Spain, [16]Senckenberg Museum of Natural History Görlitz, PF 300 154, 02806 Görlitz, Germany, [17]Department of Microbiology and Plant Pathology, Forestry Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria 0001, South Africa, [18]Real Jardín Botánico, RJB-CSIC, Plaza de Murillo 2, 28014 Madrid, Spain, [19]Department of Forest and Conservation Sciences, Faculty of Forestry, The University of British Columbia, 2424 Main Mall, Vancouver, BC, Canada, [20]Institute of Animal Physiology and Genetics, Czech Academy of Sciences, v.v.i., Videnska 1083, Prague, Czech Republic, [21]Department of Biology, University of Toronto, 3359 Mississagua Road, Mississagua, Ontario L5L 1C6, Canada, [22]Institute of Biological, Environmental and Rural Sciences, Prifysgol Aberystwyth, Aberystwyth, Ceredigion Wales SY23 3DD, UK, [23]Institute of Plant Sciences, Karl-Franzens-University, Holteigasse 6, 8010 Graz, Austria, [24]Department of Plant Sciences, University of the Free State, P.O. Box 339, Bloemfontein 9300, South Africa, [25]CSIRO-Plant Industry, CANBR, GPO Box 1600, Canberra ACT 2601, Australia, [26]State Key Laboratory of Mycology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China, [27]Biodiversity (Mycology and Microbiology), Agriculture and Agri-Food Canada, 960 Carling Avenue, Ottawa, Ontario, Canada, [28]Department of Botany, Swedish Museum of Natural History, P.O. Box 50007, SE-104 05, Stockholm, Sweden, [29]Agroscope Changins-Wädenswil Research Station ACW, Département de recherche en Protection des végétaux grandes cultures et vigne/Viticulture et oenologie, CP 1012, CH-1260 Nyon, Switzerland, [30]Korean Agricultural Culture Collection, National Academy of Agricultural Science, RDA, Suwon, 441-707, Korea, [31]University of California, Davis Department of Plant Pathology Davis, CA 95616, USA, [32]Landcare Research, Private Bag 92170, Auckland 1142, New Zealand, [33]Eötvös Loránd University, Institute of Biology, Department of Plant Anatomy, Pázmány Péter sétány 1/c, 1117 Budapest, Hungary, [34]Plant Protection Institute, Centre for Agricultural Research, Hungarian Academy of Sciences, Budapest, H-1525, Hungary, [35]Science and Education, The Field Museum, 1400 S. Lake Shore Drive, Chicago, IL 60605, USA, [36]University of Szeged, Faculty of Science and Informatics, Department of Microbiology, Kozep fasor 52, Szeged, H-6726, Hungary, [37]Bacterial Foodborne Pathogens and Mycology Research Unit, U.S. Department of Agriculture, National Center for Agricultural Utilization Research, Agricultural Research Service, 1815 North University Street, Peoria, IL 61604, USA, [38]Natural History Museum, P.O. Box 1172 Blindern, 0318 Oslo, Norway, [39]Department of Biological Sciences, The University of Alabama, Tuscaloosa, AL 35487, USA, [40]Plant Biology, Department of Biosciences, P.O. Box 65, 00014 University of Helsinki, Finland, [41]USDA Forest Service, NRS, PO Box 1377, Luquillo, Puerto Rico, [42]National Center for Genetic Engineering and Biotechnology (BIOTEC), 113 Paholyothin Road, Pathum Thani 12120 Thailand, [43]Department of Medical Microbiology and Immunology, University of Wisconsin-Madison, 1550 Linden Drive, Madison, WI 53706, USA, [44]Department of Natural History, Royal Ontario Museum,

and Department of Ecology Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada, [45]Institute of Sciences of Food Production, National Research Council (CNR), Via Amendola 122/O, Bari, Italy, [46]Center for Forest Mycology Research, Northern Research Station, U.S. Forest Service, One Gifford Pinchot Drive, Madison, WI 53726-2398, USA, [47]Botanical Museum, Finnish Museum of Natural History, FI-00014 University of Helsinki, Finland, [48]The University of North Carolina at Greensboro, Department of Chemistry and Biochemistry, 457 Sullivan Science Building, P.O. Box 26170, Greensboro, NC 27402-6170, USA, [49]Université de Bourgogne, UMR1347 Agroécologie, BP 86510, F-21000 Dijon, France, [50]Institute of Life Science and Biotechnology, Sungkyunkwan University,Suwon 440-746, Korea, [51]Mycology and Botany Program, American Type Culture Collection (ATCC), 10801 University Blvd., Manassas, VA 20110, USA, [52]Faculty of Agriculture and Life Science, Hirosaki University, 3 Bunkyo-cho, Hirosaki, Aomori 036-8561, Japan, [53]Department of Biology, Brandon University, Brandon, Manitoba, Canada, [54]Jena Microbial Resource Collection, Leibniz Institute for Natural Product Research and Infection Biology and University of Jena, Jena, Germany, [55]Department of Natural Sciences, The University of Findlay, Findlay, OH 45840, USA, [56]Institute of Evolution and Ecology, University of Tübingen, Auf der Morgenstelle 1, D-72076 Tübingen, Germany, [57]Department of Biology, McMaster University, Hamilton, Ontario L8S 4K1, Canada and [58]Department of Plant Biology and Pathology, Rutgers University, New Brunswick, NJ 08901, USA

*Corresponding author: Tel: 301-402-1502; Fax: 301-480-2918; Email: schoch2@ncbi.nlm.nih.gov

[†]These authors contributed equally to this work.

Citation details: Schoch,C.L., Robbertse,B., Robert,V. *et al.* Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. *Database* (2014) Vol. 2014: article ID bau061; doi:10.1093/database/bau061

Received 7 April 2014; Revised 19 May 2014; Accepted 28 May 2014

## Abstract

DNA phylogenetic comparisons have shown that morphology-based species recognition often underestimates fungal diversity. Therefore, the need for accurate DNA sequence data, tied to both correct taxonomic names and clearly annotated specimen data, has never been greater. Furthermore, the growing number of molecular ecology and microbiome projects using high-throughput sequencing require fast and effective methods for en masse species assignments. In this article, we focus on selecting and re-annotating a set of marker reference sequences that represent each currently accepted order of Fungi. The particular focus is on sequences from the internal transcribed spacer region in the nuclear ribosomal cistron, derived from type specimens and/or ex-type cultures. Re-annotated and verified sequences were deposited in a curated public database at the National Center for Biotechnology Information (NCBI), namely the RefSeq Targeted Loci (RTL) database, and will be visible during routine sequence similarity searches with NR_prefixed accession numbers. A set of standards and protocols is proposed to improve the data quality of new sequences, and we suggest how type and other reference sequences can be used to improve identification of Fungi.

**Database URL:** http://www.ncbi.nlm.nih.gov/bioproject/PRJNA177353

## Introduction

Fungi encompass a diverse group of organisms ranging from microscopic single-celled yeasts to macroscopic multicellular mushrooms. This implies that many of the challenges necessary to document fungal diversity overlap with those faced by researchers in other fields. Although yeast researchers share the challenges of other microbiologists to obtain viable cultures to study, macrofungal researchers often document species from dried specimens and face obstacles comparable with those of botanists. The majority of described fungal species still lacks any DNA sequence data, but it is also apparent that the vast majority of fungal diversity will have to be assessed solely by comparing DNA sequences, without accompanying cultures or physical specimens (1).

DNA sequence comparisons have demonstrated that many traditionally used phenotypic characters in Fungi are the result of convergent evolutionary processes and do not necessarily predict relatedness. Therefore, cryptic species continue to be discovered with phylogenetic methods even after examining well-studied species. Since the 19th century, it has also been accepted that fungi can occur in several morphological forms (morphs) arising from sexual, asexual or vegetative reproduction. Because these morphs often do not occur together in time and space, DNA characters greatly enhance the efficiency to confirm that separate morphs constitute a single species. This contributed to the declaration that different species names that have traditionally been applied to sexual and asexual morphs of the same fungal species are redundant (2). This redundancy is reflected in the most recent set of the rules guiding how fungal species are named, the International Code of Nomenclature for algae, fungi and plants (ICN) (3).

Improvements in how electronic data are disseminated also prompted changes in the ICN, namely, a requirement to register all new fungal taxonomic names at one or more online repositories. Recently, three candidates, Fungal Names, Index Fungorum or MycoBank were proposed (4). These databases provide an invaluable source of important information on vouchers that facilitate fungal identification. This, in turn, will aid the large-scale reassessment of taxonomic names required as part of the transition to use one name for each fungal species (5, 6). It will also improve the integration to a sequence-based classification (7).

For effective DNA-based identification to be implemented, the scientific community needs a continuously expanding, public and well-annotated set of DNA sequences. Each of these sequences needs to be associated with accurate specimen data and a current species name. Just as the current ICN addresses the requirements for a common nomenclature of species names, improved standards related to DNA sequences and specimens will improve the ability to communicate diversity effectively in ecological and microbiome studies. This infrastructure will provide the framework required to further our understanding of biology across all groups of Fungi.

### Current state of sequence databases

GenBank, together with its collaborative partners in the International Nucleotide Sequence Databases Collaboration (INSDC), i.e. the DNA Data Bank of Japan and the European Nucleotide Archive (ENA) has long been the most comprehensive resource of nucleotide data (8). It is tasked with archiving the world's genetic data as an open resource to all researchers. In spite of an extensive review of user submissions, GenBank essentially relies on users to accurately name their sequences. This results in a significant number of sequences deposited under erroneous or imprecise names, so-called 'dark taxa' (9, 10). This complicates efforts to clearly assign taxonomic names to unknowns. Mycologists have long been a vocal group in arguing for improving the accuracy of names used in GenBank (11, 12). In addition, biologists have expressed concern about the lack of associated voucher data in many GenBank entries (13). Although a specimen voucher qualifier has been available and promoted by GenBank since 1998 (14), this remains poorly used by submitters. To improve this, GenBank now recommends applying a version of the Darwin Core standards (15), which intends to facilitate the sharing of information about biological diversity through reference definitions (e.g. a standardized specimen voucher format) for relevant data. Where feasible, this will apply to any biorepository data shown in the 'specimen_voucher', 'culture_collection' and 'bio_material' qualifiers of a GenBank sequence accession (14). This format also allows for vouchers to be linked directly from a sequence accession to a dedicated specimen or culture page at the relevant biorepository (where available), and it improves traceability across different databases.

A number of additional specialized databases focused on specific marker sequences have been built to further enhance sequence accuracy. Mycologists have used DNA sequence data for testing species-rank hypotheses for over 20 years. The internal transcribed spacer (ITS) region containing two spacers (ITS1 and ITS2) flanking the nuclear ribosomal 5.8S gene has been an especially popular marker (16). A curated ITS database focused on human and animal pathogenic fungi was established at www.mycology-lab.org for the International Society of Human and Animal Mycology (ISHAM). Initially, the UNITE database (http://unite.ut.ee/) had a similar functional focus on ectomycorrhizal ITS sequences (17). Since then, it has expanded to

provide tools for assessing sequence quality and Web-based third-party sequence annotation (PlutoF) to published sequences for all Fungi. The UNITE database now acts as a GenBank mirror for all fungal ITS sequences and has a particular focus on integrating sequences from environmental samples into reproducible taxonomic frameworks (18). Among other databases with similar aims, ITSoneDB focuses on ITS1 sequences (19), whereas the ITS2 database houses ITS2 sequences and their 2D structures (20). A number of additional publicly available online databases favor other sequence markers for fungal identification, e.g. the large and small nuclear ribosomal subunits (18S, 28S) and fragments from the translation elongation factor 1-alpha gene (21, 22). Several of these databases are focused on specific taxonomic groups (23–25). The DNA barcoding movement made an important impact on sequence accuracy by promoting a clear set of standards for DNA barcodes: raw sequence reads and reliable sequence data combined with a correct taxonomic name as well as collection and voucher information (26, 27). The Barcode of Life Data System (BOLD; 28) has a significant amount of sequence data that overlap with GenBank and was explicitly set up for DNA barcoding. The CBS-KNAW Fungal Biodiversity Centre, MycoBank and the recently launched BOLD mirror, EUBOLD, are also proposing online identification tools that can compare unknown sequences simultaneously against several reference databases.

Despite its long history of usage, mycologists have only recently proposed the ribosomal ITS region as a universal DNA barcode marker for Fungi (29). This means that regardless of several limitations (30), ITS will likely remain the main marker of choice for fungal identification in the immediate future. Since 2012, the ITS region has specifically been used for species identification in numerous DNA barcoding studies on a variety of fungal groups ranging from mucoralean fungi (31) to common molds such as *Aspergillus* and *Penicillium* (32). Broader-scale studies have evaluated the utility of generating fungal barcodes for a wide variety of fungal specimens (33–35). Extracted DNA can reliably be amplified by means of the polymerase chain reaction (PCR) and sequenced for most dried fungal specimens up to 30 years old. In several cases, much older specimens have been successfully sequenced (36–39) opening the possibilities of generating fungal barcodes from some legacy type specimens. The current age record for a fungal sequence from a type specimen stands at 220 years for a mushroom species, *Hygrophorus cossus,* collected in 1794 (40).

The use of multigene analysis has now become common in defining phylogenetic species boundaries within mycology. The standard for species delimitation in mycology remains the genealogical concordance species recognition concept first advocated by Taylor *et al*. (41). This relies on character comparisons from at least three unlinked loci. In comparison, the DNA barcode approach relies on less rigorous analysis techniques using universally sampled sequences from only one or rarely two markers. The focus in DNA barcoding is on obtaining limited sequence data from the largest possible number of specimens. It is most efficient in specimen identification, used in concert with a well-validated database containing accurate species delimitations (42). However, where information on species boundaries are lacking, it can also be used for initial species discovery. In line with this, it is our intention to maximize the accuracy of the sequences available for specimen identification and to emphasize where more sampling is required during species discovery.

## Selecting reference sequences

DNA barcodes per definition, have to be backed up by publicly accessible raw data files (trace data), and, if they are linked to type material, have the potential to act as reference sequences that also provide a higher confidence in sequence accuracy. However, many other important sequences are already available in the public databases that would not be qualified as official barcodes. The need to communicate specific levels of confidence in sequence accuracy has yielded proposals for a quality scale in sequences (43), but establishment of such a system remains elusive. Since 1 January 1958, any validly published species name is connected to a type, which should be treated as primary reference. A type can principally be an original depiction of a species, though then it is good practice to designate a separate specimen as a neotype, or epitype where appropriate. However, in the majority of cases, a type will be a physical specimen. Type specimens are the only specimens to which one can reliably apply the original name, thereby providing a physical link to all other associated information. Having an ITS sequence or other marker sequences connected to a reliable publicly accessible representative of the species thus provides researchers with a reference point to a specific name. This makes it possible to unambiguously communicate findings with the research community and provide the opportunity to generate additional related data and expand current collections.

At GenBank, a particular challenge has always been the annotation of sequences derived from type material. Until recently, there was no standard means to specify type-related information during the process of data submission. Notes can be added to individual accessions, but they remain cumbersome to uncover in queries. In this article, we describe efforts to address this shortfall by expanding fields

to indicate type material in the National Center for Biotechnology Information (NCBI) Taxonomy database. A number of ITS sequences were re-annotated and formatted in a separate curated database at NCBI, RefSeq Targeted Loci (RTL). This database was originally set up as a repository for bacterial type sequences obtained from the databases RDP, SILVA and Greengenes. It was subsequently expanded to Fungi—initially using a divergent set of sequences generated by the collaborative Assembling the Fungal Tree of Life project (AFToL) (44).

We release an initial reference sequence set of nearly 2600 ITS accessions covering ~2500 species for inclusion in RTL. These records have been extensively verified with input from collaborators at Index Fungorum, MycoBank and UNITE, as well as a large group of taxonomic specialists. The existing set was chosen to represent most currently accepted orders (45) with eventual expansion to lower hierarchical taxa. It is intended that this new reference sequence set will continue to be widely used, adapted and expanded by the research community.

## Materials and Methods

**Verification steps for RefSeq data set:**

Verification was done in the following order with each step building on the information of the previous step (Figure 1).

1) *Collecting ITS records for evaluation*. Lists of potential ITS accessions from type and verified specimens for display in RefSeq were generated in several ways via Entrez queries in the NCBI Nucleotide database, daily taxonomy curation and collaboration with experts in the fungal research community.

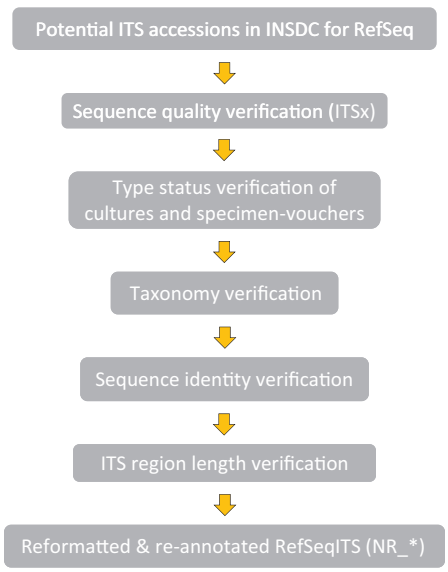


**Figure 1.** Workflow of the ITS verification for RTL ITS.

2) *Sequence quality*. All accessions were verified with the Perl script ITSx (46) to ensure sequence continuity including the ITS1, 5.8S and ITS2 regions. A record was excluded if it had an incomplete ITS2 region (as inferred by ITSx) and incomplete ITS1 region, which had no conserved CATTA-like motif at the 5′ end (within 40 bases of the end), and the length was <80% of the average complete ITS region (annotation as inferred by ITSx) for the taxonomic order to which it belonged. If there was no order defined then the class statistic was used, and if no class was defined then the statistic of the complete ITS region at kingdom rank (Fungi) was considered. In addition, sequences were also verified for non-ATGC characters [i.e. IUPAC DNA ambiguity symbols (47)], which often indicates poor quality. Their presence was limited to <0.5% of the ITS region. In some cases, exceptions to this rule and the length requirement were made for sequences representing underrepresented lineages.

3) *Type material definitions*. *Type*: The ICN defines a type as 'that element to which the name of a taxon is permanently attached' (Article 7.1). In addition, it states that types are not necessarily defined as the best representatives of the taxon (Article 7.2). We thus attempted to distinguish between the various types and annotate type status in the organism note field in each sequence record. We only considered one of the following types per species: holotype, isotype, lectotype, neotype, epitype, syntype and paratype. The holotype is a single specimen designated by the original author at the time of a species description. The other types indicate a variety of relationships to that specimen or can serve as replacements in certain circumstances (see glossary of the ICN for details: http://www.iapt-taxon.org/nomen/main.php?page=glo). Where we could not clearly distinguish the kind of type, these are annotated only as type. For the verification of type status, we relied mainly on the information at culture collection databases listed in Table 1 and the nomenclatural databases, MycoBank and Index Fungorum, as well as experts in the fungal research community. The main source of type status information was publications. Type status information can currently not be extracted from publications in a high-throughput manner, and the documents themselves are not always freely accessible, making curation efforts time consuming and heavily dependent on manual curation. Where possible, types tied to the original species description (protolog) of the currently accepted name were selected.

Ex-type: Living cultures do not have the formal nomenclatural status of a type specimen, but sequences obtained from cultures that were derived from type specimens were also indicated; where possible, it was indicated from what kind of type collection these originated. Details on such ex-type cultures and type specimens are both included under

**Table 1.** List of collection databases with specimen pages to which links were established from records in GenBank

| Acronym | Collection Institute | Database link |
|---|---|---|
| ACBR | Austrian Center of Biological Resources and Applied Mycology | http://www.acbr-database.at/BioloMICS.aspx |
| ATCC | American Type Culture Collection | http://www.atcc.org/Products/Cells_and_Microorganisms/Fungi_and_Yeast.aspx |
| BCRC | Bioresource Collection and Research Center | https://catalog.bcrc.firdi.org.tw/BSAS_cart/controller?event=WELCOME |
| BPI | US National Fungus Collections, Systematic Botany and Mycology Laboratory | http://nt.ars-grin.gov/fungaldatabases/specimens/Specimens.cfm |
| CBS | Centraalbureau voor Schimmelcultures, Fungal and Yeast Collection | http://www.cbs.knaw.nl/Collections/Biolomics.aspx?Table=CBS%20strain%20database |
| CFMR | Center for Forest Mycology Research | http://www.fpl.fs.fed.us/search/mycology_request.php |
| DSM | Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH | http://www.dsmz.de/catalogues/catalogue-microorganisms.html |
| FRR | Food Science Australia, Ryde | http://www.foodscience.csiro.au/fcc/search.htm |
| ICMP | International Collection of Microorganisms from Plants | http://scd.landcareresearch.co.nz/Search/Search/ICMP |
| JCM | Japan Collection of Microorganisms | http://www.jcm.riken.jp/JCM/catalogue.shtml |
| MA | Real Jardín Botánico de Madrid Herbarium | http://www.rjb.csic.es/jardinbotanico/jardin/index.php?Cab=109&len=es |
| MAFF | MAFF Genebank, Ministry of Agriculture, Forestry and Fisheries | http://www.gene.affrc.go.jp/databases-micro_search_en.php |
| MICH | University of Michigan | http://quod.lib.umich.edu/h/herb4ic?page=search |
| MTCC | Microbial Type Culture Collection and Gene Bank | http://mtcc.imtech.res.in/catalogue.php |
| MUCL | Mycotheque de l'Universite Catholique de Louvain | http://bccm.belspo.be/db/mucl_search_form.php |
| NBRC | NITE Biological Resource Center | http://www.nbrc.nite.go.jp/NBRC2/NBRCDispSearchServlet?lang=en |
| NRRL | Agricultural Research Service Culture Collection | http://nrrl.ncaur.usda.gov/cgi-bin/usda/index.html |
| PDD | New Zealand Fungal and Plant Disease Herbarium | http://nzfungi2.landcareresearch.co.nz |
| PYCC | Portuguese Yeast Culture Collection | http://pycc.bio-aware.com/BioloMICS.aspx?Table=PYCC%20strains |
| SAG | Sammlung von Algenkulturen at Universitat Gottingen | http://sagdb.uni-goettingen.de/ |
| UAMH | University of Alberta Microfungus Collection and Herbarium | https://secure.devonian.ualberta.ca/uamh/searchcatalogue.php |

Unique acronyms were taken from the GenBank collections database and, where possible, agree with labels used by Index Herbariorum, WFCC and GRBio

'type material' in the NCBI Taxonomy database. Type identifiers in the NCBI Taxonomy database can include both heterotypic synonyms (also referred to as taxonomic or facultative synonyms) and homotypic synonyms (also referred to as nomenclatural or obligate synonyms). A simplified description of homotypic and heterotypic synonyms are indicated in Supplementary Material (Supplementary Figure S1).

*Verified*: This label was used to label placeholder sequences for important lineages in the fungal tree of life until sequences derived from type material are available. We relied on the advice from acknowledged taxonomic experts and input from large collaborative projects such as the AFToL project.

4) *Current taxonomic name*. ITS records from type specimens were selected only for current names where a single type applies, i.e. homotypic names. This means all associated obligate synonyms can effectively be traced to a single type specimen. Records associated with types from names that were synonymized subjectively were excluded where possible (heterotypic names). However when possible, we combined and annotated heterotypic types from asexual and sexual morphs (anamorphs and teleomorphs) from the same species in order to promote nomenclatural stability. An example is indicated in Supplementary Material (Supplementary Figure S1). The taxonomic names in current use were identified by consulting the latest publications, acknowledged taxonomic experts, culture collection databases as well as MycoBank and Index Fungorum. Where possible, a script using cURL (http://curl.haxx.se/) was used to extract type status and names from databases such as CBS and MycoBank.

5) *Sequence identity*. This is not the first attempt at verifying data in INSDC, and thus we relied on the data from the UNITE (version 5) and ISHAM databases to help verify sequences. Also, the sequence identity of selected INSDC records that could potentially be represented in RefSeq was compared with other sequences from type specimens. These were identified via type specimen identifiers obtained from MycoBank (compared with the isolate, strain, collection and specimen voucher fields) and from daily taxonomy curation. Finally, any type material data were uploaded as permanent name types in the NCBI Taxonomy database.

Sequence identity was considered accurate, and a sequence was considered to be associated with the type specimen if one of these conditions were met:

- There was >99.5% identity over >90% of the ITS region in the potential RefSeq sequence compared with another type specimen sequence of the same TaxID in GenBank using megablast alignments. Instead of using 100% identity, we used 99.5% to accommodate for a small number of non-ATGC characters. Each sequence record is associated with one TaxID, and the TaxID represents one taxon that in NCBI taxonomy can

accommodate several synonymous names (e.g. http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=48490&lvl=3&lin=f&keep=1&srchmode=1&unlock).
- The same accession was in UNITE's list of representative sequences (RepSs) or reference sequences (RefSs) with the same TaxID. Any synonymous taxon names used in UNITE were resolved, and the TaxID were identified with the name status tool in NCBI taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi).
- There was >99.5% identity over >90% of the ITS region in the potential RefSeq sequence compared with RepS or RefS with the same TaxID from UNITE using megablast alignments.

Possible misidentifications/labeling of accessions (not verified above) were investigated for the following:

- Sequences that were >99.5% identical over >90% of the ITS region to more than one type sequence identified in GenBank or RefS/RepS from UNITE of a different TaxID using megablast alignments.
- Sequences that were <98.5% identical over >90% of the ITS region to RefS/RepS from UNITE of the same TaxID using megablast alignments.
- Same accessions associated with different TaxIDs in GenBank and UNITE.

Further investigation was necessary if more than one ITS accession were available for a type, and one or more copies were <99.6% identical to the sequence selected for representation in RefSeq. This was done to ensure that the selected sequence was not the outlier in the group, which may be the result of low sequence quality or mislabeling. Sequence copies were aligned using MAFFT (http://www.ebi.ac.uk/Tools/msa/mafft/) and viewed in BioEdit (48) to determine which sequence contained the bases that are at odds with the rest. If only two sequence copies from the specimen were available to compare, then additional sequences from the same TaxID were aligned. If the uncertainty could not be resolved, then no sequence was selected for RefSeq.

### Reformating of accessions for RefSeq

Each ITS record was re-annotated with ribosomal RNA (rRNA) and miscellaneous RNA (misc_RNA) features representing the boundaries of the rRNA and the ITSs as predicted by the ITSx Perl script. Lists of ITS records with metadata provided by experts were compared with metadata in the original GenBank submission. Source features were reformatted, corrected and augmented with information where needed. Culture collection specimens entered in the strain or isolate fields were moved to the culture collection field. Similarly, any herbarium specimen information was moved to the specimen voucher field. If the original GenBank submission contained no identifier from a collection in the NCBI Collections database, then the appropriate public collection identifier obtained from the original species description was added to the RefSeq record. Collection codes used in the culture collection and specimen voucher fields followed the acronym format used by GenBank indexing (the NCBI Collections database available at http://www.ncbi.nlm.nih.gov/projects/BioCollection/search_collection.cgi). Both these fields were formatted appropriately with the code separated with a colon from the collection's correct identifier. If a dedicated specimen page was available online, with the collection identifier in the URL, direct links to the culture collection database could be made available when the correct format was used. Google searches for the presence of online databases of all collections associated with this data set were performed and specimen specific pages identified. The note field for each record was augmented with type status information, which included the type category (holotype, isotype, etc.) and the species name associated with the first description of the specimen.

### Centrality analysis and clustering

A centrality analysis was performed with BioloMICS (from BioAware, Hannut, Belgium) to find the most central sequence to a given group, which is the sequence having the highest average similarity to other members of the group (49). Because sequences selected for RefSeq were limited to only one record per species (a few species with known internal variation had multiple records from one specimen), the group was not defined at species rank but at genus rank. The centrality analysis shows the diversity in a designated group.

A multidimensional cluster analysis was performed to visualize the distribution of the data. The distance between every pair of sequences was calculated based on similarity, and a distance matrix was created. Using the multidimensional scaling tool (BioloMICS) with the distance matrix, the data points were visualized in 3D and colored according to the classification rank specified.

## Results

### Sequence quality

Of a set of ~3100 accessions considered for inclusion, we removed 16% for a variety of reasons, and currently, 2593 accessions were selected for RefSeq. The most commonly

encountered problem was lack of sufficient and reliable metadata associated with a sequence record to confirm a type specimen in a timely manner. During the process of verifying sequence quality, some records were identified by ITSx as being problematic and were excluded, e.g. containing assembly chimeras, incomplete sequences (e.g. a missing 5.8S gene), sequences not from the ITS region or not of fungal origin. Figure 2 shows the length variation of accessions destined for RefSeq and with a complete ITS region (which accounted for ~70% of the RefSeq ITS records). When the nuclear ribosomal 18S end or 28S start was within the first or last ~25 bases of the sequence in the record, it became difficult for the ITSx script to identify it with confidence. This was due to the fact that the probability of a hidden Markov model score influenced by chance alone became much higher. Sequences with a CATTA motif within 40 bases of the 5′ end of a sequence but with an ITSx annotation of 'ITS1 partial' were considered complete. Some sequences contained more than one CATTA motif. These were compared with closely related sequences with a complete ITS region as defined by ITSx to confirm that these were complete for the ITS1 spacer. The boundary of the 18S was mostly (in 84% of the sequences with an 18S fragment) defined by the CATTA motif, although not all sequences contained this motif. Rather, variations of the CATTA motif were observed in some sequences, such as CATTC (e.g. in Mortierellaceae), CATTG (e.g. in Diaporthales), CACTA (e.g. in Cystofilobasidiales) or CAGTA (e.g. in Tremellales). The submitted sequence toward the end of the ITS2 spacer was frequently long enough to identify the 28S start with confidence, and no additional effort was made to identify conserved motifs within the last few bases. The majority (95% of 2593 sequences) of the RefSeq-selected ITS sequences had no undetermined bases, and the rest mostly had only one non-ATGC character, but none had more than four non-ATGC characters.

## Type status

Metadata associated with accessions in lists provided by mycology experts were compared with source metadata of these accessions in GenBank at NCBI. Conflicting information (e.g. collection/specimen identifiers) was resolved by updating the GenBank record (if submitted to GenBank and the original submitter was involved) or the RefSeq record. When the correct information was not rapidly

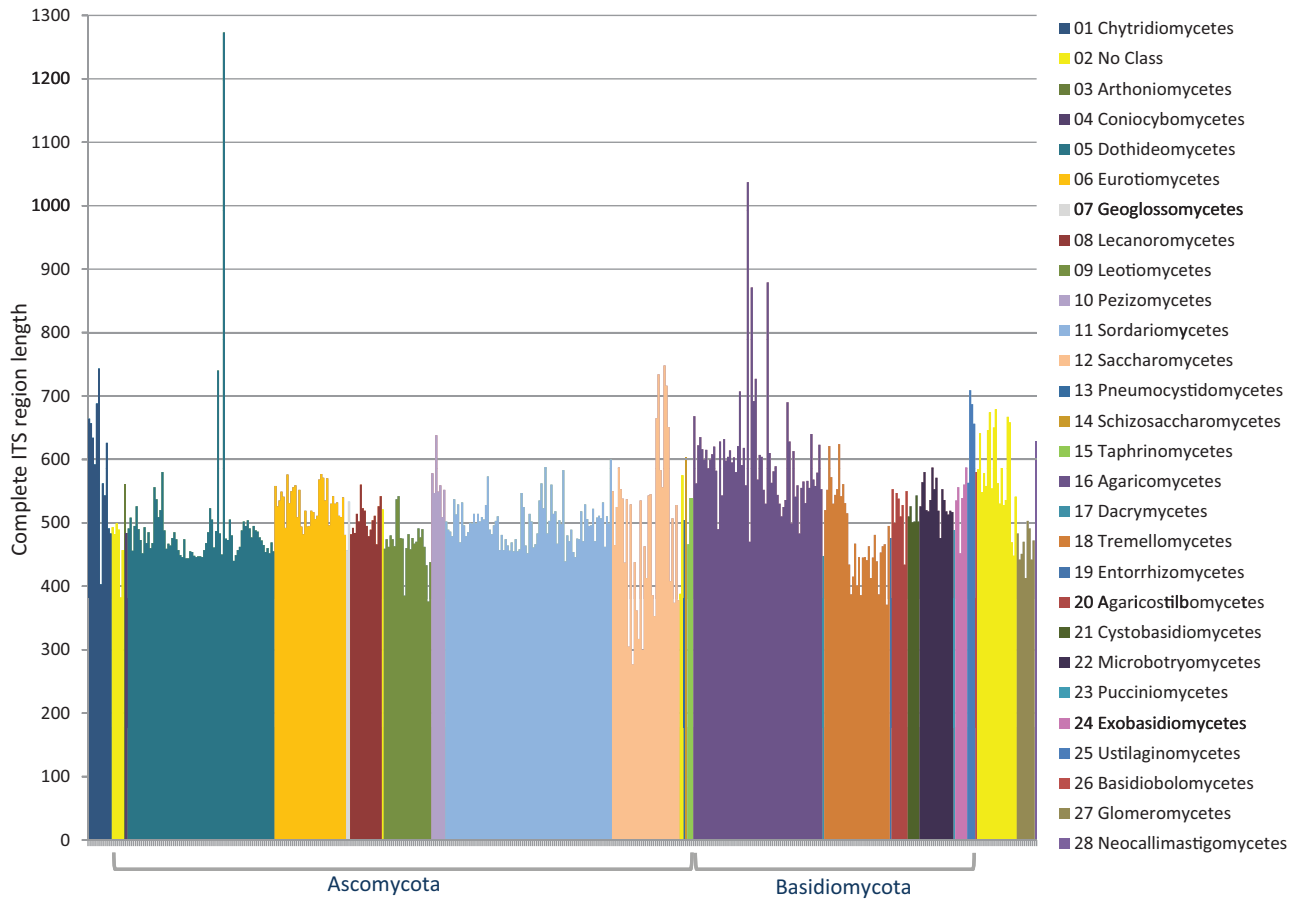


**Figure 2.** ITS length variation of complete ITS regions in the RTL data set according to class.

discernable, records were excluded from further analysis. Type status information of culture collection identifiers or specimen vouchers provided by the community was also compared with taxonomy/collection databases and publications. This curation process involved the research community and curators at culture/herbarium collections to resolve conflicting or missing information. Similar to the process in GenBank, culture/herbarium information from recently published research is not released publicly until collection curators receive notification or find the associated publication in the public domain. In addition, even though more and more culture/herbarium information is digitized, backlogs often exist. Thus, the absence of culture/herbarium numbers at an accessible online database does not necessarily imply a dead, contaminated or misidentified specimen. Only a small portion of the conflicting type specimen identifier information or absence was attributed to typographical errors or dead/contaminated cultures (14 identifiers). Where possible, we excluded types of heterotypic synonyms (see Supplementary Figure S1), as explained in following curation steps.

## Current taxonomic name

Identifying homotypic synonyms involved collecting the name by which a type specimen was first described, the name with which this specimen is currently associated and finally the original name (basionym) of that current name. Names provided by experts in the research community were compared with those in GenBank records, as well as other collection and taxonomy databases. The original name of the type specimen had to be the same as the current name or basionym and, if not, it needed to be a homotypic synonym of the current name to be considered for inclusion in this RefSeq set. The majority of this current name information (92%) was easily accessible with a script from MycoBank, and ~56% of the original names of type specimens were accessible from collection databases. However, any remaining information required a manual labor-intensive effort to obtain or verify names from publications or less accessible databases. This step in the curation process revealed many discrepancies between databases (and publications), which included orthographic variants, a need for taxonomic updates and spelling or labeling mistakes. Most discrepancies were resolved by addressing these issues at NCBI Taxonomy, external taxonomy and collection databases involving the respective curators. At publication time, 94% of RefSeq ITS records used the same current name that MycoBank or Index Fungorum used, and the rest used published names that were not public at both databases (1%) and different/not designated as current name at Index Fungorum (5%).

## Sequence identity

Sequence identity of UNITE's curated list of RepSs or RefSs were compared with those in GenBank selected for RefSeq curation. The UNITE database uses a centrality test to verify sequence identity and evaluate curation. However, because of filtering steps at UNITE and newly described species with a unique ITS sequence, not all RefSeq accessions are present in UNITE or associated with a species hypothesis (SH) (Figure 3). Comparisons against selected type sequences in GenBank verified the sequences selected for RefSeq at type specimen level making sure the best sequence for the specimen was selected. Comparisons identified classification discrepancies between MycoBank, Index Fungorum (used by UNITE) and NCBI Taxonomy, which have been communicated among the different curation databases. This comparison step also identified discrepancies in voucher or species names between the GenBank record and the publication, which could then be corrected. Sequences that were >99.5% identical and had over 90% overlap of the ITS region with more than one type sequence in GenBank or RefS/RepS from UNITE under a different TaxID were investigated. Discrepancies mostly revealed the existence of closely related species, which have been noted in a publication or by experts in the fungal research community. Thus, for these cases, there was no problem with the identity of the specimen under the classification point. Sequences that were <98.5% identical and had over 90% overlap of the ITS region to RefSs/RepSs from UNITE of the same TaxID

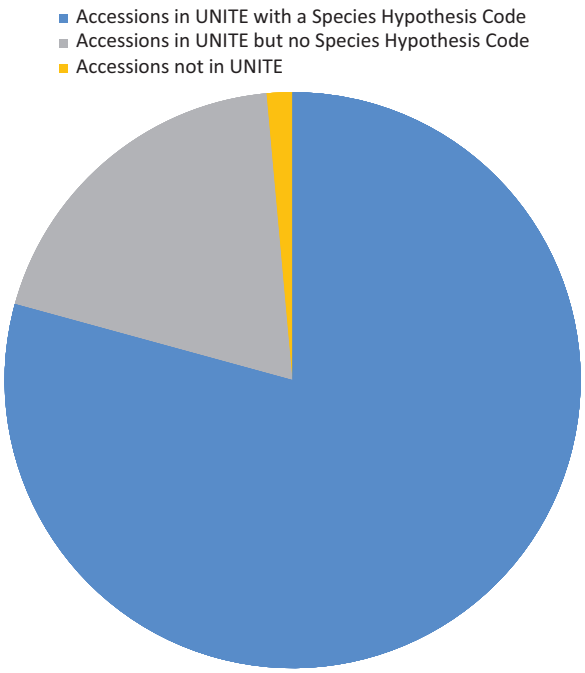


**Figure 3.** Diagram showing the proportion of accessions associated with UNITE (version 6) data.

were also investigated. Discrepancies mostly revealed the unclear indication of types from heterotypic synonyms, effect of non-ATGC characters, incorrect type specimens or sequences incorrectly associated with the culture or specimen. In two records, the difference between sequences from the same type material (same collection) was as great as 13 bases in a taxon not known for variation within ITS copies, and these records were excluded. Intragenomic variation is a known phenomenon in the ITS region (50). Such variation may typically be encountered when sequences were derived from cloned PCR products. Where needed, multiple ITS records for a single species were added. For example, Fungi from Glomeromycota are often represented with more than one ITS record. A few similar cases with multiple ITS sequences were also indicated in Basidiomycota (*Megacollybia subfurfuracea*, *Mucidula mucida* and *Ponticulomyces kedrovayae*).

### Reformatting accessions for RefSeq

All ITS accession numbers in RefSeq start with NR_, and are associated with an RTL Bioproject number. This allows RefSeq users to easily find all curated records (http://www.ncbi.nlm.nih.gov/nuccore/?term=PRJNA177353) and view a summary of the project (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA177353). The definition line (which appears in the output of the sequence similarity search tool BLAST in GenBank) has been simplified to the following format: '[species name] [culture collection/specimen voucher identifier] ITS region; from TYPE/verified material' (for example: *Penicillium expansum* ATCC 7861 ITS region; from TYPE material). All records were re-annotated, and the 34% that did not have annotation in INSDC now have annotation in the RefSeq version. Only ~25% of the selected INSDC records had culture collection or/and specimen voucher information that was correctly fielded and formatted. Culture collection information was moved to the culture collection field and formatted correctly for just over 1000 records. By doing so, these records could potentially be linked to more metadata at a collection's database. The 'rRNA' feature key was used to indicate the boundaries of the 18S, 5.8S and 28S rRNA and the 'misc_RNA' feature key to indicate the position of the two ITSs (Figure 4).

About 250 records were edited to correct collection/specimen voucher information or to add a collection number from a public collection. After curation, all records contained a culture collection or/and specimen voucher identifier in the correct field. Sequences originated from material kept at 159 collections of which only 32% had a searchable public database. However, most (~75%) of the records were associated with material from collections with a public database (Figure 5). A small number of collections had a specimen page URL that includes the collection number and to which a link can easily be formatted. Before curation started, links existed to five biorepositories. Additional links were added, and the full list of biorepositories with their acronyms indicated is in Table 1. More links will be added as this becomes possible. Recently, LinkOut features linking to SH pages (maintained by UNITE) also became available for individual NCBI ITS sequence records.

### Results of centrality analysis

To visualize the taxonomic diversity in our currently selected data set, we present a profile of the RefSeq data set at class in using multidimensional scaling clustering (Figure 6). The centrality analysis at genus rank of the curated RefSeq ITS data set has shown that the ITS variation around a central sequence differs greatly among genera as visualized for those with ≥20 sequences (Figure 7). The centrality score range from 0 to 1, where a score of 1 reflected a sequence identical to the calculated central sequence. Based on this score, most species in some genera (e.g. *Penicillium*, *Colletotrichum*) form a tight group in relation to their central sequences (Figure 7). It was clear that in some genera, species cannot be distinguished by comparing ITS sequences only. Centrality scores of 1 indicated where the ITS region did not show variation to distinguish it from the central sequence, and these included a number of taxa, mainly from *Cladosporium*. The inability of the ITS region to distinguish between many, but not all, species has been reported before in several species, including *Cercospora* (51) and *Cladosporium* (52, 53). However, the distribution of centrality scores (Figure 7) shows that some genera are either diverse in terms of ITS sequence similarity or are in need of taxonomic revision (e.g. *Candida*, *Cryptococcus*). It is already well known that the large genera of asexual species *Candida* and *Cryptococcus* are polyphyletic (54). Other large genera, like *Mortierella* (55) and *Mucor* (56), also require revision. Thus, given the poorly defined boundaries of some genera and lack of ITS variability in several species, classifying an unknown ITS sequence to species, and sometimes genus rank will not always produce a definitive answer.

## Discussion

### Changes to NCBI databases

The NCBI Taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy) acts as the standard nomenclature and classification repository for the INSDC. It is a central core where taxonomic information for the entries in other

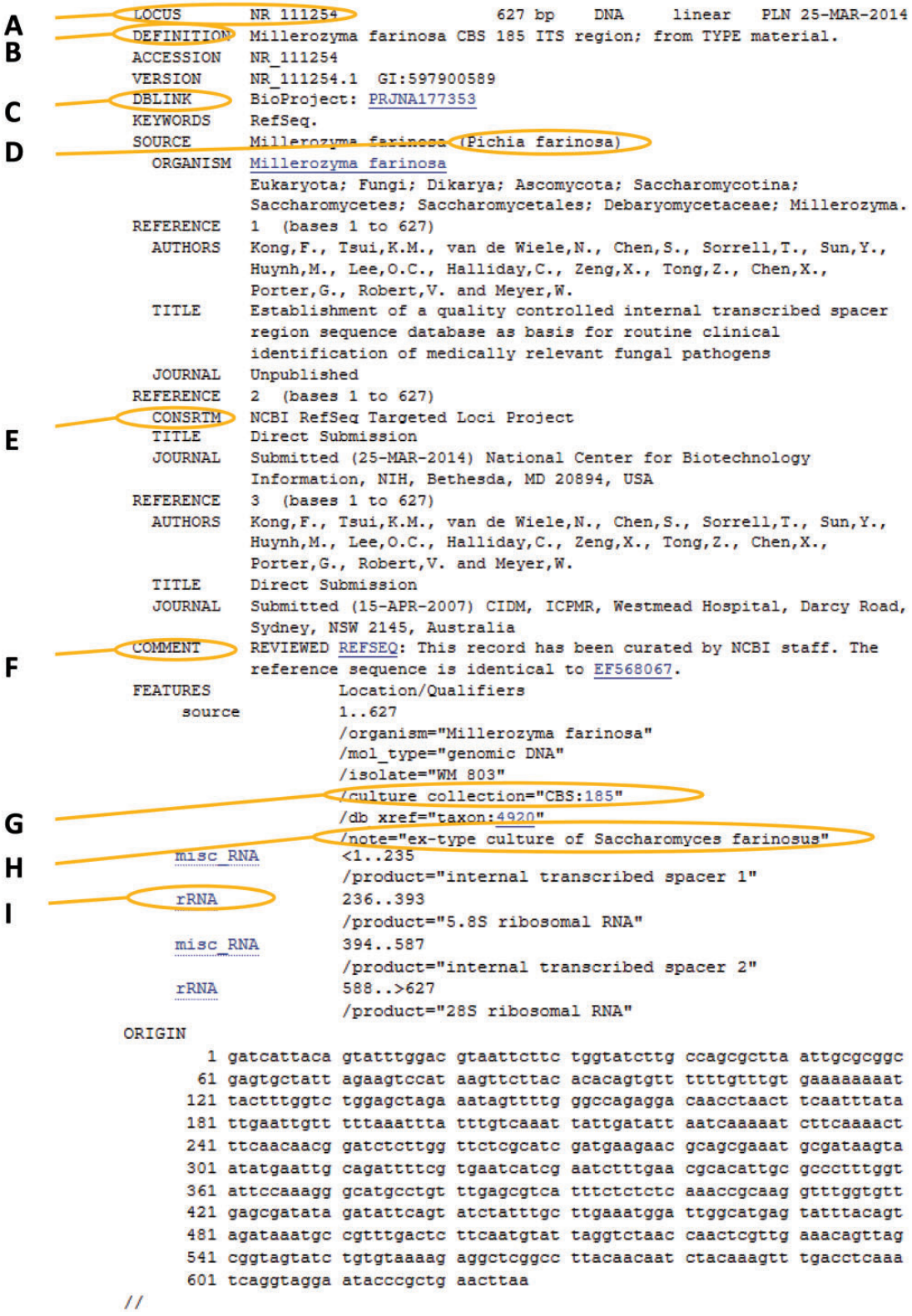


**Figure 4.** Anatomy of an RTL record. The marked areas indicate most common additions to the original nucleotide record. (**A**) New RTL accession number; (**B**) new simplified definition line; (**C**) Bioproject number for the ITS-targeted loci project; (**D**) GenBank synonym of current taxonomic name (used in cases of common usage); (**E**) label indicating that this is a RefSeq record; (**F**) comment regarding the source of the record; (**G**) the culture collection or specimen voucher presented as a validated structured triplet or doublet that can link directly to a relevant outside culture or specimen page; (**H**) additional information on the type and basionym name; (**I**) the ITS entry of all records was re-annotated to indicate the spacers and ribosomal genes.

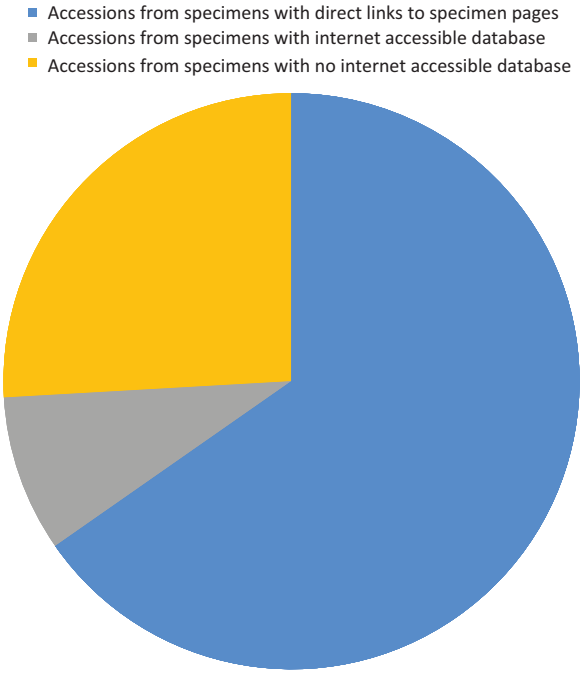


**Figure 5.** Diagram showing the proportion of accessions that originated from specimens associated with a collection that has an online database.

databases—such as GenBank—is stored. NCBI Taxonomy uses an array of name classes, e.g. 'scientific name', 'synonym', 'equivalent name' to express various taxonomic attributes (57). An example of a taxonomic record is shown in the Supplementary Material (Supplementary Figure S2). Two specific name types unique to fungi, 'anamorph' and 'teleomorph', but falling out of favor (58), will slowly be phased out as fungal classification adapts to a new nomenclatural system. An additional name type was recently added to the taxonomy database, 'type material'. This information is indexed so that related sequence records annotated with type specimen or ex-type culture identifiers with synonymous (homotypic and heterotypic) species names can be found with an Entrez query.

The following Entrez query 'sequence from type[filter] AND fungi[orgn]' will list all fungal taxonomic entries (from all genes and genetic markers) with type material attributes. The same query can be used to do a limited BLAST search on sequences from type material. Currently (March 2014), this covers over 150 000-nt sequence accessions in INSDC databases, including several additional regions besides ITS. This includes genome sequences and RefSeq messenger RNA records. In the era of phylogenomics, researchers may also be interested to know which

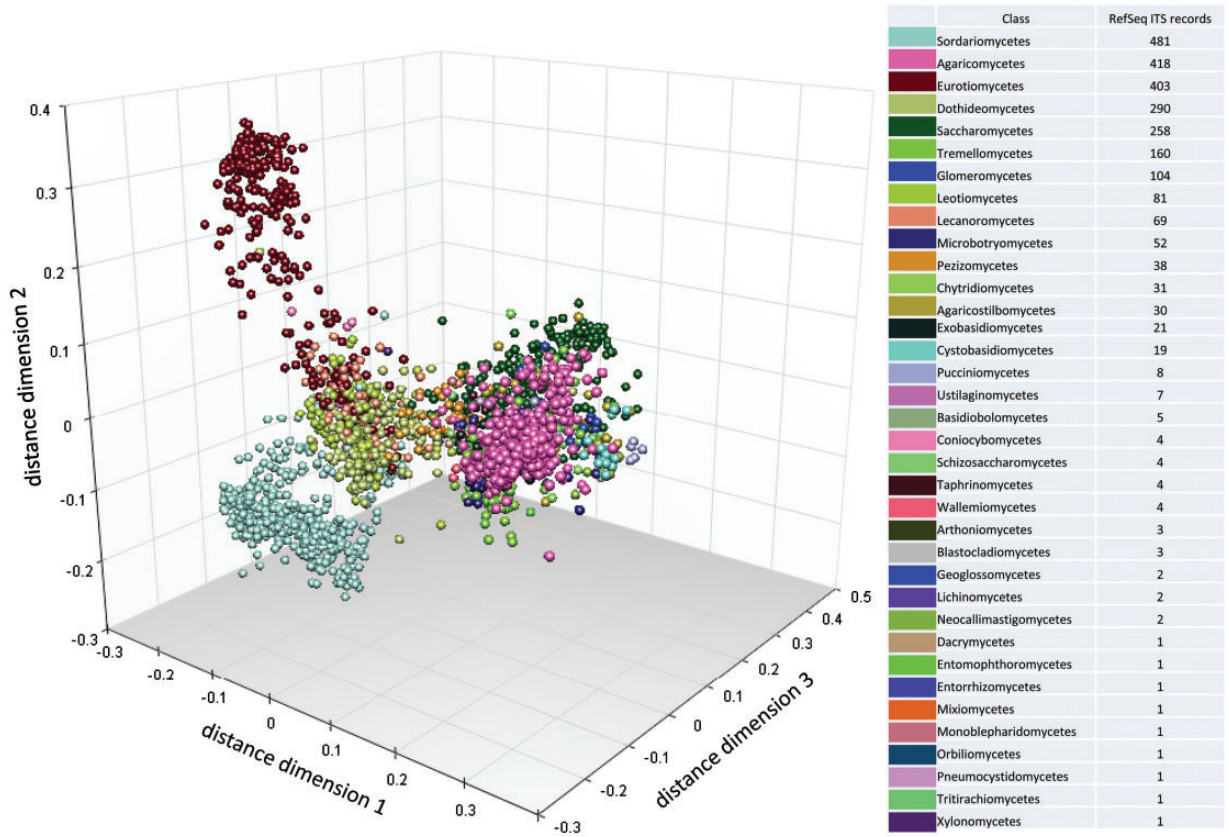


**Figure 6.** Multidimensional scaling clustering of RTL ITS sequences and coloring, according to the NCBI Taxonomy classification at class rank. Each marker represents an individual sequence.

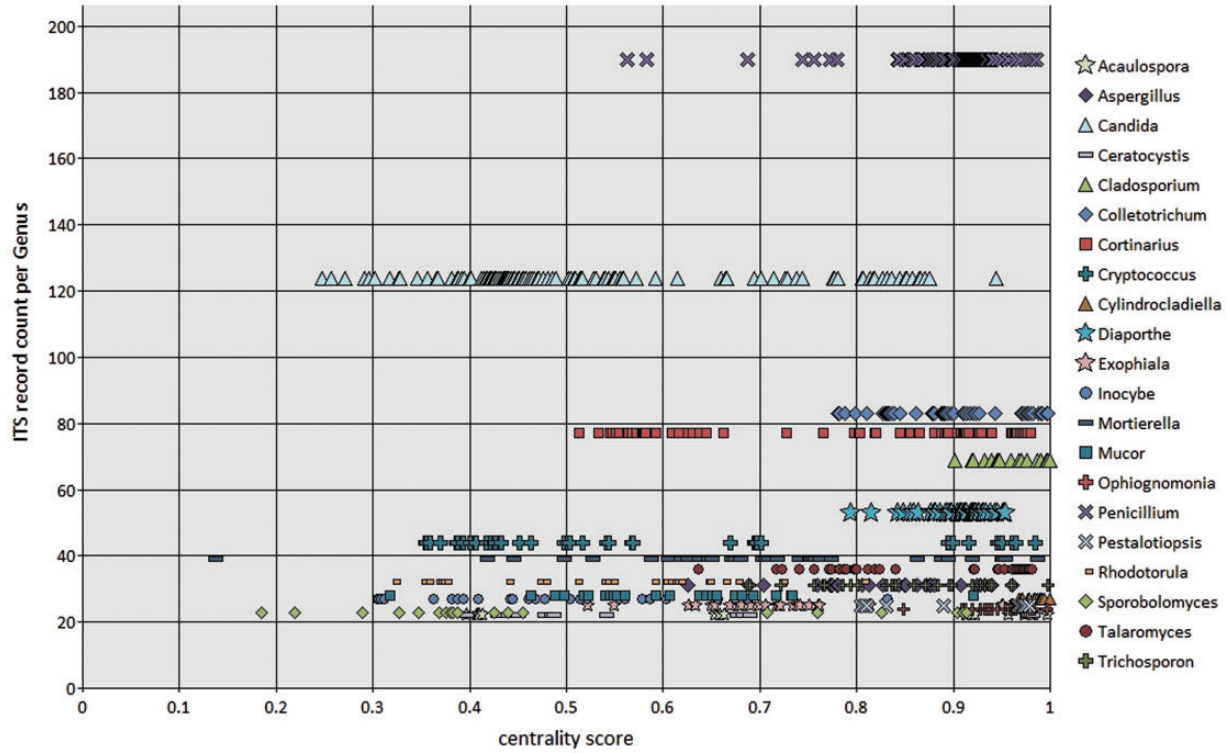


**Figure 7.** BioloMICS centrality scores of ITS sequences at genus rank, showing genera with ≥20 ITS records in the RefSeq data set. Each marker represents an individual sequence.

genomes are from type specimens and the associated proteins. In the future, the Entrez Protein Clusters database will also include fungal protein accessions, and those originating from type specimens can be marked as such. In addition to fungal type data, there are now >500 type-associated entries for metazoa and already >500 000 for bacteria.

NCBI Taxonomy currently lists >28 000 binomial fungal names at species rank, and 56% had good quality (not chimeric or broken) ITS records in GenBank (including synonyms). Our current data set of RefSeq sequences represents 16% of binomials with clean ITS data. In terms of classification (regardless of presence or absence of ITS data), the RefSeq set covers 660 of 4387 possible fungal genera, 249 of 514 possible families, 120 of 153 possible orders and 36 of 37 possible classes. With continued curation, more types will be identified and the associated sequences added to the RefSeq database.

The presence of curated type material improved the efficiency of taxonomic updates at NCBI. The validation of sequences and type material released in RefSeq allowed >300 taxonomic names to be merged, rectified or updated in NCBI Taxonomy. Several taxonomic names that were submitted with a genus and strain identifier only and not updated upon publication could easily be verified and updated by relying on accurate specimen data present in their sequence accessions. Similarly, curating and knowledge of synonyms are important because it can greatly influence the accuracy of any microbiome or ecological study. Recent studies on the oral microbiome provide a good example (59). Several researchers still continue to use large polyphyletic genus names to discuss species of clinical importance (60).

## Standards for traceability of specimen vouchers

The most time-consuming step in this curation process was to identify and verify type specimens and cultures. It was useful to import type identifiers from, for example, MycoBank, but identifiers must still agree with the metadata in the GenBank records. Using the same identifier from a specific collection (especially when specimen vouchers from herbarium material are involved) and a standard structure among various sources such as taxonomy databases, collections, publications and sequence records will contribute tremendously to improve this process. Listing type specimen identifiers in the abstract of a paper represents another helpful measure to avoid having type information hidden behind a paywall. Using an Entrez query such as this: (collection cbs[prop] OR cbs[title] AND fungi[orgn]) AND (2014/01/01: 3000[PDAT]) can help CBS collection curators, for example, to identify newly released sequence records since the beginning of 2014. The search term ‘CBS’ is just an example and can be replaced with any other acronym in the NCBI collections.

## NCBI Collections

The ability to provide direct links between GenBank and biorepositories (herbaria and fungaria, natural history collections, zoos, botanical gardens, biobanks, culture collections and others) relies on using unique identifiers to denote cultures and specimens. The potential pages to target with links have been expanded by several projects aimed at increasing the digital presence of a number of institutions. For example, the Mycology Collections data Portal (MyCoPortal; http://mycoportal.org/) provides direct access to digitized specimens records provided by The Macrofungi Collection Consortium, a collaboration of 35 institutions in 24 states in the USA (http://mycoportal.org/portal/index.php). The Global Plants Initiative (GPI) is another such effort, housed at the Royal Botanic Gardens, Kew. This is an international partnership of more than 300 herbaria in 72 countries. GPI's goal is to digitize and provide access to type specimens of plants, fungi and algae through community-contributed JSTOR Global Plants online database (http://plants.jstor.org). Other resources include Straininfo, which databases information related to cultures and strains (61).

NCBI has retained a record of all biorepositories to assist indexing of submissions in the NCBI Collections database (http://www.ncbi.nlm.nih.gov/projects/BioCollection/search_collection.cgi). The majority of fungal-related acronyms rely on unique identifiers of herbaria indexed at Index Herbariorum (62), whereas the majority of culture collections are listed in the directory of the World Federation for Culture Collections (WFCC). At NCBI, these unique identifiers are also used for museum collections. These multiple sources often contain redundant identifiers, so it is necessary to provide unique versions. This was achieved by adding a country abbreviation in angular brackets. For example, BR<BEL> was used to distinguish the National Botanic Garden of Belgium from the Embrapa Agrobiology Diazothrophic Microbial Culture Collection, BR<BRA>. For the present, it is more practical for NCBI to continue usage of this resource for its own curating and indexing functions. Another effort, the Global Registry of Biorepositories (GRBio) has been supported by the Consortium for the Barcode of Life (CBOL). This currently lists >7000 biorepository records by combining data from CBOL, Index Herbariorum and the Biodiversity Collections Index. It also lists >20 personal collections and allows for registrations online.

## Application of Darwin Core and other standards

Darwin Core is a data standard for publishing and integrating biodiversity information (15). The Darwin Core standard triplet format for specimen data consists of a structured string containing an institutional ID, collection code and catalog ID, all separated by colons. Currently, NCBI uses unique labels from the Collections database (14). In many cases, a secondary collection code (such as a collection devoted to Fungi or Plants at a specific institution) is not necessary. In the example given above, the ex-type culture of *Colletotrichum brevisporum* is indicated as a doublet only, e.g. /specimen_voucher="BR<BEL>:70109".

It is now possible to register typification events at MycoBank. MycoBank Typification numbers for the designation of lectotypes, epitypes and neotypes can be obtained and referred to in publication (63). The challenge remains to standardize voucher data, so it can be tracked consistently among multiple databases. In a future release of the MycoBank Web site scheduled for 2014, GenBank sequence identifiers will be requested upon deposition of new fungal names and/or associated type specimens. Some changes to the ICN to clarify the circumstances for epitypification have also been proposed (64).

## DNA barcoding and standards for GenBank submissions

The ITSx script (46) has been a helpful and time-saving tool in curating the ITS records. It has also provided an important quality control tool for anyone downloading and submitting ITS sequences. The script is efficient to confirm complete ITS regions if enough nucleotides are present in the 18S and 28S region, otherwise curation time needs to be spent to verify the coverage. In addition, sequences were also screened for non-ATGC characters, but ideally one would also like to be able to view sequence traces and be assured about the quality of the base calls. Currently, this information is not available for any ITS sequence. The standard for a DNA barcode (http://www.barcoding.si.edu/PDF/DWG_data_standards-Final.pdf) contains a set of sequence quality requirements in addition to increased scrutiny of specimen data. Part of this involves the deposit of trace data in addition to the sequence deposit at the INSDC. Currently, GenBank will assign a BARCODE keyword to sequences that meet these standards. However, many sequences continue to be referred to as DNA barcodes in the literature without meeting all these requirements. The deposit of sequence traces is a crucial missing element, and it is not likely to see an increase in the foreseeable future. Many highly significant sequences from types and other important specimens already exist in the INSDC databases. Sequences selected as part of this article should meet all the standards for a DNA barcode except for the deposition of trace data. It should therefore be also possible to use these sequences as 'barcode-like' or reference

sequences, although they would not formally qualify for barcode status.

## Effectiveness of ITS as barcode marker

The nuclear rRNA cistron consists of multiple copies ranging from a single copy to >200 in Fungi (65, 66). A number of processes can cause within-individual sequence heterogeneity in the ribosomal repeat, which complicates any analysis using ITS sequences. This includes intra- and intertaxon hybridization accompanied by lack of homogenization (concerted evolution) of the ribosomal repeat at some level in a wide range of species (67–72). Often the rate at which homogenization occurs and whether this varies from taxon to taxon is unknown. However, the process can be rapid (73). In genetically diverse interbreeding populations, however, the ribosomal repeat may never completely homogenize. Other heterogeneity can be due to variation between chromosomes in diploid or heterokaryotic specimens. It is also feasible that more than one ribosomal repeat could exist in some taxa as a consequence of hybridization or horizontal gene transfer (74). Collections selected as types or as exemplars for a species are often not completely homogenic. When heterogeneity is low, this has been handled by creating a consensus barcode using ambiguity codes as is commonly done for members of the *Glomeromycota* (75). In many cases, however, the level of genetic divergence between haplotypes or between copies of the ribosomal repeat can be significant (≥3% sequence divergence) (50, 76).

In addition to overestimating diversity, the ITS region can also underestimate diversity for several species groups (77). The search for alternative regions has already yielded several markers with equal or improved performance in specific lineages. During the last decade, phylogenetics has moved on from analyzing multiple genes to full genomes in a search for the true species phylogeny (42, 78–82). DNA barcodes have different criteria from phylogenetic markers, although they can often be used interchangeably (83). So the search for a single marker sequence that could represent an idealized phylogeny will most likely also yield a good candidate for a DNA barcode that could identify all Fungi.

## Defining types and reference sequences

Currently, the public sequence databases include a mix of sequences derived from type and non-type strains and with various degrees of curation and certainty. An improved and expanded nomenclature for sequences has been proposed elsewhere (41), based on an earlier proposal for 'gene types' (84). This work was done with a zoological perspective, addressing concepts formulated under the International Code of Zoological Nomenclature. To continue this discussion and present a system applicable to species codified under the ICN, we propose a simplified framework for consideration. Following this concept, species can be divided into several categories according to the combination of the type/reference strain status and of the sequence length/quality.

Sequences from type material should, when possible, be used preferentially for identification purposes, whereas the other sequences can be used for the description of the beta diversity within the species. However, we also introduced a concept of verified sequences to allow for small and manageable subset of taxonomically important sequences to be included in RTL. In the current data set, they constitute 5% of the total. Another factor influencing the reliability of molecular data—sequence length and quality. Given the enormous variability of fungal ITS sequences (Figure 2), it is difficult to establish a universal length threshold. Instead, such a threshold will be easier to define separately for taxa at the family or order rank and above. Similarly, low-quality sequences with too many degenerate sites can lead to non-authentic identifications. On a preliminary basis, it will be advantageous to calculate a guiding parameter, net sequence length, or the actual length reduced by the number of degenerated sites. A comprehensive length/quality index (LQI) could be defined by a simple equation:

$$\mathrm{LQI} = (\mathrm{SL} - \mathrm{DS})/\mathrm{LT}$$

where SL is the actual sequence length, DS the number of degenerated sites and LT the minimal sequence length to obtain a sound classification.

According to this LQI parameter, all the sequences currently presented in the RefSeq database exceed the minimal requirement for robust identification. In general, sequence databases could be managed according to simple rules, defining a hierarchy of sequences according to their origin, for example:

1. Type/reference sequences with high LQI are used for any purpose and serve as potential targets for the RefSeq database.
2. Type/reference sequences with low LQI are used for identification with a warning on the identification quality until they can be replaced with better sequences.
3. Non-type/reference sequences with high LQI can be used for any purpose, other than species identification.
4. Non-type/reference sequences with low LQI until better sequences are obtained.

The UNITE database for molecular identification of fungi represents another approach to improve sequence accuracy and fungal species identification. It comprises all fungal ITS sequences in INSDC and offers extended

functionalities for their curation and analysis (http://unite.ut.ee) (18). All sequences are clustered into SHs variously designated at 97–100% similarity (at 0.5% intervals) to seek to reflect the species rank. The SHs are assigned unique identifiers of the accession number type—e.g. SH133781.05FU—and are resolved with URLs such as 'http://unite.ut.ee/sh/SH158651.06FU'. All INSDC sequences that belong to an SH are hyperlinked from GenBank/ENA directly to that SH in UNITE through a LinkOut feature. More than 205 000 ITS accessions in the UNITE database can be accessed by using the query 'loprovunite[filter]' in GenBank. The SH concept is also implemented in the next-generation sequencing pipelines QIIME (85, 86), mothur (87), SCATA (http://scata.mykopat.slu.se/), CREST (88) and in the recently launched EU BOLD mirror (www.eubold.org). A total of ~21 000 SH or operational taxonomic units (OTUs) (excluding singletons) at 98.5% similarity are indexed in the current (sixth) release of UNITE. By default, a sequence from the most common sequence type in each SH is chosen to represent the SH and to form part of its name. It is also possible to change the chosen representative where there is a need to exercise extended control. Sequences from type material, in particular, are given priority whenever available and of satisfactory length and technical quality (18, 89).

## Conclusions

The Linnaean binomial has been a constant anchor in biology, and it remains central to communication in biology (90). It is intuitive to the way humans process information regarding the natural world, if not always in concert with shifting evolutionary concepts. Given the huge genetic diversity found within the kingdom Fungi, coupled with often cryptic and convergent morphologies, attempts to clearly delineate species boundaries remains a substantial challenge. This has led to a view that taxonomists might be better served by not focusing on fungal species names until more is known about their general biology (91). Although this might be a provocative view, even with ample DNA sequence data, debates about species boundaries will likely persist. A single name, linked to a specific specimen without dispute, following the rules and standards set down in the ICN will remain essential. It follows logically that if the same link can be made for DNA sequences, these sequences can provide reliable reference points for names in computational comparisons.

In this article, we have focused on the re-annotation of a taxonomically diverse set of marker sequences such that a clear link between a species name, a specimen or culture and its sequences can be established with a high level of certainty. The most important part of this process is the increased focus on specimen and culture annotations using a standardized format that can be traced across multiple databases. We used a number of redundant steps in the curation process to remove errors. Yet as is true for any database, some will remain. Because RefSeq is a fully curated database, relying on selections made by taxonomists at NCBI in consultation with a range of experts, it will also be simple to remove questionable sequences as soon as we are aware of them. Feedback about incorrect RefSeq records can be received here: http://www.ncbi.nlm.nih.gov/projects/RefSeq/update.cgi.

There is a substantial and growing increase in the number of sequences being deposited in public sequence databases without scientific binomials (10). To better distinguish truly novel lineages from poorly identified ones, an accurate set of reference sequences will be essential. The manual curation performed in this study relied on mining the information from a variety of resources, including the associated literature. This scales poorly beyond a few thousand entries. For this reason, we focused on a manageable subset of reference sequences focused on ties to type material. It is hoped that machine learning techniques, as already applied to taxonomic names from literature (92), can also improve specimen and, specifically, type material annotation in the future. Our initial data set of ~2600 ITS records should provide a valuable training set for such techniques.

It is projected that there are ~400 000 fungal names already in existence. Although only 100 000 are accepted taxonomically, it still makes updates to the existing taxonomic structure a continuous task. It is also clear that these named fungi represent only a fraction of the estimated total, 1–6 million fungal species (93–95). Moving forward, as new species are being described, this process must be documented in a more efficient manner, keeping track of the type specimen information in association with its sequence data. Submitters of newly generated fungal ITS sequences are also asked to consider previously published guidelines (96, 97).

We propose the following steps in submitting future type-related data as part of a normal submissions process to GenBank and the nomenclature databases. It is important to emphasize that Refseq selections will only happen after submission to INSDC databases and does not require a separate user-directed process.

1. Where possible, submitters can alert GenBank indexers to the presence of type material and include a table during their submissions:

<species name>\t <type strain/specimen>\t <type of type>, for example:

*Aspergillus niger* \t ATCC 16888 \t ex-neotype
*Agaricus chartaceus* \t PERTH 07582757 \t holotype
*Saccharomyces cerevisiae* \t CBS 1171 \t ex-type

2. If using ITS sequence data, use Figure 4 annotation as an exemplar during GenBank submission, applying annotations determined by the ITSx script.
3. Register typification with correctly annotated specimen numbers in an available database, e.g. MycoBank.
4. Verify that the format for specimens and ex-type cultures in GenBank match that in the published record as far as possible.
5. Extend the principles for traceable specimen and culture data during species descriptions in mycological journals.

Correctly formatted type specimen identifiers from public culture/herbarium collections should not be limited only to ITS records but also any other sequence records including genome records. Genome sequencing centers use ITS regions to confirm the identity of the fungus being sequenced. It is a good practice to include this with a genome assembly. However, the ribosomal cistron is often omitted because of difficulty to establish the exact copy number and the positions of the multiple copies in the genome. The RefSeq ITS set has already been applied in improving genome assembly quality at NCBI by identifying contamination in genome assemblies, especially in obligate biotroph genomes.

The increasing digitization of the biological literature and the growing availability of tools to search the literature and biorepositories are improving ways to link and contextualize sequences and biological data (9). The ability to semantically enhance journals will also allow future taxonomic papers to be mined for valuable taxonomic information (98). Type information is often found in a variety of formats that makes it challenging for machine reading. PubMed Central already has an initial species description extension in XML that could serve as a purpose for linking taxonomic data to additional metadata. This could include barcode data, and some shortened machine-searchable version could be placed in abstracts, so it is easily indexed in various openly accessible literature services like PubMed, PubMed Central and others without residing behind a paywall (99). We also advocate a newly available option to comment on papers in PubMed, PubMed Commons, by registering third-party opinions on sequences and species contained within the relevant publications (http://www.ncbi.nlm.nih.gov/pubmedcommons/).

In the immediate future, we will explore ways to streamline the expansion of RTL for Fungi. We currently only rely on selection by NCBI Taxonomy and RefSeq curators in consultation with numerous taxonomists. Cooperation with the nomenclature databases, MycoBank and Index Fungorum, as well as annotation and specimen databases such as UNITE and MyCoPortal should be expanded where possible. The focus on specimen and culture designations can be extended to include reliable standardized geographical data. We have also collected a smaller set of sequence accessions from the 28S nuclear ribosomal gene and will work to expand the set of accessions to include this and several other markers, using the re-annotated bio collection data where possible. Finally, working with partners to collect and sequence rare species in developing countries should be explored as well, ensuring the availability of annotated reference sequences to all potential users.

It seems likely that nomenclature will face increasingly radical changes in the future. DNA sequencing technology is rapidly revealing biodiversity information. Sequences obtained from environmental sampling can potentially be named under the current ICN with a DNA sample as a physical specimen, but this will not apply in many cases. This will require additional means to standardize labeling and to improve communication. Addressing this unsampled diversity may be 'the next major challenge for fungal taxonomy' (6). However, as we show here, much needs to be done to improve the way sampled diversity data are currently disseminated. During the next few years, several conversations will commence on ways to label sequences in public databases to facilitate sequence-based taxonomy (7). We hope the topics covered in this article will contribute to those discussions.

## Acknowledgements

B.R. and C.L.S. acknowledge support from the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

## Funding

Funding for open access charge: The Intramural Research Programs of the National Center for Biotechnology Information, National Library of Medicine and the National Human Genome Research Institute, both at the National Institutes of Health.

*Conflict of interest*. None declared.

## References

1. Hibbett,D.S., Ohman,A., Glotzer,D., *et al.* (2011) Progress in molecular and morphological taxon discovery in Fungi and options for formal classification of environmental sequences. *Fungal Biol. Rev.*, 25, 38–47.
2. Hawksworth,D.L., Crous,P.W., Redhead,S.A., *et al.* (2011) The Amsterdam declaration on fungal nomenclature. *IMA Fungus*, 2, 105–112.
3. McNeill,J., Barrie,F.R., Buck,W.R., *et al.* (2012) International Code of Nomenclature for algae, fungi, and plants (Melbourne Code). In: McNeill,J. (ed). *Regnum Vegetabile*, Vol. 154. Koeltz Scientific Books, Königstein, p. 240.

4. Redhead,S.A. and Norvell,L.L. (2013) Report of the Nomenclature Committee for Fungi 19: official repositories for fungal names. *Taxon*, 62, 173–174.
5. Kirk,P.M., Stalpers,J., Braun,U., *et al.* (2013) A without-prejudice list of generic names of fungi for protection under the International Code of Nomenclature for algae, fungi, and plants. *IMA Fungus*, 4, 381–443.
6. Hibbett,D.S. and Taylor,J.W. (2013) Fungal systematics: is a new age of enlightenment at hand? *Nat. Rev.,* 11, 129–133.
7. Taylor,J.W. and Hibbett,D.S. (2013) Toward sequence-based classification of fungal species. *IMA Fungus*, 4, 33–34.
8. Nakamura,Y., Cochrane,G. and Karsch-Mizrachi,I. (2013) The international nucleotide sequence database collaboration. *Nucleic Acids Res.,* 41, D21–D24.
9. Page,R.D. (2013) BioNames: linking taxonomy, texts, and trees. *PeerJ*, 1, e190.
10. Parr,C.S., Guralnick,R., Cellinese,N., *et al.* (2012) Evolutionary informatics: unifying knowledge about the diversity of life. *Trends Ecol. Evol.*, 27, 94–103.
11. Bidartondo,M.I., Bruns,T.D., Blackwell,M., *et al.* (2008) Preserving accuracy in GenBank. *Science*, 319, 1616.
12. Nilsson,R.H., Ryberg,M., Kristiansson,E., *et al.* (2006) Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS One*, 1, e59.
13. Pleijel,F., Jondelius,U., Norlinder,E., *et al.* (2008) Phylogenies without roots? A plea for the use of vouchers in molecular phylogenetic studies. *Mol. Phylogenet. Evol.,* 48, 369–371.
14. Federhen,S., Hotton,C. and Mizrachi,I. (2009) Comments on the paper by Pleijel *et al.* (2008): vouching for GenBank. *Mol. Phylogenet. Evol.,* 53, 357–358.
15. Wieczorek,J., Bloom,D., Guralnick,R., *et al.* (2012) Darwin Core: an evolving community–developed biodiversity data standard. *PLoS One*, 7, e29715.
16. Gardes,M. and Bruns,T.D. (1993) ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts. *Mol. Ecol.*, 2, 113–118.
17. Kõljalg,U., Larsson,K.H., Abarenkov,K., *et al.* (2005) UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytol.,* 166, 1063–1068.
18. Kõljalg,U., Nilsson,R.H., Abarenkov,K., *et al.* (2013) Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.*, 22, 5271–5277.
19. Santamaria,M., Fosso,B., Consiglio,A., *et al.* (2012) Reference databases for taxonomic assignment in metagenomics. *Brief. Bioinform.*, 13, 682–695.
20. Koetschan,C., Hackl,T., Muller,T., *et al.* (2012) ITS2 database IV: interactive taxon sampling for internal transcribed spacer 2 based phylogenies. *Mol. Phylogenet. Evol.*, 63, 585–588.
21. Liu,K.L., Porras-Alfaro,A., Kuske,C.R., *et al.* (2012) Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes. *Appl. Environ. Microb.*, 78, 1523–1533.
22. Mahe,S., Duhamel,M., Le Calvez,T., *et al.* (2012) PHYMYCO-DB: a curated database for analyses of fungal diversity and evolution. *PLoS One*, 7, e43117.
23. Öpik,M., Vanatoa,A., Vanatoa,E., *et al.* (2010) The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytol.,* 188, 223–241.
24. Park,B., Park,J., Cheong,K.C., *et al.* (2011) Cyber infrastructure for *Fusarium*: three integrated platforms supporting strain identification, phylogenetics, comparative genomics and knowledge sharing. *Nucleic Acids Res.,* 39, D640–D646.
25. Druzhinina,I.S., Kopchinskiy,A.G., Komon,M., *et al.* (2005) An oligonucleotide barcode for species identification in *Trichoderma* and *Hypocrea*. *Fung. Genet. Biol.*, 42, 813–828.
26. Hebert,P.D.N., Ratnasingham,S., deWaard,J.R. (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. Roy. Soc. Lond. B Bio.*, 270, S96–S99.
27. Hebert,P.D.N., Cywinska,A., Ball,S.L., *et al.* (2003) Biological identifications through DNA barcodes. *Proc. Roy. Soc. Lond. B Bio.*, 270, 313–321.
28. Ratnasingham,S. and Hebert,P.D.N. (2007) BOLD: the Barcode of Life Data System. *Mol. Ecol. Notes*, 7, 355–364. www.barcodinglife.org
29. Schoch,C.L., Seifert,K.A., Huhndorf,S., *et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. USA*, 109, 6241–6246.
30. Kiss,L. (2012) Limits of nuclear ribosomal DNA internal transcribed spacer (ITS) sequences as species barcodes for Fungi. *Proc. Natl. Acad. Sci. USA*, 109, E1811–E1811.
31. Walther,G., Pawlowska,J., Alastruey-Izquierdo,A., *et al.* (2013) DNA barcoding in Mucorales: an inventory of biodiversity. *Persoonia*, 30, 11–47.
32. Peterson,S.W. (2012) *Aspergillus* and *Penicillium* identification using DNA sequences: barcode or MLST? *Appl. Microbiol. Biotechnol.*, 95, 339–344.
33. Brock,P.M., Doring,H., Bidartondo,M.I. (2009) How to know unknown fungi: the role of a herbarium. *New Phytol.,* 181, 719–724.
34. Dentinger,B.T.M., Didukh,M.Y. and Moncalvo,J.M. (2011) Comparing COI and ITS as DNA barcode markers for mushrooms and allies (Agaricomycotina). *PLoS One*, 6, e25081.
35. Osmundson,T.W., Robert,V.A., Schoch,C.L., *et al.* (2013) Filling gaps in biodiversity knowledge for macrofungi: contributions and assessment of an herbarium collection DNA barcode sequencing project. *PLoS One*, 8, e62419.
36. Quaedvlieg,W., Kema,G.H., Groenewald,J.Z., *et al.* (2011) *Zymoseptoria* gen. nov.: a new genus to accommodate *Septoria*-like species occurring on graminicolous hosts. *Persoonia*, 26, 57–69.
37. Schell,W.A., Lee,A.G. and Aime,M.C. (2011) A new lineage in Pucciniomycotina: class Tritirachiomycetes, order Tritirachiales, family Tritirachiaceae. *Mycologia*, 103, 1331–1340.
38. Sohrabi,M., Myllys,L. and Stenroos,S. (2010) Successful DNA sequencing of a 75 year-old herbarium specimen of *Aspicilia aschabadensis* (J. Steiner) Mereschk. *Lichenologist*, 42, 626–628.
39. Trappe,J.M., Kovács,G.M. and Claridge,A.W. (2010) Comparative taxonomy of desert truffles of the Australian outback and the African Kalahari. *Mycol. Prog.*, 9, 131–143.
40. Larsson,E. and Jacobsson,S. (2004) Controversy over *Hygrophorus cossus* settled using ITS sequence data from 200-year-old type material. *Mycol. Res.,* 108, 781–786.
41. Taylor,J.W., Jacobson,D.J., Kroken,S., *et al.* (2000) Phylogenetic species recognition and species concepts in fungi. Fungal Genet. Biol., 31, 21–32.

42. Collins,R.A. and Cruickshank,R.H. (2012). The seven deadly sins of DNA barcoding. *Mol. Ecol. Resour.*, 6, 969–975.
43. Chakrabarty,P., Warren,M., Page,L., *et al.* (2013) GenSeq: an updated nomenclature and ranking for genetic sequences from type and non-type sources. *ZooKeys*, 346, 29–41.
44. Lutzoni,F., Kauff,F., Cox,C.J., *et al.* (2004) Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits. *Am. J. Bot.*, 91, 1446–1480.
45. Hibbett,D.S., Binder,M., Bischoff,J.F., *et al.* (2007) A higher-level phylogenetic classification of the Fungi. *Mycol. Res.,* 111, 509–547.
46. Bengtsson-Palme,J., Ryberg,M., Hartmann,M., *et al.* (2013) Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods Ecol. Evol.,* 4, 914–919.
47. Cornishbowden,A. (1985) Nomenclature for incompletely specified bases in nucleic-acid sequences—recommendations 1984. *Eur. J. Biochem.*, 150, 1–5.
48. Hall,T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.*, 41, 95–98.
49. Antonielli,L.V., Robert,L., Corte,L., *et al.* (2011) Centrality of objects in a multidimensional space and its effects on distance-based biological classifications. *Open Appl. Inform. J.*, 5, 11–19.
50. Simon,U.K., Weiß,M. (2008) Intragenomic variation of fungal ribosomal genes is higher than previously thought. *Mol. Biol. Evol.,* 25, 2251–2254.
51. Groenewald,J.Z., Groenewald,M., Braun,U., *et al.* (2010) *Cercospora* speciation and host range. In: Lartey,R.T., Weiland,J.J., Panella,L., *et al.* (eds.), *Cercospora Leaf Spot of Sugar Beet and Related Species*. APS Press, Minnesota, pp. 21–37.
52. Bensch,K., Braun,U. and Groenewald,J.Z., *et al.* (2012) The genus. Cladosporium. Stud. Mycol., 72, 1–401.
53. Braun,U., Crous,P. and Dugan,F., *et al.* (2003) Phylogeny and taxonomy of *Cladosporium*-like hyphomycetes, including *Davidiella* gen. nov., the teleomorph of *Cladosporium* s. str. *Mycol. Prog.*, 2, 3–18.
54. Kurtzman,C.P. (2014) Use of gene sequence analyses and genome comparisons for yeast systematics. *Int. J. Syst. Evol. Microbiol.*, 64, 325–332.
55. Petkovits,T., Nagy,L.G. and Hoffmann,K., *et al.* (2011) Data partitions, Bayesian analysis and phylogeny of the zygomycetous fungal family Mortierellaceae, inferred from nuclear ribosomal DNA sequences. *PLoS One*, 6, e27507.
56. Budziszewska,J., Piatkowska,J. and Wrzosek,M. (2010) Taxonomic position of *Mucor hiemalis* f. luteus. Mycotaxon, 111, 75–85.
57. Federhen,S. (2012) The NCBI taxonomy database. *Nucleic Acids Res.,* 40, D136–D143.
58. Hawksworth,D.L. (2013) Mycospeak and Biobabble. *IMA Fungus*, 4, 1–1.
59. Dupuy,A.K., David,M.S., Li,L., *et al.* (2014) Redefining the human oral mycobiome with improved practices in amplicon-based taxonomy: discovery of *Malassezia* as a prominent commensal. *PloS One*, 9, e90899.
60. Mukherjee,P.K., Chandra,J., Retuerto,M., *et al.* (2014) Oral mycobiome analysis of HIV-infected patients: identification of *Pichia* as an antagonist of opportunistic fungi. *PLoS Pathogens*, 10, e1003996.
61. Verslyppe,B., De Smet,W., De Baets,B., *et al.* (2014) StrainInfo introduces electronic passports for microorganisms. *Syst. Appl. Microbiol.*, 37, 42–50.
62. Thiers,B. (2014) *Index Herbariorum: A Global Directory of Public Herbaria and Associated Staff*. Botanical Garden's Virtual Herbarium, New York.
63. Robert,V., Vu,D., Amor,A.B.H., *et al.* (2013) MycoBank gearing up for new horizons. *IMA Fungus*, 4, 371–379.
64. Hawksworth,D.L. (2014) Possible house-keeping and other draft proposals to clarify or enhance the naming of fungi within the International Code of Nomenclature for algae, fungi, and plants (ICN). *IMA Fungus* 5, 31–37.
65. Baldrian,P., Vetrovsky,T., Cajthaml,T., *et al.* (2013) Estimation of fungal biomass in forest litter and soil. *Fungal Ecol.*, 6, 1–11.
66. Howlett,B.J., Rolls,B.D. and Cozijnsen,A.J. (1997) Organisation of ribosomal DNA in the ascomycete *Leptosphaeria maculans*. *Microbiol. Rev.,* 152, 261–267.
67. Garbelotto,M., Gonthier,P. and Nicolotti,G. (2007) Ecological constraints limit the fitness of fungal hybrids in the *Heterobasidion annosum* species complex. *Appl. Environ. Microbiol.*, 73, 6106–6111.
68. Hughes,K.W., Petersen,R.H., Lodge,D.J., *et al.* (2013) Evolutionary consequences of putative intra- and interspecific hybridization in agaric fungi. *Phytopathology*, 103, 63.
69. Lindner,D.L., Carlsen,T., Nilsson,R.H., *et al.* (2013) Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. *Ecol. Evol.,* 3, 1751–1764.
70. Schardl,C.L., Craven,K.D. (2003) Interspecific hybridization in plant-associated fungi and oomycetes: a review. *Mol. Ecol.*, 12, 2861–2873.
71. Dujon,B., Sherman,D., Fischer,G., *et al.* (2004) Genome evolution in yeasts. *Nature*, 430, 35–44.
72. Kovács,G.M., Balázs,T.K., Calonge,F.D., *et al.* (2011) The diversity of *Terfezia* desert truffles: new species and a highly variable species complex with intrasporocarpic nrDNA ITS heterogeneity. *Mycologia*, 103, 841–853.
73. Ganley,A.R.D. and Kobayashi,T. (2007) Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Res.,* 17, 184–191.
74. Xie,J., Fu,Y., Jiang,D., *et al.* (2008) Intergeneric transfer of ribosomal genes between two fungi. *BMC Evol. Biol.*, 8, 87.
75. Stockinger,H., Krüger,M. and Schüßler,A. (2010) DNA barcoding of arbuscular mycorrhizal fungi. *New Phytol.,* 187, 461–474.
76. Hughes,K.W., Petersen,R.H. and Lickey,E.B. (2009) Using heterozygosity to estimate a percentage DNA sequence similarity for environmental species' delimitation across basidiomycete fungi. *New Phytol.,* 182, 795–798.
77. Gaziz,S., Rehner,S. and Chaverri,P. (2011) Species delimitation in fungal endophyte diversity studies and its implications in ecological and biogeographic inferences. *Mol. Ecol.*, 20, 3001–3013.
78. Hibbett,D.S., Stajich,J.E. and Spatafora,J.W. (2013) Toward genome-enabled mycology. *Mycologia*, 105, 1339–1349.

79. James,T.Y., Kauff,F., Schoch,C.L., *et al.* (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, 443, 818–822.
80. Kurtzman,C.P. and Robnett,C.J. (2003) Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Res.,* 3, 417–432.
81. Rokas,A., Williams,B.L., King,N., *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425, 798–804.
82. Lewis,C.A., Satpal,B., Robert,V., *et al.* (2011) Identification of fungal DNA barcode targets and PCR primers based on Pfam protein families and taxonomic hierarchy. *Open Appl. Inform. J.*, 5, 30–44.
83. Wang,Z., Nilsson,R.H., Lopez-Giraldez,F., *et al.* (2011) Tasting soil fungal diversity with earth tongues: phylogenetic test of SATe alignments for environmental ITS data. *PLoS One*, 6, e19039.
84. Chakrabarty,P. (2010) Genetypes: a concept to help integrate molecular phylogenetics and taxonomy. *Zootaxa*, 2632, 67–68.
85. Caporaso,J.G., Kuczynski,J., Stombaugh,J., *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, 7, 335–336.
86. Bates,S.T., Ahrendt,S., Bik,H.M., *et al.* (2013) Meeting report: fungal ITS workshop (October 2012). *Stand. Genomic Sci.*, 8, 118–123.
87. Schloss,P.D., Westcott,S.L., Ryabin,T., *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, 75, 7537–7541.
88. Lanzen,A., Jorgensen,S.L., Huson,D.H., *et al.* (2012) CREST—classification resources for environmental sequence tags. *PLoS One*, 7, e49334.
89. Abarenkov,K., Tedersoo,L., Nilsson,R.H., *et al.* (2010) PlutoF-a web based workbench for ecological and taxonomic research, with an online implementation for fungal ITS sequences. *Evol. Bioinform.*, 6, 189–196.
90. Patterson,D.J., Cooper,J., Kirk,P.M., *et al.* (2010) Names are key to the big new biology. *Trends Ecol. Evol.,* 25, 686–391.
91. Money,N.P. (2013) Against the naming of fungi. *Fung. Biol.*, 117, 463–465.
92. Akella,L.M., Norton,C.N., Miller,H. (2012) NetiNeti: discovery of scientific names from text using machine learning methods. *BMC Bioinformatics*, 13, 211.
93. Hawksworth,D.L. (1991) The fungal dimension of biodiversity - magnitude, significance, and conservation. *Mycol. Res.,* 95, 641–655.
94. Blackwell,M. (2011) The Fungi: 1, 2, 3… 5.1 million species? *Am. J. Bot.*, 98, 426–438.
95. Taylor,D.L., Hollingsworth,T.N., McFarland,J.W., *et al.* (2013) A first comprehensive census of fungi in soil reveals both hyper-diversity and fine-scale niche partitioning. *Ecol. Monogr.*, 84, 3–20.
96. Nilsson,R.H., 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.
97. Hyde,K.D., Udayanga,D., Manamgoda,D.S., *et al.* (2013) Incorporating molecular data in fungal systematics: a guide for aspiring researchers. *Mycosphere Online*, 3, 1–32.
98. Penev,L., Agosti,D., Georgiev,T., *et al.* (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. *Zookeys*, 50, 1–16.
99. Satoh,K., Maeda,M., Umeda,Y., *et al.* (2013) *Cryptococcus lacticolor* sp. nov. and *Rhodotorula oligophaga* sp. nov., novel yeasts isolated from the nasal smear microbiota of Queensland koalas kept in Japanese zoological parks. *Antonie Van Leeuwenhoek*, 104, 83–93.