Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi

Conrad L. Schoch¹, Keith A. Seifert², Sabine Huhndorf, Vincent Robert³, John L. Spouge, C. André Levesque, Wen Chen, and Fungal Barcoding Consortium*²

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892; ‡Biodiversity (Myology and Microbiology), Agriculture and Agri-Food Canada, Ottawa, ON, Canada K1A 0C6; ³Department of Botany, The Field Museum, Chicago, IL 60605; and ⁴Centraalbureau voor Schimmelcultures Fungal Biodiversity Centre (CBS-KNAW), 3508 AD, Utrecht, The Netherlands

Edited* by Daniel H. Janzen, University of Pennsylvania, Philadelphia, PA, and approved February 24, 2012 (received for review October 18, 2011)

Six DNA regions were evaluated as potential DNA barcodes for Fungi, the second largest kingdom of eukaryotic life, by a multinational, multilaboratory consortium. The region of the mitochondrial cytochrome c oxidase subunit 1 used as the animal barcode was excluded as a potential marker, because it is difficult to amplify in fungi, often includes large introns, and can be insufficiently variable. Three subunits from the nuclear ribosomal RNA cistron were compared together with regions of three representative protein-coding genes (largest subunit of RNA polymerase II, second largest subunit of RNA polymerase II, and minichromosome maintenance protein). Although the protein-coding gene regions often had a higher percent of correct identification compared with ribosomal markers, low PCR amplification and sequencing success eliminated them as candidates for a universal fungal barcode. Among the regions of the ribosomal cistron, the internal transcribed spacer (ITS) region has the highest probability of successful identification for the broadest range of fungi, with the most clearly defined barcode gap between inter- and intraspecific variation. The nuclear ribosomal large subunit, a popular phylogenetic marker in certain groups, had superior species resolution in some taxonomic groups, such as the early diverging lineages and the ascomycete yeasts, but was otherwise slightly inferior to the ITS. The nuclear ribosomal small subunit has poor species-level resolution in fungi. ITS will be formally proposed for adoption as the primary fungal barcode marker to the Consortium for the Barcode of Life, with the possibility that supplementary barcodes may be developed for particular narrowly circumscribed taxonomic groups.

DNA barcoding | fungal biodiversity

The absence of a universally accepted DNA barcode for Fungi, the second most speciose eukaryotic kingdom (1, 2), is a serious limitation for multitaxon ecological and biodiversity studies. DNA barcoding uses standardized 500- to 800-bp sequences to identify species of all eukaryotic kingdoms using primers that are applicable for the broadest possible taxonomic group. Reference barcodes must be derived from expertly identified vouchers deposited in biological collections with online metadata and validated by available online sequence chromatograms. Interspecific variation should exceed intraspecific variation (the barcode gap), and barcoding is optimal when a sequence is constant and unique to one species (3, 4). Ideally, the barcode locus would be the same for all kingdoms. A region of the mitochondrial gene encoding the cytochrome c oxidase subunit 1 (CO1) is the barcode for animals (3, 4) and the default marker adopted by the Consortium for the Barcode of Life for all groups of organisms, including fungi (5). In Oomycota, part of the kingdom Stramenopila historically studied by mycologists, the de facto barcode internal transcribed spacer (ITS) region is suitable for identification, but the default CO1 marker is more reliable in a few clades of closely related species (6). In plants, CO1 has limited value for differentiating species, and a two-marker system of chloroplast genes was adopted (7, 8) based on portions of the ribulose 1,5-biphosphate carboxylase/oxygenase large subunit gene and a maturing-encoding gene from the intron of the tmk gene. This system sets a precedent for reconsidering CO1 as the default fungal barcode.

CO1 functions reasonably well as a barcode in some fungal genera, such as Penicillium, with reliable primers and adequate species resolution (67% in this young lineage) (9); however, results in the few other genera examined experimentally are inconsistent, and cloning is often required (10). The degenerate primers applicable to many Ascomycota (11) are difficult to assess, because amplification failures may not reflect priming mismatches. Extreme length variation occurs because of multiple introns (9, 12–14), which are not consistently present in a species. Multiple copies of different lengths and variable sequences occur, with identical sequences sometimes shared by several species (11). Some fungal clades, such as Neocallimastigomycota (an early diverging lineage of obligately anaerobic, zoosporic gut fungi), lack mitochondria (15). Finally, because most fungi are microscopic and inconspicuous and many are unculturable, robust, universal primers must be available to detect a truly representative profile. This availability seems impossible with CO1.

The nuclear rRNA cistron has been used for fungal diagnostics and phylogenetics for more than 20 y (16), and its components are most frequently discussed as alternatives to CO1 (13, 17). The eukaryotic rRNA cistron consists of the 18S, 5.8S, and 28S rRNA genes transcribed as a unit by RNA polymerase I. Posttranscriptional processes split the cistron, removing two internal transcribed spacers. These two spacers, including the 5.8S gene, are usually referred to as the ITS region. The 18S nuclear ribosomal small subunit rRNA gene (SSU) is commonly used in phylogenetics, and although its homolog (16S) is often used as a species diagnostic for bacteria (18), it has fewer hypervariable


The authors declare no conflict of interest.

* This Direct Submission article had a prearranged editor. Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in GenBank. Sequences are listed in Dataset S1.

† To whom correspondence may be addressed. E-mail: schoch2@ncbi.nlm.nih.gov or Keith.Seifert@AGR.GC.CA.

A complete list of the Fungal Barcoding Consortium can be found in the SI Appendix. This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1117018109/-DCSupplemental.
domains in fungi. The 28S nuclear ribosomal large subunit rRNA gene (LSU) sometimes discriminates species on its own or combined with ITS. For yeasts, the D1/D2 region of LSU was adopted for characterizing species long before the concept of DNA barcoding was promoted (19–21).

Currently, ~172,000 full-length fungal ITS sequences are deposited in GenBank, and 56% are associated with a Latin binomial, representing ~15,500 species and 2,500 genera, derived from ~11,500 scientific studies in ~500 journals. An important fraction of the sequences lacking binominals is from environmental samples (22, 23). In a smaller number of environmental studies, ITS has been used combined with LSU (24, 25). ITS is also used in some fungi for providing an indication of delimitation by a measure of the genetic distances (26). However, phylogenetic approaches are also being used to identify taxonomic units in environmental sampling of fungi (27) and are often more effective in comparison (28).

Protein-coding genes are widely used in mycology for phylogenetic analyses or species identification. For Ascomycota (including mold genera such as Aspergillus), they are generally superior to rRNA genes for resolving relationships at various taxonomic levels (29). Specialized identification databases use several markers [e.g., translation elongation factor 1-α for Fusarium (30) and β-tubulin for Penicillium (31)], but there is little standardization. Available primers for such markers usually amplify a narrow taxonomic range. Among protein-coding genes, the largest subunit of RNA polymerase II (RPB1) may have potential as a fungal barcode; it is ubiquitous and single copy, and it has a slow rate of sequence divergence (32). Its phylogenetic use was shown in studies of Basidiomycota, Zygomycota, Microsporidia (33–36), and some protists (37). RPB1 primers were developed for the Assembling the Fungal Tree of Life (AFToL) project, and the locus is included in the subsequent AFToL2 (38). However, its use as a barcode remains untested.

This paper stems from a multilaboratory, multinational initiative to formalize a standard DNA barcode for kingdom Fungi (excluding nonfungal organisms traditionally treated as fungi). We compared barcoding performance of three nuclear ribosomal regions (ITS, LSU, and SSU) and one region from a representative protein-coding gene, RPB1, based on probability of correct identification (PCI) and barcode gap analysis using newly generated sequences for representatives of the 17 major fungal lineages (Fig. 1). Contributors used standard primers and protocols developed by AFToL and submitted sequences to a customized database for analysis. Some also contributed sequences from regions of two additional optional genes, namely the second largest subunit of RNA polymerase II (RPB2; also an AFToL marker) (39) and a gene encoding a minichromosome maintenance protein (MCM7), which were chosen based on their usefulness in phylogenetic studies and ease of amplification across Ascomycota (40–42).

**Results**

We compared the barcoding performance of four markers using newly generated sequences from 742 strains or specimens, with two additional protein-coding markers analyzed for a smaller subset of about 200 fungi. Our taxon sampling was comprehensive and covered the main fungal lineages, with heavier sampling in the most speciose clades. Comparisons of PCI for all combinations of ITS, LSU, SSU, and RPB1 for all Fungi are shown in Figs. S1 and S2. We attempted to include Glomeromyces in the four-marker comparison, but RPB1 could only be amplified for some species of Glomeraceae. A simplified analysis of ITS vs. LSU in Glomeromyces (Fig. S3) indicated high levels of intraspecific variation in this group. We were unable to include Neocallimastigomycota and Microsporidia clades; arguments for and against their inclusion within Fungi continue (43, 44), although they presently are classified within the kingdom. For practical reasons, we had to assume that species concepts used by the taxonomists in the consortium were accurate and consistent, relying on the current circumscription of each species as assessed by the participants’ expertise. Genealogical concordance phylogenetic species recognition is commonly applied in mycology (45).

**PCR Success.** The survey (Fig. S4) showed that PCR amplifications of ribosomal RNA genes were more reliable across the Fungi than the protein-coding markers (Fig. 1). As expected, the success varied by taxonomic group [e.g., ITS PCR amplification success ranged from 100% (Saccharomycotina) to 65% (early

**Fig. 1.** Dendrogram of 17 fungal lineages sampled in this study showing consensus relationships and sampling. Relationships with high levels of uncertainty are indicated by stippled lines. Lineages are labeled and listed together with the approximate number of currently described species. The currently accepted node for delineating Fungi is indicated by F. The phyla Ascomycota and Basidiomycota are indicated by A and B, respectively. Gray bars to the left indicate numbers of strains in the barcode database, with the longest bar equal to 1,176 strains. Black bars indicate the proportions selected for a PCI analysis. The four datasets analyzed for PCI are numbered 1–4: 1, Pezizomycotina; 2, Saccharomycotina; 3, Basidiomycota; 4, early diverging lineages. Pie charts indicate the proportion of success from attempts to amplify the four-marker regions in the following order: ITS, LSU, SSU, and RPB1. Black, successful PCRs and sequences; gray, uncertain cases where no report was given; white, unsuccessful PCR.
diverging lineages). Ranges for the other ribosomal markers were similar. In comparison, success for RPB1 varied from 80% (Saccharomycotina) to 14% (basal lineages). About 80% of respondents reported no problems with PCR amplification of ITS, 90% scored it as easy to obtain a high-quality PCR product, and 80% reported no significant sequencing concerns. In comparison, >70% reported PCR amplification problems for RPB1; 40–50% reported primer failure as the biggest problem.

**Species Identification.** We performed several analyses to allow direct comparison of the barcoding use of the four main markers under consideration (i.e., ITS, LSU, SSU, and RPB1) (Figs. 2 and 3). To assess the PCI, data were divided into subsets according to taxonomic affinity. The combined four-marker PCI comparisons (Fig. 2) included 742 samples, with 142 species represented by more than one sample and 84 species represented with one sample. With all taxa considered, the PCI of ITS (0.73) was marginally lower than RPB1 (0.76). RPB1 consistently yielded high levels of species discrimination in all of the fungal groups except the early diverging lineages, which is comparable with multigene combinations (Fig. 2). Within Dikarya, ITS had the most resolving power for species discrimination in Ascomycotina (0.77 vs. 0.67 for RPB1). For Pezizomycotina, the PCI of RPB1 (0.80) outperformed ITS (0.71). ITS had lower discriminatory power than SSU and LSU in early diverging lineages, but margins of error were high. LSU had variable levels of PCI (0.66–0.75) among all groups but was generally lower than RPB1 or ITS (Fig. 2). In Saccharomycotina, LSU had the lowest PCI (0.67), but all four markers performed similarly. SSU was consistently the worst performing marker, with the lowest species discrimination in Pezizomycotina (Fig. 2) and Basidiomycota (Fig. 2). In the early diverging lineages (Fig. 2), SSU had a better PCI, on par with LSU and better than both ITS and RPB1.

In the multigene combinations, the most effective two genes in the combined analysis were either ITS and RPB1 or LSU and RPB1, both yielding a PCI of 0.78. This finding represented an increase of 0.02 from the highest-ranked single gene. The highest-ranked three- and four-gene combinations gave comparable increases.

Two supplementary three-marker comparisons expanded diversity for some major clades underrepresented in the four-gene analysis. For lichen-forming fungi, SSU was often absent, because the protocols favored amplicons from the photobiont rather than the fungus. Eliminating the requirement for SSU allowed more intensive sampling, yielding 683 sequences that included 179 species represented by more than one sample and 117 species represented by one sample (Fig. S5A). There was no apparent difference in ranking of the four candidate barcodes for the Pezizomycotina compared with the four-gene comparison in this analysis. Similarly, early diverging lineages yielded only 43 RPB1 sequences, and a comparison of ribosomal markers (ITS, SSU, and LSU) allowed inclusion of a larger set of 152 samples, with 34 species represented by more than one sample and 50 species by one sample. In this dataset, all sequences were unique to their species (Fig. S5B), and there was again no difference from the original four-gene comparison.

The barcode gap analyses (Fig. 3) largely confirmed the trends seen in the PCI analysis. The clearest indication of a barcode gap is seen for RPB1 followed by ITS. LSU and SSU performed poorly, each lacking a significant barcode gap.

---

**Fig. 2.** Barcode gap probability of identification for the four-marker datasets of ITS, LSU, SSU, and RPB1. The plots show the combinations of barcode markers investigated on the y axis, I, ITS; L, LSU; S, SSU; R, RPB1. The x axis shows the barcode gap PCI estimate for Ascomycota, Pezizomycotina (142 species), Basidiomycota (43 species), Ascomycota, Saccharomycotina (13 species), early diverging lineages (8 species), and combined groups (206 species). The error bars indicate 95% confidence intervals for the PCI estimate.

**Fig. 3.** Barcode gap analyses using distance histograms for each marker. Histograms display intraspecific variation in light gray and interspecific variation in dark gray. Inserts summarize distance data.
To test whether other single-copy protein-coding markers might have a similar barcoding performance to \( \text{RPB1}, \ \text{RPB2} \) and \( \text{MCM7} \) sequences were tested for a subset of taxa. Neither yielded data from the early diverging lineages, but a combination of remaining groups yielded 207 strains, including 55 species with more than one sample and 23 species with one sample (Fig. S6). For both markers, all sequences were unique to their species. The two supplementary genes had a similar barcoding performance to \( \text{RPB1} \), with \( \text{RPB2} \) yielding slightly superior results followed by \( \text{RPB1} \) and \( \text{MCM7} \).

### Discussion

Overall, ribosomal markers had fewer problems with PCR amplification than protein-coding markers (Fig. 1 and Fig. S4). Based on overall performance in species discrimination, SSU had almost no barcode gap (46) and the worst combined PCI, and it can be eliminated as a candidate locus (Figs. 2 and 3). LSU, a favored phylogenetic marker among many mycologists, had virtually no amplification, sequencing, alignment, or editing problems, and the barcode gap was superior to the SSU. However, across the fungal kingdom, ITS was generally superior to LSU in species discrimination and had a more clearly defined barcode gap (Fig. 3). The overall probability of correct species identification using ITS is comparable with the success reported for the two-marker plant barcode system (0.73 vs. 0.70) (7). Higher species identification success can be expected in the major macrofungal groups in \( \text{Basidiomycota} \) (0.79), and slightly lower success can be expected in the economically important microfungal groups in filamentous \( \text{Ascomycota} \) (0.75). ITS performed as a close second to the most heavily sampled of our protein-coding markers, \( \text{RPB1} \). However, the much higher PCR amplification success rate for ITS is a critical difference in its performance as a barcode (Fig. 1). ITS primers used in this study were applied to a range of fungal lineages, and several primers function as almost universal primers. However, all primer sets have a range of biases, and an appropriate solution will be to use more than one primer combination (47).

Taking all these arguments into account, we propose ITS as the standard barcode for fungi. The proposal will satisfy most fungal biologists but not all. Given the fungal kingdom’s age and genetic diversity, it is unlikely that a single-marker barcode system will be capable of identifying every specimen or culture to species level. Furthermore, the limitations of ITS sequences for identifying species in some groups and the failure of the universal ITS primers to work in a minority of other groups will have to be carefully documented (14, 43, 48). ITS sequences shared among different species have already been documented in species-rich \( \text{Pezizomycotina} \) genera with shorter amplicons, such as the economically important genera \( \text{Cladosporium} \) (49), \( \text{Penicillium} \) (50), and \( \text{Fusarium} \) (51). In \( \text{Aspergillus} \), ITS sequences are identical in several complexes of critical mycotoxicogenic, industrial, and medically important species, and additional markers are necessary (52). Although the ITS region is a potentially effective DNA barcode in several lichenized lineages (53), genetic drift may prevent lineage sorting of ancestral polymorphisms in some slowly evolving groups.

Intrageneric variation, such as the existence of multiple paralogous or nonhomologous copies within single fruiting bodies of basidiomycetes (54, 55) and ascomycetes (56) or within axenic cultures (51), may lead to higher estimates of infraspecific variability (57, 58) or generation of barcodes that act only as representative sequences of multiple variable repeats (59, 60). Highly variable lengths and high evolutionary rates for the nuclear ribosomal cistron in species of \( \text{Cryptococcus, Tulasnella (Cantherelles, Basidiomycota)} \) (61–63), and some lichens (53) may provide challenges for sequencing and analysis. The upper range of this ITS region variation is likely found in the \( \text{Glomeromycota} \), with up to 20% divergence within a single multicellulate spore (64, 65).

We acknowledge that species delimitations vary from one fungal group to another and are often influenced by scarcity of sampling and lack of detailed biological knowledge (43, 45). This influence is reflected when ITS distances are compared between phyla, subphyla, and species (Figs. S7 and S8). In an expanded dataset of ITS sequences from our fungal DNA barcoding database, the highest variation was most often found in the early diverging lineages. This finding confirms the fact that fungal diversity remains poorly sampled with DNA sequences for these lineages (43, 48). It is, therefore, very likely that high divergence reflects the presence of multiple cryptic species, indicating important focal points for additional study. Despite these challenges, ITS combines the highest resolving power for discriminating closely related species with a high PCR and sequencing success rate across a broad range of fungi.

In addition to ITS, LSU may also be applicable as a barcode for other organisms. ITS has already been shown in \( \text{Chlorophyta} \) and \( \text{Glomeromycota} \) (66, 67) as well as in \( \text{Oomycota} \) (68). The possibility of multi-kingdom analyses of complex ecosystems like soil using the species-informative, stable, high copy number ITS mirrors the original vision of DNA barcoding, and it already seems feasible, for example, to amplify \( \text{Fungi} \) and other eukaryotes from soil (23).

Protein-coding genes are popular phylogenetic markers in mycology, and they are used as de facto barcodes for limited taxonomic scope in several groups of fungi. We chose \( \text{RPB1} \) as a representative marker to include in our broad comparisons, with \( \text{RPB2} \) and \( \text{MCM7} \) analyzed for a smaller sampling. In general, such protein markers had more species resolving power, but PCR and sequencing failures eliminate them as potential universal barcodes for the broad phylogenetic scope of the kingdom \( \text{Fungi} \). Reliable kingdom-wide PCR amplification needs to be tested for other widely used protein-coding markers, such as translation elongation factor 1-α, β-tubulin, or actin.

The possibility of a two-marker barcoding system for fungi, as adopted for plants, is often discussed among mycologists, particularly researchers working on ascomycetous yeasts (19–21) and \( \text{Glomeromycota} \) (68) who prefer a system combining ITS and LSU. Data from this study (Fig. S5) indicate that ITS and LSU perform very similarly as barcodes and that differences in these sequences correlate well with current species concepts. Combinations of both ITS and LSU sequences are also applied in environmental sampling (69), where tandem amplification can allow simultaneous species identification with ITS and phylogenetic analysis with LSU. Our analyses with two-, three-, or four-marker barcode systems (Fig. 2) reveal only a modest increase in the PCI over a single-marker ITS barcode. The need for a second marker depends on the intended purpose of an investigation (i.e., whether a broad and general survey is intended or whether particular critical species are being monitored). If these taxa are taxa with low ITS interspecific variability, secondary markers must be used to accurately report genetic diversity (70). Genome mining efforts have identified a few single-copy genes that might be amenable for broad-range priming, and these efforts should continue (71, 72).

Although the genome diversity of fungal species is studied with increasing intensity, the vast majority of fungal species remains unknown. The recent discovery of a ubiquitous fungal class from soil (73) and a diverse early diverging phylum, \( \text{Cryptomycota} \), tied to \( \text{Rozella} \) (74–76) from riverine and marine sites illuminates this fact. More than 90% of \( \text{Fungi} \) may be awaiting discovery, posing a tremendous pressure to increase the pace of fungal species discovery (1, 2). In addition to this, the Melbourne Botanical Congress has recently approved large-scale changes to the process of naming fungi (77), and sequence data from type specimens will increasingly be essential to the stability of fungal nomenclature. Continuing discovery of novel biodiversity while classifying knowledge already available will demand well-coordinated initiatives, and DNA barcoding has a crucial role to play.
Materials and Methods
DNA Isolation, Amplification, and Sequencing. DNA was isolated and purified from cultures or specimens using the methods routinely used by the participating laboratories. Similarly, PCR protocols (Table 5) and thermocyclers varied from laboratory to laboratory and were those from the AFTOL project (Table 5). Several samples were sent by contributors for PCR amplification and sequencing at LifeTech. For PCR at LifeTech, 1–2 μL fungal DNA were amplified in a final volume of 30 μL with 15 μL AmpliTaq Gold 360 Mastermix, PCR primers, and water. All forward primers contained the M13F-20 sequencing primer, and reverse primers included the M13R-27 sequencing primer. PCR products (3 μL) were electroeluted cleaned before sequencing with 1 μL ExoSap-IT and 1 μL Tris EDTA and incubated at 37 °C for 20 min followed by 80 °C for 15 min. Cycle sequencing reactions contained 5 μL cleaned PCR product, 2 μL BigDye Terminator v3.1 Ready Reaction Mix, 1 μL 5x Sequencing Buffer, 1.6 pmol M13F or M13R sequencing primer, and water in a final volume of 10 μL. The standard cycle sequencing protocol was 27 cycles of 10 s at 96 °C, 5 s at 50 °C, 4 min at 60 °C, and hold 4 °C. Sequence cleaning was performed with the BigDye XTerminator Purification Kit as recommended by the manufacturer for 10-μL volumes. Sequence reactions were analyzed on a 3730xl Genetic Analyzer.

Sampling. Closely related but separately named asexual and sexual species were coded with one genus name and sex and sexual species were coded with one genus name and sex and sexual species were coded with one genus name and sex and sexual species were coded with one genus name and sex and sexual species were coded with one genus name and sex and sexual species were coded with one genus name and sex and sexual species were coded with one genus name and sex and sexual species were coded with one genus name and sex and sexual species were coded with one genus name and sex. All alignments used Mol Ecol Resour (with ITS, LSU, and RPB2 as a test case).

Participants recorded their experience on the success of PCR amplification and sequencing. Closely related but separately named asexual and sexual species are the greatest distance apart that two species can be used to identify a species. In this study, they also documented species and their overall taxonomic and phylogenetic use in specialized taxonomic groups. Comments were parsed to identify taxon-specific problems and are summarized in Fig. S4.

Data Analyses. Database. A query-based BioLomics database (78) was established for 2,920 strains (1,022 species including subspecies) provided by >70 members of the consortium (www.fungalbarcoding.org). The complete database sets consist of 213 different genera and 915 unique species; there are an average of four species per genus and three strains per species. Approximately one-third (1,029) of the strains were scored as sibling species of other species in the sample, with 156 unique sibling species groups. All data are based on deposited voucher specimens or cultures identified by taxonomic specialists. The database allowed pairwise sequence alignments or polyphasic identifications using one or any combination of the six genes used in this study. The taxon sampling covered 15 of 17 major lineages attributed to the Fungi (Fig. 1) that were weighted to species-rich higher taxa such as the Pezizomycotina (the largest group of Ascomycota) and the Agaricomycotina (mushrooms and other macrobasidiomycetes).

PCI. For each dataset, we calculated the barcode gap PCI. All alignments used the BLAST default DNA scoring system (79, 80). Two kinds of alignment were calculated between every sample pair, namely (i) a global alignment using the Needleman–Wunsch algorithm, which aligns the entire sequence length with penalties for gaps at the alignment ends (81), and (ii) a semiglobal alignment using a variant Needleman–Wunsch algorithm that includes both ends of one sequence and finds the alignment with the highest score without penalizing end gaps in the other sequence. The latter algorithm does the same for the other sequence, returning the alignment with the higher of the two scores. Thus, the global alignment matches the whole length of two sequences, and the semiglobal alignment matches one sequence to a subset of the other and then vice versa. Semiglobal alignment checks whether disparate sequence lengths degrade species identification; if they do not, global and semiglobal alignment should result in similar identifications. For the two types of alignment, the p-distance (the proportion of aligned nucleotide pairs consisting of differing nucleotides) was calculated. The sequence diameter of a species is defined as the greatest p-distance between any two samples from within a species. Based on the sequence diameter, the correct identification of a species occurs if, for every sample in the species, no sample from another species lies within the sequence diameter. The corresponding barcode gap PCI is the fraction of species correctly identified (7). The Wilson score interval yielded 95% confidence intervals for each PCI estimate (82). PCI was also calculated for all possible combinations of two, three, or four genes to evaluate the potential payoff of a multigene barcoding system.

Sequence divergence and DNA gap analyses. Using the same dataset as for the PCI analysis, a DNA barcode gap analysis was performed using matrix algebra and this was supported with the Intramural Research Program of the National Institutes of Health, the National Institutes of Science, and Statistic Analysis Software (SAS Institute) as described previously (6) except that the lower triangular uncorrected distance matrix was calculated using mom8 (83). The results are indicated in Fig. 3. Additional comparisons were done and are described in Figs. S2, S3, and S7–59.

ACKNOWLEDGMENTS. We thank David L. Hawksworth, Martin Bidartondo, and numerous other colleagues for critical comments. This work was organized under the fungal Working Group of the Consortium for the Barcode of Life (CBOL), which provided support from its funding from the Alfred P. Sloan Foundation. Support was also provided by the Intramural Research Program of the National Institutes of Health, the National Institutes of Science, and Statistic Analysis Software (SAS Institute) as described previously (6) except that the lower triangular uncorrected distance matrix was calculated using mom8 (83). The results are indicated in Fig. 3. Additional comparisons were done and are described in Figs. S2, S3, and S7–59.


