Formal description of sequence-based voucherless Fungi: promises and pitfalls, and how to resolve them

Robert Lücking¹ and David L. Hawksworth²

¹Botanischer Garten und Botanisches Museum, Freie Universität Berlin, Königin-Luise-Strasse 6–8, 14195 Berlin, Germany; corresponding author e-mail: r.luecking@bgbm.org
²Department of Life Sciences, The Natural History Museum, Cromwell Road, London SW7 5BD, UK; and Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Surrey TW9 3DS, UK; Jilin Agricultural University, Changchun, Jilin Province, 130118 China

Abstract: There is urgent need for a formal nomenclature of sequence-based, voucherless Fungi, given that environmental sequencing has accumulated more than one billion fungal ITS reads in the Sequence Read Archive, about 1,000 times as many as fungal ITS sequences in GenBank. These unnamed Fungi could help to bridge the gap between 115,000 to 140,000 currently accepted and 2.2 to 3.8 million predicted species, a gap that cannot realistically be filled using specimen or culture-based inventories. The Code never aimed at placing restrictions on the nature of characters chosen for taxonomy, and the requirement for physical types is now becoming a constraint on the advancement of science. We elaborate on the promises and pitfalls of sequence-based nomenclature and provide potential solutions to major concerns of the mycological community. Types of sequence-based taxa, which by default lack a physical specimen or culture, could be designated in four alternative ways: (1) the underlying sample (‘bag’ type), (2) the DNA extract, (3) fluorescent in situ hybridization (FISH), or (4) the type sequence itself. Only (4) would require changes to the Code and the latter would be the most straightforward approach, complying with three of the five principal functions of types better than physical specimens. A fifth way, representation of the sequence in an illustration, has been ruled as unacceptable in the Code.

Potential flaws in sequence data are analogous to flaws in physical types, and artifacts are manageable if a stringent analytical approach is applied. Conceptual errors such as homoplasy, intragenomic variation, gene duplication, hybridization, and horizontal gene transfer, apply to all molecular approaches and cannot be used as a specific argument against sequence-based nomenclature. The potential impact of these phenomena is manageable, as phylogenetic species delimitation has worked satisfactorily in Fungi. The most serious shortcoming of sequence-based nomenclature is the likelihood of parallel classifications, either by describing taxa that already have names based on physical types, or by using different markers to delimit species within the same lineage. The probability of inadvertently establishing sequence-based species that have names available is between 20.4 % and 1.5 % depending on the number of globally predicted fungal species. This compares favourably to a historical error rate of about 30 % based on physical types, and this rate could be reduced to practically zero by adding specific provisions to this approach in the Code. To avoid parallel classifications based on different markers, sequence-based nomenclature should be limited to a single marker, preferably the fungal ITS barcoding marker; this is possible since sequence-based nomenclature does not aim at accurate species delimitation but at naming lineages to generate a reference database, independent of whether these lineages represent species, closely related species complexes, or infraspecies. We argue that clustering methods are inappropriate for sequence-based nomenclature; this approach must instead use phylogenetic methods based on multiple alignments, combined with quantitative species recognition methods. We outline strategies to obtain higher-level phylogenies for ITS-based, voucherless species, including phylogenetic binning, ‘hijacking’ species delimitation methods, and temporal banding. We conclude that voucherless, sequence-based nomenclature is not a threat to specimen and culture-based fungal taxonomy, but a complementary approach capable of substantially closing the gap between known and predicted fungal diversity, an approach that requires careful work and high skill levels.

Article info: Submitted: 2 July 2017; Accepted: 15 May 2018; Published: 22 May 2018.
BACKGROUND


The structured query <Fungi[organism] AND (5.8S[title] OR ITS1[title] OR ITS2[title] OR ITS[title])> returned over 1 million (1,042,545) ITS sequences from GenBank (Benson et al. 2013, https://www.ncbi.nlm.nih.gov/genbank) on 19 Oct. 2017). The unstructured query <(Fungi or fungal) AND (5.8S OR ITS1 OR ITS2 OR ITS or “internal transcribed spacer”)> returned only a slightly higher number (1,065,267). This impressive number corresponds to approximately 30 years of sequencing work. Since 2009, the Sequence Read Archive (SRA; Leinonen et al. 2011, Kodama et al. 2012) stores data obtained from environmental sequencing studies. Using the unstructured query above (since the structured query does not work in the SRA), the SRA returned 246 studies, 2144 biosamples (= environmental samples), and 20,879 experiments (= NGS runs). Excluding 71 experiments with zero sequences and weighting the remaining 20,818 experiments as 1 (exclusively fungal), 0.5 (mixed fungal and bacterial), and 0 (likely low presence of fungal sequences); we estimated that these data contained over 1 billion fungal ITS reads (1,222,062,203), with an average length of 375 bases (SRA: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=search_obj) on 19 Oct. 2017; see Suppl. File S1). Thus, at present there are 1,000 times more NGS reads in the SRA than sequences in GenBank for the fungal barcoding marker (Fig. 1). Only three years ago, this ratio amounted to about 20:1 (Lücking 2014), which means it has grown by the factor 50 in this short time period and is expected to further increase exponentially, considering the growth rate of SRA data (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi).

A substantial proportion of the approximately 1 million fungal ITS sequences in GenBank is unidentified or wrongly labeled or represents unrecognized contaminants (Harris 2003, Vilgays 2003, Nilsson et al. 2005, 2006, 2012, 2014, Meier 2008, Bidartondo et al. 2008, Lücking et al. 2012, Köljalg et al. 2013). Sixty percent of these correspond to ‘uncultured’ Fungi at best identified to genus level, but most often not identified at all. The number of taxa sequenced is only a portion of all currently accepted Fungi, about 35,000 out of 120,000 (C. Schoch, pers. comm., Hawksworth & Lücking 2017). Only properly identified and labeled sequences can be used as reference for accurate fungal identification using ITS barcoding, and clearly there is a need to quickly increase and eventually complete this reference library for all Fungi (Meier 2008, Begerow et al. 2010, Köljalg et al. 2013, Tanabe & Toju 2013). While the situation is bad in GenBank, the over 1 billion fungal ITS reads in the SRA are not named at all. These data encompass thousands, perhaps tens or hundreds of thousands of novel taxonomic units, from species to class level, and hence provide a substantial source of potential reference sequences far beyond GenBank. To serve as such, they need to be named (Hibbett et al. 2011, 2016, Hawksworth et al. 2011, 2016, Hibbett & Taylor 2013, Lücking 2014, Minnis 2015, Hibbett 2016). An informal naming system that is not compatible with formal nomenclature, such as the ‘species hypotheses’ in UNITE (Köljalg et al. 2013), is impractical as a reference library, as informal names or numbers remain obscure without a broadly accepted, formal naming framework. Another shortcoming of a curated database is the amount of data to be handled. UNITE has approximately 800,000 fungal ITS sequences, close to 75% of what is deposited in GenBank, and corresponding to over 70,000 species hypotheses at 98.5% similarity threshold (https://unite.ut.ee) on 19 Oct. 2017. To deal with SRA reads in a similar way, UNITE would have to add about 1,000 times that number, with an exponential increase in the foreseeable future, an amount of data that is virtually ‘incuratable’ as so-called ‘species hypotheses’. Also, clusters based on a fixed threshold do not necessarily correspond accurately to species (see below).

How many Fungi exist is unknown. The number of accepted species estimated oscillates between 115,000 and
140 000 (Roskov et al. 2016, Species Fungorum http://www.speciesfungorum.org), with a figure of 120 000 presumed reasonable (Hawksworth & Lücking 2017); these variations are attributable largely to allowances made for synonymy and separately named morphs for the same species. Estimates for global fungal species richness range from 611 000 to 10 million (Hawksworth 1991, 2001, 2012, O’Brien et al. 2005, Schmitt & Mueller 2007, Blackwell 2011, Bass & Richards 2011, Mora et al. 2011), with an often-cited number of 1.5 million and a recent estimate of 2.2–3.8 million (Hawksworth & Lücking 2017). Even with an estimate of 1.5 million, a complete inventory of all Fungi on Earth using traditional methods within a reasonable time frame is impossible, given that it took 250 years to discover and describe less than 10 % of that diversity. Furthermore, natural habitats harbouring unknown species are being destroyed at an accelerated rate before they can be inventoried, as a result of the Sixth Mass Extinction (Leakey & Lewin 1995, Wake & Vredenburg 2008, Barnosky et al. 2011).

While molecular approaches have revolutionized our understanding of fungal diversity, they have not substantially increased the speed of discovery and formal description of new species. In the two decades prior to the onset of molecular systematics, the average number of newly described species per year was about 1250, slightly increasing to about 1300 in the two decades between 1990 and 2010. With the growth of species delimitation approaches around 2010, this number stands now at about 1750 per year (Hawksworth & Lücking 2017). To classify most or all Fungi within a reasonable time, this rate would have to increase by an order of magnitude, an impossible prospect considering the already limited resources of the mycological community and the diminishing number of fungal taxonomists (Gams 1997, Korf 2005, Meier 2008, Hawksworth 2009, Gryzenhout et al. 2012, Rambold et al. 2013). NGS offers a new approach to fungal inventories, allowing fast detection of a broad range of taxa in a relatively short time and at low cost (Hibbett et al. 2011, Grantham et al. 2015, Hibbett 2016). Numerous novel Fungi have already been discovered from environmental samples, including at higher taxonomic ranks (Jones et al. 2011a, b, Rosling et al. 2011, Livermore & Mattes 2013, Glass et al. 2014, Tedersoo et al. 2014, 2017, Lazarus & James 2015, De Beer et al. 2016, Nilsson et al. 2016). The setback of this approach is that the only manifestation of these Fungi are sequence data, unless taxa are successfully cultured, resequenced, and matched to previously obtained sequences (Rosling et al. 2011, Menkis et al. 2014, De Beer et al. 2016).

Box 1. What is a reasonable time frame to classify most or all Fungi?

With a rate of 2000 species described per year, it would take 680, 1030, or 1830 years to complete cataloguing 1.5, 2.2, or 3.8 million species respectively. Increasing our current efforts by an order of magnitude, it would take between 68 and 183 years. A more realistic increase, e.g. 5,000 species per year, would result in 272 to 732 years. A “reasonable” time frame is arbitrary, but several factors can be considered. For one, conversion of natural ecosystems is accelerating, and the Sixth Mass Extinction is well under way. Second, global climate change will have substantial effects predicted within the next few decades. Third, there is the life-span of a scientific career, usually 50 years or less. Facing accelerated habitat destruction, global climate changes, and the fact that we would like to see substantial progress in terms of cataloguing fungal diversity within the life span of our own careers, including young mycologists now starting at the age of about 25, we can set a reasonable time frame at 50 years. At the current rate, we would then formally describe another 100,000 species if we stick to specimen- and culture-based inventories, for a total of 240 000, 16 % of 1.5 million and 6 % of 3.8 million. To get much closer to the goal, we would have to increase our efforts by an order of magnitude, achieving 1 million new species within 50 years or an impressive 76 % of 1.5 million, but still only 30 % of 3.8 million. A 10-fold increase of specimen- or culture-based inventories is, however, not going to happen; at best we could reach 5000 species per year, for a total of 250 000 in 50 years. With that, we approach 26 % of 1.5 million but only 10 % of 3.9 million, and we would barely reach the number of plant species. Therefore, if the mycological community cannot convincingly argue that we will reach a 10-fold increase of our specimen- and culture-based efforts almost instantly, there is no alternative to including voucherless, sequence-based formal nomenclature as an additional method of classifying fungal diversity if we want to see substantial progress during the next five decades.

An example illustrating the problem is the class Archaeorhizomycetes, originally established on a single genus and species, Archaeorhizomycetes finlayi, with a second species described later, both based on permanently preserved cultures (Rosling et al. 2011, Menkis et al. 2014). Based on additional sequences from GenBank and a limited sample from the SRA, this class was estimated to contain close to 500 species (Menkis et al. 2014). SRA blast search on 9 Nov. 2014 using the ITS sequence of the type species retrieved 106 563 reads from environmental samples belonging to this class (Smith & Lücking, unpubl.; Suppl. File S2). With the overall increase in fungal ITS reads in the SRA by the factor 50 since 2014, this number could now potentially amount to about 5 million. Analysis of this data set using clustering through USEARCH (Edgar 2010) suggests the presence of between 28 435 species at 99 % threshold level and 2,658 species at 95 %; with the UNITE ‘species hypothesis’ threshold of 98.5 %, the estimated number of species would be 16 231. Preliminary phylogenetic analysis based on multiple sequence alignment suggests around 1000 taxa, apparently corresponding to separate genera, families, and perhaps orders within the class Archaeorhizomycetes. Irrespective of the accurate number, the magnitude of the problem is illustrated by the fact that, since the original
discovery of this class, only two valid names have been established (a rate of 0.3 per year). Therefore, adhering to the requirement of physical type specimens, including cultures, for the valid description of novel Fungi detected through environmental sequencing, is illusory. It is inconceivable that a sizable proportion of Fungi from environmental sequencing reads will ever be documented through specimen-based fungal inventories or culturing. Culturing only detects a portion of the fungal diversity present in a sample (Arnold et al. 2007, De Beer et al. 2016), and the examples of Archaeorhizomycetes, Hawksworthiomycetes, and Cryptomycota (Livermore & Mattes 2013, Letcher et al. 2013, Lazarus & James 2015, De Beer et al. 2016), show that culturing hardly makes a dent into the huge number of Fungi to be formally described from environmental samples, simply because there are no capacities for a global approach to catalogue millions of species that way. The Westerdijk Fungal Biodiversity Institute (CBS; formerly the CBS-KNAW Fungal Biodiversity Institute and Centraalbureau voor Schimmelcultures) and the ARS Culture Collection (NRRL) are the largest public service fungal culture collections in the world, with about 50,000 and 68,000 strains, respectively (for CBS see below, for NRRL data provided by T. Adkins, pers. comm. 2017). Both have contributed substantially to fungal ITS sequences in GenBank (https://www.ncbi.nlm.nih.gov/biocollections/?term=cbs; https://www.ncbi.nlm.nih.gov/biocollections/3689). The search string <Fungi AND CBS AND (5.8S OR ITS1 OR ITS2 OR ITS OR “internal transcribed spacer”)> returned 37 680 fungal entries, including almost 10 % of sequences from type material on 19 Oct. 2017; just <Fungi AND CBS> returns 863 723 fungal entries. For NRRL, there are 5340 ITS sequences and 209 624 fungal entries overall. Over 90 % of the CBS entries are identified to species, corresponding to nearly 9000 taxa. While this level of resolution is impressive, the identified taxa constitute just 7 % of the currently accepted Fungi, a proportion that decreases to far less than 1 % if we assume up to 3.8 million predicted species.

Even if CBS and other large fungal culture collections could increase their efforts by an order of magnitude, culture-based fungal inventories would still be incapable of dealing with even the most conservative species-richer predictions in a reasonable time frame. CBS had 51 908 fungal strains corresponding to 15 526 species on 5 Dec. 2017 (www.westerdijkinsitute.nl/Colletions/localfiles/CBSStrain- s.July21st2016.zip). A ten-fold increase to about 500 000 strains, apart from being logistically challenging, if not impossible, may increase the number of taxa to about 150 000. If we assume three large culture collections, with a taxonomic overlap of 50 %, we would stand at 300 000 taxa. Thus, an already impossible effort by the two cited large culture collections to augment their capacities by the factor ten, plus adding a third such collection, would increase the proportion of known species to just 20 % (if we assume 1.5 million) or even less than 10 % (if we assume 3.8 million). Clearly, the bulk of Fungi detected through environmental sequencing cannot be formally named if not also based on sequence data. Whatever reservations there may be against this approach, it seems impossible to conceive a practical alternative. Leaving this diversity unnamed and unclassified is not an option, as it would continue to be an enormous and increasing impediment to communication and research in the field.

In order to address this problem, a proposal had been put forward to modify the Code to allow sequences as types (Hawksworth et al. 2016). This proposal was not supported by the Nomenclature Committee for Fungi (Turland & Wiersema 2017) and was rejected by the Nomenclature Section of the International Botanical Congress in Shenzhen 2017. However, the Congress established a Special Committee to examine the matter for all groups of organisms which is due to report to the next Congress in 2023 (Hawksworth et al. 2017). Some authors have nevertheless already described new species based on an environmental sample type, such as Piromyces cryptodigmaticus (Fliegerová et al. in Körk 2012), or with a sequence type, such as Hawksworthiomycetes sequentia (De Beer et al. 2016), a currently invalid name established in anticipation of changes to the Code. A potential loophole for the formal description of voucherless, ecologically cryptic microfungi based on sequence data was posited by invoking the ‘illustration clause’ in Art. 40.5 (De Beer et al. 2016, Lücking & Moncada 2017, Turland & Wiersema 2017). However, this led to a suggestion during the Nomenclature Section meeting to redefine what constitutes an illustration allowed as type; an example was inserted into the Code to close this potential loop-hole (Turland et al. 2018: Art.40.5 Ex. 5), making clear that a representation of a sequence was not to be interpreted as an illustration for the purposes of typification; this option cannot now be the subject of a proposal until the next International Botanical Congress in 2023.

In order to stimulate discussion of this issue prior to the XIth International Mycological Congress (IMC11) in Puerto Rico in July 2018, to which a similar proposal to allow sequences to serve as types of fungal names has been submitted (Hawksworth et al. 2018), we elaborate here on the promises and pitfalls of formal, sequence-based, voucherless nomenclature. We offer solutions to problems at hand that could lead to specific provisions being made in the Code at IMC11 to allow formal, sequence-based nomenclature for voucherless fungi.

THE TYPE CONCEPT IN SEQUENCE-BASED NOMENCLATURE

The purpose of a type is to fix the application of a name. Presently, this has to be a physical specimen (including a permanently preserved, metabolically inactive culture in the case of Fungi), or if none can be preserved, in certain circumstances an illustration (Lücking & Moncada 2017, Turland et al. 2018). Apart from linking a name to a specimen, a name-bearing type has the following functions:

- Depiction of the phenotype, including morphological, anatomical, chemical, and physiological characters.
- Long-term (ideally perpetual) conservation of the original material.
Description of sequenced-based voucherless *Fungi*

Table 1. Possible alternative types for sequence-based nomenclature and their advantages and disadvantages. Desirability refers to a purely scientific viewpoint, without considering the actual necessity; risk refers to the possibility of undesired outcomes, such as artifactual taxa or newly described synonyms; feasibility refers to the efforts required to designate and store a type; and effectivity refers to the proportion of effort versus gain in closing the gap of undescribed fungal diversity.

<table>
<thead>
<tr>
<th>Specimen</th>
<th>Culture</th>
<th>DNA extract</th>
<th>Environmental sample</th>
<th>FISH</th>
<th>Illustration</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Comparability</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>limited</td>
<td>limited</td>
</tr>
<tr>
<td>New characters</td>
<td>yes</td>
<td>yes</td>
<td>limited</td>
<td>limited</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Quality control</td>
<td>yes</td>
<td>yes</td>
<td>limited</td>
<td>limited</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Artifactual taxa</td>
<td>rare</td>
<td>impossible</td>
<td>possible</td>
<td>possible</td>
<td>impossible</td>
<td>possible</td>
</tr>
<tr>
<td>Technical requirements</td>
<td>low</td>
<td>high</td>
<td>medium</td>
<td>medium</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Storage requirements</td>
<td>medium</td>
<td>high</td>
<td>medium</td>
<td>high</td>
<td>medium</td>
<td>low</td>
</tr>
<tr>
<td>Long-term preservation</td>
<td>limited</td>
<td>limited</td>
<td>limited</td>
<td>limited</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Capture of diversity</td>
<td>low</td>
<td>low</td>
<td>medium</td>
<td>medium</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>Code compliance</td>
<td>yes</td>
<td>yes</td>
<td>limited</td>
<td>limited</td>
<td>yes</td>
<td>limited</td>
</tr>
<tr>
<td>Changes to the Code</td>
<td>none</td>
<td>none</td>
<td>“mixtotype”?</td>
<td>“mixtotype”?</td>
<td>none</td>
<td>clarify “sequence type”</td>
</tr>
<tr>
<td>Overall desirability</td>
<td>high</td>
<td>very high</td>
<td>medium</td>
<td>low</td>
<td>high</td>
<td>very low</td>
</tr>
<tr>
<td>Overall risk</td>
<td>low</td>
<td>very low</td>
<td>medium</td>
<td>high</td>
<td>very low</td>
<td>high</td>
</tr>
<tr>
<td>Overall feasibility</td>
<td>very low</td>
<td>low</td>
<td>high</td>
<td>medium</td>
<td>very high</td>
<td>very high</td>
</tr>
<tr>
<td>Overall effectivity</td>
<td>very low</td>
<td>low</td>
<td>medium</td>
<td>high</td>
<td>very high</td>
<td>very high</td>
</tr>
</tbody>
</table>

- Reassessment of characters whenever necessary.
- Comparison with other specimens to establish their identity.
- Assessment of additional and new characters, including through new technology.

Seven possible types could be conceived to formally describe new *Fungi* from environmental sequencing data (Table 1). These can be divided into four groups: (1) physical type specimens (dried specimen, metabolically inactive culture); (2) undefined mixed samples (DNA extract, environmental sample); (3) a novel physical type derived via FISH technology (Spribille et al. 2016); and (4) the sequence data itself. Physical type specimens fulfil all five principal functions of a type, are Code compliant, and score high in terms of quality control and assessment of phenotype and novel characters (Table 1). However, to obtain physical types from taxa detected through environmental sequencing is not feasible at a large scale and this approach would defy the concept, since ultimately the type sequence would be obtained by resequencing the specimen and not from the original environmental sequence data. Thus, by default, sequence-based nomenclature cannot operate with traditional physical types, which in effect leaves only the five options in categories two to four above.

Some workers proposed using the environmental sample from which sequences were obtained as type material, in what could be referred to as a ‘bag’ type (Kirk 2012, Hawksworth et al. 2011, Hibbett & Taylor 2013, Minnis 2015, De Beer et al. 2016). This complies with the Code in having a physical type and hence fulfils a formal requisite for valid description: according to Art. 40.2, “… indication of the type as required by Art. 40.1 can be achieved by reference to an entire gathering, or a part thereof, even if it consists of two or more specimens as defined in Art. 8 …” (Turland et al. 2018). Although valid, for practical purposes this is not feasible, for three reasons: (1) the precise specimen to which a sequence belongs cannot be located within the sample, except for techniques such as fluorescent *in situ* hybridization (FISH; e.g. Spribille et al. 2016); (2) it is uncertain whether a fungus detected in the portion of the sample used up in the study is actually present in the remaining sample (De Beer et al. 2016); and (3) samples would have to be stored in long-term preservation in a frozen state to allow for further access of DNA material, to render the type material actually useful. In any case, such a type would be ambiguous, which would require subsequent lectotypification, generating the very problems outlined above, in that a precise lectotype cannot be designated.

Another option is to designate the DNA extract from which a type sequence originated as type. While permanent storage of a DNA extract is more feasible than the corresponding environmental sample, and the DNA that produced the sequence is likely to be contained in the remaining extract, a DNA extract type has the same problems as a ‘bag’ type, in that the precise piece of DNA corresponding to a particular taxon cannot be located within the extract. In addition, it might be argued whether a DNA extract type is still in compliance with the Code since, contrary to a ‘bag’ type, it does not contain an actual fungus. A type based on fluorescent *in situ* hybridization (FISH), a technique for instance performed in Cryptomycota and Cyphobasidium (Jones et al. 2011a, Spribille et al. 2016), would appear to be an ideal compromise between the extremes of a physical type and a sequence type. This technique would use the type sequence of a clade recognized as new taxon to precisely locate and visualize the corresponding physical structures (cells or hyphae) in the underlying sample, which could then be photographically documented and stored as a
permanent slide (similar to a metabolically inactive culture). The immediate advantage of this approach would be the implicit cross-check of the sequence data, since only real sequences would lead to a positive result. On the other hand, it would be difficult to validate this approach \textit{a posteriori} unless the fluorescent effect is permanent in the type slide. In case of a ‘bag’ type, a DNA extract type, or a FISH type, valid description of new, sequence-based taxa would only be possible with simultaneous access to the original material and its subsequent deposition in an institutional collection where it would be permanently accessible to researchers.

A voucherless sequence type fails on depicting a phenotype and assessment of additional and new characters (by default, all of its characters have been assessed through initial analysis), but fulfills the other three criteria better than a physical type. Whereas a physical type degrades over time, a sequence type can be stored as a digital file in perpetuity without quality loss. Digital data may be subject to technical failure and cyberattacks, but this applies to any electronic data and is not specific to the issue at hand. There is an equal likelihood of damage to physical type specimens, e.g. through pests, mould, humidity, water damage, and fire (Metsger 1999), as evidenced by the loss of most of the Berlin (B) collections in World War II (Hiepko 1987). Since a digital type can be stored in multiple identical copies, the risk of a complete loss of the information is much lower than for any physical type material. Effectively, a digital sequence type is an ‘exsiccate’ with an unlimited number of copies. Type sequences are universally accessible and their restudy is not destructive. In contrast, physical types need to be located, borrowed, or require a visit to study them. In addition, restudy is destructive, and so reduces their value as a reference point over time, for example as sporing structures are removed and samples taken for thin-layer-chromatography. Study of physical types is also dependent on methodology, can be open to interpretation and can lead to ambiguous results. In contrast, a sequence corresponds to a defined set of features (four possible states per character expressed by the universal IUPAC letter code), and the characters and their features underlie specific rules for their assessment, for example by checking against an original trace file. Two workers assessing ascospore dimensions in the same type may obtain different results; in sequence data, at a given position, an ‘A’ is an ‘A’.

Box 2. Environmental samples are real
There is a misconception that environmental sequences are not real but consensus sequences. This is not correct. Environmental reads deposited in e.g. the SRA, with an individual accession number, are as real as Sanger contigs deposited in GenBank. OTUs in analyses of environmental studies that are derived as consensus sequences from clustering are not real sequences and these are not acceptable as type sequences. A type sequence is always a real sequence and ideally it is the best sequence representing the lineage (or cluster). This is comparable to physical types: when a new species is described based on more than one collection, the description is a consensus of all specimens, but the type cannot be a consensus and instead must be a single collection or specimen, usually the one that is best developed. Another argument is that sequences do not correspond to real organisms but are trace manifestations. While this is true, there is another example of formal nomenclature of \textit{Fungi} in which the types are trace manifestations. With some exceptions, fossils are traces of organisms obtained through impression or crystallization, and their interpretation is often speculative. Yet, their nomenclature is fully integrated into the \textit{Code}, with some specific provisions.

An important argument regarding nomenclatural types is repeated, unlimited and free access. For physical types this criterion does not apply, although the problem is in part remedied by the availability of digital type images, e.g. through the \textit{Global Plants Initiative} on JSTOR (http://gpi.myspecies.info; https://plants.jstor.org). In contrast, type sequences can be accessed and compared to other sequence data in unlimited ways and in reproducible fashion, using quantitative methods such as automated alignment, assessment of alignment ambiguity, phylogenetic analysis, and species recognition methods. As a consequence, while the ideal situation is to have physical types plus sequence data in order to apply a consolidated species concept (Quaedvlieg \textit{et al.} 2014), type sequences, while not displaying phenotype features or harboring potential new characters, are superior to physical types in three of the five criteria listed above.

**POTENTIAL PITFALLS OF DNA SEQUENCE TYPES**

**Sequence errors**
Physical types may have flaws. A type does not usually encompass the phenotypic variation of a species. It need not be “typical”, since a species usually becomes much better known after its original description. It might not exhibit all characters that define the species, especially if the taxon occurs in various sexual, asexual and vegetative morphs, or it might be a mix of more than one taxon or else aberrant or an monstrosity. Sequence data may have errors analogous to those of physical types.

One of the most serious problems is chimeras, occurring both in Sanger and NGS techniques, as well as base flow (homopolymer) errors, most typical of Roche 454 and Ion Torrent platforms, and tag switching (Carlson \textit{et al.} 2012, Luo \textit{et al.} 2012, Yergeau \textit{et al.} 2012, Salipante \textit{et al.} 2014, Goodwin \textit{et al.} 2016). Chimeras arise from template DNA representing more than one taxon during PCR (Haas \textit{et al.} 2011). A mixed template in a Sanger PCR will cause double
or multiple peaks at a given position. The trace files are easily recognized and dismissed. In a rare constellation, a primer pair may have differential affinity to one template or another, generating clean trace files of different taxa for each primer. Sequence assembly will result in numerous ambiguous base calls except for conserved regions, making such chimeras again easily detectable. Very unlikely, but not impossible, is the above case but with reduced read length due to particular cycle conditions. For instance, in the case of ITS, the forward primer would sequence the ITS1 region and the reverse primer the ITS2. Through the conserved 5.8S region, such reads would be assembled into chimeras without immediate detection, but they can be identified in a phylogenetic context: since they have unique, artificial sequence patterns and are pulled towards two separate lineages simultaneously, they will appear on long branches with low basal node support. Dividing a data set into ITS1 and ITS2 and analyzing these separately (e.g. Blaalid et al. 2013), with a subsequent test for topological conflict, is thus a straightforward strategy to detect chimeras.

NGS chimeras arise from mixed DNA templates in amplicon PCR, when the amplicon from one template finishes prematurely and in the next cycle another template attaches (Haas et al. 2011). Since this is a stochastic process, the PCR product results in a mix of templates with regions of close correspondence and regions of disparate base calls. It is unlikely that such a mixed PCR product produces a sequence read passing through quality filters, since many positions will have subpar signal due to the presence of mixed bases in a given position between individual DNA fragments. In the unlikely event that the PCR combines two different templates at the same position, a true chimera corresponding to a high quality read similar to Sanger chimeras would result, with the difference that the joining point was not obvious; the only way to spot such chimeras would be to divide the read into variable portions starting from the centre and simultaneously blast both. The proportion of chimeras in NGS amplicon sequencing typically ranges between 8 % and 17 % for raw reads, and there are various tools for chimera detection and removal that reduce the proportion of chimeric reads to about 1 % (Huber et al. 2004, Ashelford et al. 2005, Edgar et al. 2011, Quince et al. 2011, Schloss et al. 2011, Porazinska et al. 2012, Kim et al. 2013, Mysara et al. 2015, Edgar 2016).

CARRY-FORWARD-INCOMPLETE-EXTENSION (CaFIE) errors are mostly generated in Roche 454 pyrosequencing and on Ion Torrent platforms, while apparently not occurring on Illumina platforms, although the latter has other sequencing errors (Minoche et al. 2011, Loman et al. 2012, Luo et al. 2012).

During sequencing, the extension in a given well follows a Poisson distribution, with most fragments fully, but a small portion only partially extended. Depending on the proportion of extended fragments, this leads to a suboptimal light signal which will be interpreted as either base not present or as a homopolymer of shorter length (Margulies et al. 2005, Huse et al. 2007, Gomez-Alvares et al. 2009, Kunin et al. 2010, Niu et al. 2010, Tedersoo et al. 2010, Balzer et al. 2011, Lücking et al. 2014). Incompletely extended homopolymer fragments become desynchronized and are completed during the next cycle of the corresponding flow base, causing a misplaced signal several bases after the homopolymer, not detectable as error but mimicking a genuine substitution. The only way to detect such errors is through alignment of reads relative to a broad reference alignment which will then place misplaced base calls in largely gapped columns (see below). CAFIE errors depends on the location of homopolymers and their length, with the consequence that phased indels can appear at the same position in independent reads; as a consequence, erroneous sequences are not necessarily singletons and they do not exhibit random patterns, which makes their automated detection close to impossible. It has been shown that such erroneous sequences can inflate taxonomic diversity computed through clustering techniques by several orders of magnitude, whereas multiple alignment-based methods are not susceptible to this problem (Lücking et al. 2014; see below)

There are various approaches available to detect, filter, and manage artificial sequences, so that the problem of inadvertently including artificial data in an analysis leading to recognition of artificial taxa can be reduced to a manageable proportion of less than 5 %. The most commonly used approaches exclude singletons and rare sequence reads or reads that cannot be mapped to reference taxa (Tedersoo et al. 2010, Nilsson et al. 2011, Caporaso et al. 2012, Edgar 2013). However, this will also exclude genuinely rare taxa (Lim et al. 2012), as was the case in an unnamed Alaskan soil fungus (Glass et al. 2014). Other tools to be tested to potentially detect aberrant and artificial sequences include assessing the secondary structure of ITS reads (Goertzen et al. 2003, Morrison 2009, Glass et al. 2013, Koetschan et al. 2014, Coleman 2015, Giudicelli et al. 2017). The likelihood of artificial sequences obtained from different studies being so similar that they form a well-defined and well-supported species-level clade is remote. Therefore, instead of simply excluding rare sequences or singletons from submitted biosample runs, the best approach to filter potential artifacts is to only allow formal description of novel species if the

Box 3. Physical vs. Sequence types

Opponents of sequence-based nomenclature argue that, in contrast to characters from physical types, environmental sequences cannot be reproduced. This is not correct. Environmental sequences come from samples and DNA extracts and if properly stored, sequences are theoretically reproducible. At the same time, physical types can have diagnostic characters that are not apparent in dried fungarium collections, such as basidiome cap colour, ascospore halos, conidial appendages which also cannot be resurrected. This is also so for metabolically inactive cultures designated as types from cultured Fungi where it can be impossible to recover some of the diagnostic features of a living culture from these types. The issue is therefore comparable to names based on physical types and cannot be upheld as an argument against sequence-based nomenclature.
sequences defining a species-level clade have been detected in a number of independent samples, with that number small enough to consider rare species and high enough to provide effective quality control. This number could be determined by a simple formula relating to the probability of sequence reads coming from N separate, independent studies, forming a supported clade, and at the same time being artificial; it can be shown that N ≥ 5 fulfills this requirement. This had been proposed as a recommendation (Hawksworth et al. 2016, 2018), but could be made a mandatory requirement for taxa based on NGS reads.

The following guidelines would help or substantially reduce the probability of describing artificial taxa based on faulty sequence data:

- **Dismiss sequences with a high proportion of ambiguous base calls (e.g. over 5%).**

- **Use filtering techniques for raw NGS sequence data to automatically detect and remove chimeras** (e.g. UCHIME, UCHIME2; Edgar et al. 2011; Edgar 2016).

- **Divide ITS data set into two portions, analyse them separately and test for topological conflict and rogue OTUs** (e.g. Blaalid et al. 2013).

- **DO NOT USE CLUSTERING TECHNIQUES!** Instead, apply multiple alignment techniques aligning reads to a reference alignment and check for gap-rich indel columns (e.g. PaPaRa, MAFFT ‘-add’; Berger & Stamatakis 2011, 2012, Katoh et al. 2017).

**Conceptual errors in sequence-based species delimitation**

Besides faulty sequences, sequence-based nomenclature is prone to conceptual errors that may lead to inaccurate recognition of taxa. The most commonly cited problems are homoplasy, intragenomic variation and gene duplication, and lack of resolution at species level (O’Donnell & Cigelnik 1997, O’Donnell et al. 1998, Hassanin et al. 1998, Källersjö et al. 1999, Inderbitzin et al. 2009, Druzhinina et al. 2010, Gazis et al. 2011, Kovács et al. 2011, Coissac et al. 2012, Kiss 2012). These problems are not unique to sequence-based taxa but apply to phylogenetic species recognition in general; therefore, the possibility of conceptual errors cannot be held exclusively against the idea of sequence-based nomenclature. The difference is that sequence-based taxa do not allow for independent, specimen-based check of sequence data; also, multi-locus approaches to detect problems with individual markers are not (yet) possible in sequence-based nomenclature (see below). However, the probability for such conceptual errors to occur is not higher in sequence-based than in specimen-based taxa. For the latter, a plethora of studies supports the value of molecular phylogeny for taxonomy, systematics and classification, in spite of the occasional shortcomings (Rossman 2007, Seifert 2008, 2009, Begerow et al. 2010, Schoch et al. 2012).

Some workers have argued that DNA homoplasy is more frequent than phenotype homoplasy (Baker et al. 1998; Wiens 2004). This is based on the misguided concept of total evidence, in which morphological homoplasy is potentially masked as incorrect phylogenetic signal. When mapping phenotype characters on trees derived from sequence data only, it can be shown that phenotype characters are often ecologically overformed and obscure true evolutionary relationships (Wake 1991, Hall 2003). An outstanding example is the molecular phylogeny of the *Fungi*, which has dramatically altered our understanding of fungal evolution and classification, from kingdom to species level. Yet, misunderstandings about sequence data persist. While the Nomenclature Committee for Fungi did not support the proposal on allowing sequences as types (Hawksworth et al. 2016), the Rapporteurs expressed their concern about a presumed "... lack of control as to the type sequence being an informative sequence. Many taxa could have the same sequence." (Turland & Wiersma 2017: 225). Homoplasy is of course present in DNA sequence data; for instance, in the third-codon position of protein-codon genes (Hassanin et al. 1998, Källersjö et al. 1999), there is a 25 % probability that the same base call arose by chance in two unrelated sequences. The same applies to ITS sequences, in which saturation effects, even if they cannot be directly measured due to ambiguous alignment positions, may occur in the highly variable ITS1 and ITS2 regions. However, contrary to phenotype data, in which characters are subjectively weighted, DNA-based phylogenies are based on simultaneous, unweighted assessment of all characters. For instance, in a 1000-bases long protein marker, there are about 300 third codon positions that may evolve freely and develop homoplasy due to saturation effects. For each individual position, the probability of homoplasy is 25 %, but the probability of two entire, unrelated sequences to evolve a similar pattern over 300 positions by chance is effectively zero. Thus, whatever the notion on DNA homoplasy may be: it is virtually impossible to imagine that two highly similar sequences of a given length evolved independently by chance. Sequence similarity is therefore always to be interpreted as indicating common descent, albeit possibly obscured by mechanisms such as hybridization or horizontal gene transfer.

There is, however, the problem of a lack of phylogenetic resolution due to homoplasy in recently and actively evolving lineages with incomplete lineage sorting (Will et al. 2005, Inderbitzin et al. 2009, Druzhinina et al. 2010, Gazis et al. 2011, Dupuis et al. 2012, O’Donnell et al. 2015). However, sequence-based nomenclature does not aim at resolving species complexes, it aims at naming novel lineages. It is therefore of no practical consequence if, in some cases, clades in voucherless taxonomy are erroneously defined as species when in reality they represent a species complex. As long as there are no associated physical specimens, there is no way of knowing, and this would only lead to an under-recognition of novel taxa. Long-branch attraction due to presumed DNA homoplasy (Bergsten 2005, Kück et al. 2012, Susko 2014) is a different issue that does not apply here. Taxa falsely clustering on long branches have both a long shared stem branch and long terminal branches, a pattern
Box 4. Species delimitation

Voucherless sequence-based nomenclature is not about accurately delimiting species in Fungi. The sole purpose of this approach is to convert countless unnamed into named sequences so that they can serve as proper references. This approach only works if done with a single marker, with the obvious choice being the fungal barcoding marker ITS. It does not matter whether ITS may not fully resolve species in particular groups or may over-resolve species as infraspecific entities in others. What matters is that a particular ITS sequence, when compared to a reference database, returns names, instead of designations such as “uncultured fungus”, “uncultured Ascomycota”, or “ectomycorrhizal fungal clone” etc. Whether such names are actual species, species complexes, or infraspecies, is irrelevant to the current debate; they will always point to the correct lineage at the given level of resolution and this will greatly facilitate correct classification of environmental or otherwise unknown sequences. Potential conflict arises when specimens or cultures become available for an initially voucherless lineage and multilocus studies may show that ITS in such lineages does not properly resolve species. However, such conflict can be handled: if ITS is too variable, with additional markers resulting in less species, formal synonymization takes place. If ITS lacks resolution and a different marker is a better barcode, then the initial name is epitypified with one of the specimens for which the additional marker is available and the name is attached to that particular clade. This process is the same as in nomenclature based on physical types.

different from species-level clades, in which the terminal branches leading to individual sequences are short; hence, long-branch attraction cannot lead to artificial species-level entities.

Intragenomic variation and gene duplication (paralogs, pseudogenes), as well as horizontal gene transfer, may be serious issues resulting in artificial topologies. Horizontal gene transfer has been demonstrated in Fungi (Schmitt & Lumbersch 2009, Soanes & Richards 2014), but is not expected to pose problems in species delimitation studies, particularly not with a multiple-copy marker such as ITS. Gene duplication has been found in various protein coding genes, such as β-tubulin, TEF1 (EF1α), and PKS genes (Schmitt et al. 2005, James et al. 2006, Aquileta et al. 2008, Hubka & Kolarik 2012), and using such markers may result in duplicate clades mimicking separate taxa. In contrast to protein-coding genes, rDNA occurs in multiple copies in large arrays in the genome and is presumed to maintain a consistent sequence pattern due to concerted evolution (Hurst & Smith 1998, Liao 1999, 2008, Ganley & Kobayashi 2007). Evidence for potential intragenomic ITS variation in Fungi is inconclusive; studies have demonstrated both presence and absence of such variation, using techniques such as RFLP, cloning, NGS amplicon sequencing, whole genome sequencing including HTS, and specifically designed primers (O’Donnell & Cigelnik 1997, O’Donnell et al. 1998, Ganley & Kobayashi 2007, Simon & Weiß 2008, James et al. 2009, Lindner & Banik 2011, Kovács et al. 2011, Kiss 2012, Lindner et al. 2013). Some methods are subject to the observer effect, in that variation is generated by methodical errors rather than being intrinsic (Keirle et al. 2011, Lücking et al. 2014, Mark et al. 2016). Lücking et al. (2014) reported that up to 99.3 % of indel variation in 454 pyrosequencing ITS reads from a single target taxon were due to sequencing errors, particularly homopolymer (CAFIE) errors, with genuine variation almost entirely ascribed to substitutions (Fig. 2). In the cloning approach by Simon & Weiß (2008), the proportion of variant base calls indicated in the supplemental figure [mbe-08-0468-File005_msn188.pdf] is 0.11 %, at reported TAG polymerase error levels (Chen et al. 1991), and the authors admit that in vitro TAG misreadings might cause such variation (Simon & Weiß 2008: 2251). Even if the rDNA cistron underlies strict concerted evolution, it can be expected that each generation of ITS copies after replication will have a natural level of variation corresponding to DNA polymerase misreadings in situ. In any case, such point mutations or single nucleotide polymorphisms (SNPs), whether real or methodological, would not result in artificial taxa when analysed in the context of multiple alignments, whereas clustering methods are highly sensitive to such variation (see below). Sanger sequencing usually gives a consistent signal corresponding to the dominant haplotype, which is supported by the numerous studies in which the ITS barcoding approach appears to work well (Rossman 2007, Begerow et al. 2010, Gazis et al. 2011, Schoch et al. 2012).

Pyrosequencing analyses demonstrated generally low intragenomic ITS variation in a broad set of fungal taxa (Lindner et al. 2013), and potential gene duplication involving the ITS has been reported from a few lineages only (O’Donnell & Cigelnik 1997, O’Donnell et al. 1998, Hughes & Petersen 2001, Ko & Jung 2002, Gomes et al. 2002, Smith et al. 2007a, Li et al. 2013). In most cases, this phenomenon is explained by past hybridization, and it appears to be highly constrained in the fungal genome (Wapinski et al. 2007) and hence would have minor impact on species delimitation approaches. In the study by Lindner & Banik (2011), considerable intragenomic ITS variation was reported for Laetiporus cincinnatus. Reanalysis of the original data (not shown) recovered these results which, however, suggest
hybridization as the cause: all “rogue” haplotypes cluster with strong support with other *Laetiporus* species and so cannot be the result of intragenomic evolution of new ITS variants. If intragenomic ITS variants are causes by hybridization, detection of such variants in vouchercless sequence data would not lead to artifactual taxa, since individual ITS clones would always belong to an existing species, even in a hybrid genome.

Many species-delimitation studies attempt to obtain fungal ITS barcoding and other markers from physical type specimens, indicating a community consensus that sequence data can properly place types within a phylogenetic framework, and hence allow for a proper application of the names attached to them. It is therefore not logical to argue that sequences as types would not work or would be inferior to physical types. DNA sequence data have already been used as sole diagnostic characters (Fliegerová et al. in Kirk 2012, Tripp & Lendemer 2012, 2014, Renner 2016, Lücking et al. 2016b). The argument that recovery and validation of a sequence from the material cannot be guaranteed is not relevant, as the same problem may exist with ephemerale phenotype characters of physical types, an issue not confined to fungi and seen, for example, in the highly diagnostic oil bodies in *Hepaticae* (von Konrat et al. 2012, He et al. 2013). There is also the ‘reverse epitype’ concept: currently, in a molecular framework, epitypes are designated based on specimens from which sequence data were obtained, to complement original physical type material. In analogy, when a fungus originally described from voucherless type sequences is eventually discovered as physical specimen, that material can be designated as an epitype to depict the phenotypic features of the fungus.

**PARALLEL CLASSIFICATIONS**

Apart from the possibility of formally establishing artifactual species based on erroneous sequence data or unrecognized conceptual pitfalls such as gene duplication, another major pitfall of sequence-based nomenclature is the establishment of parallel species-level classifications, either by describing new species that potentially have a name among the numerous unsequenced *Fungi* or by separately using different markers that cannot be traced back to a single taxon.

**Accidental de novo descriptions of the same species**

The number of fungal species has been estimated conservatively at 1.5 million, with other estimates ranging between 611 000 and up to 10 million (Hawksworth 1991, 2001, 2012, O’Brien et al. 2005, Schmit & Mueller 2007, Blackwell 2011, Mora et al. 2011, Hawksworth & Lücking 2017). With 120 000 species currently accepted, this means that in the best case scenario, over 500 000 species are still to be discovered; using the recently proposed range of 2.2 to 3.8 million, at least over 2 million await formal recognition. About 240 000 species-level names have been described in *Fungi*, apart from the 120 000 accepted species, another 120 000 considered synonyms or orphans1 (Hawksworth & Lücking 2017). Presently, about 35 000 species have sequence data available. Thus, if we assume a scenario of 120,000 accepted species, of which 35 000 have been sequenced, with a total of 1 million species existing and half of the presumably synonymous names or orphans not being conspecific with any of the 120 000 accepted species, a random set of environmental sequencing data would resolve as follows if a random representation of fungal diversity was assumed: 3.5 % of the sequences would cluster with accepted species and 96.5 % appearing novel; of the latter, 8.5 % would correspond to accepted, yet unsequenced species, 6 % to species with names potentially available but not currently in use, and 82 % to genuinely novel taxa, resulting in a probability of 14.5 % of newly describing species that already have names. This probability decreases assuming a higher total of fungal species (Table 2). If among the 240 000 existing names, in addition to the 120 000 accepted species, there are further 60 000 hidden in synonyms and orphans, the overall error rate for taxonomy based on physical types over the past 250 years is 33 %. Therefore appears that a projected, statistical rate of between 14.5 % and 1.5 % newly generated synonyms for sequence-based nomenclature would be a considerable improvement over specimen-based nomenclature.

There are ways to deal with this problem. Apart from unknown fungal lineages, environmental sequencing techniques frequently yield *Fungi* in well-known taxa, such as ectomycorrhizal species of, e.g. the genus *Russula* in soil samples or species of *Xylaria* in endophyte studies (Arnold et al. 2003, Davis et al. 2003, O’Brien et al. 2005, Geml et al. 2010). Since only a fraction of described species in such genera has been sequenced, sequence-based nomenclature would allow establishment of new, voucherless taxa that may already have a name. This is not the objective of sequence-based nomenclature, which should aim at formally classifying genuinely novel taxa and not interfering with other, integrative approaches to classify *Fungi*. For instance, the genera *Russula* and *Xylaria* contain around 750 and 300 accepted species, respectively (Kirk et al. 2008), with 2 673 and 791 species-level names described in each (Index Fungorum). Of these, 135 (*Russula*) and 17 (*Xylaria*) have been sequenced (GenBank), i.e. in both cases there are numerous described species that have not been sequenced, plus hundreds of synonyms that may correspond to yet unrecognized species. As a consequence, until all these names have been sorted out in a phylogenetic or taxonomic context (e.g. as synonyms in other genera), establishing new species based on sequence data only should be avoided. In contrast, *Archaeorhizomyces*, *Hawksworthiomyces* and *Lawreyomyces* are novel genera based on environmental sequencing or similar approaches and thus had no existing species names available prior to their description (Rosling et al. 2011, Menkis et al. 2014, De Beer et al. 2016, Lücking & Moncada 2017). The same applies to species of *Cyphobasidium detected* by Spribille 1

1*Orphan*: in taxonomy, a species name described in a genus to which it does not belong, and the placement of which has not been reassessed. Examples include species in genera such as *Agaricus*, *Lichen*, *Sphaeria*, and *Verrucaria* not congenic with the type species of those generic names.
et al. (2016), as only two species based on physical type specimens have been described in this genus (Millanes et al. 2016). Two complementary or alternative provisions could take care of this concern among the mycological community.

Parallel classification based on different markers

One of the central issues of sequence-based nomenclature is a community-wide agreement which markers to use. Current NGS technologies do not yet allow sequencing different markers from the same template or entire genomes, and maximum read lengths on Illumina MiSeq and HiSeq and Ion-Torrent PGM platforms do not exceed 300–600 bases (100–200 bases on HTS platforms), compared to the phased-out Roche 454 Titanium platform with up to 700 bases or 1500 bases and more on PacBio RS (Loman et al. 2012, Luo et al. 2012, Quail et al. 2012, Yergeau et al. 2012, Salipante et al. 2014, Goodwin et al. 2016). Sequence data corresponding to different markers, or fragments thereof, obtained from the same environmental sample cannot be concatenated to produce multilocus phylogenies since they cannot be traced back to particular individuals.

The ITS has been selected as the universal fungal barcoding marker (Schoch et al. 2012), inspite of some shortcomings, such as potential infragenomic variation and lack of resolution in evolving species complexes (see above). For instance, ITS data suggest the mushroom Schizophyllum commune represents a single species, whereas IGS indicates several, geographically separated lineages (James et al. 2001). Intron-rich protein-coding markers such as TEF1 have been shown to be superior to ITS in delimiting species in Fusarium (O’Donnell et al. 2015). Notably, while arguments against ITS include potential intragenomic variation, TEF1 has been shown to contain paralogs (James et al. 2006, Aguileta et al. 2008).

Some workers argue not to limit sequence-based nomenclature to a single marker and instead select the best possible marker in each instance (De Beer et al. 2016, Hibbett et al. 2016). Hawksworth et al. (2016) proposed as recommendation 8C.3: “DNA sequence data used for typification should be drawn from the molecular regions that are appropriate for delimiting species, based on prevailing best practices as determined by the relevant taxonomic communities.” This suggests the ITS barcoding locus as principal marker for the mycological community, but leaves the ultimate choice open to the specialists of a given taxonomic group. One could envision a scenario where ITS would be the default marker and more variable markers would be used in specific lineages. However, this could potentially lead to

<table>
<thead>
<tr>
<th>Predicted global species richness</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>B + C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proportion of accepted, sequenced species</td>
<td>Proportion of accepted, not sequenced species</td>
<td>Proportion of species hidden among synonyms</td>
<td>Proportion of genuinely novel taxa</td>
<td>Estimated error rate</td>
</tr>
<tr>
<td>712 000</td>
<td>4.9 %</td>
<td>11.9 %</td>
<td>8.4 %</td>
<td>74.7 %</td>
<td>20.4 %</td>
</tr>
<tr>
<td>1 000 000</td>
<td>3.5 %</td>
<td>8.5 %</td>
<td>6.0 %</td>
<td>82.0 %</td>
<td>14.5 %</td>
</tr>
<tr>
<td>1 500 000</td>
<td>2.3 %</td>
<td>5.7 %</td>
<td>4.0 %</td>
<td>88.0 %</td>
<td>9.7 %</td>
</tr>
<tr>
<td>2 000 000</td>
<td>1.8 %</td>
<td>4.3 %</td>
<td>3.0 %</td>
<td>91.0 %</td>
<td>7.3 %</td>
</tr>
<tr>
<td>2 200 000</td>
<td>1.6 %</td>
<td>3.9 %</td>
<td>2.7 %</td>
<td>91.8 %</td>
<td>6.6 %</td>
</tr>
<tr>
<td>3 000 000</td>
<td>1.2 %</td>
<td>2.8 %</td>
<td>2.0 %</td>
<td>94.0 %</td>
<td>4.8 %</td>
</tr>
<tr>
<td>3 800 000</td>
<td>0.9 %</td>
<td>2.2 %</td>
<td>1.6 %</td>
<td>95.3 %</td>
<td>3.8 %</td>
</tr>
<tr>
<td>5 000 000</td>
<td>0.7 %</td>
<td>1.7 %</td>
<td>1.2 %</td>
<td>96.4 %</td>
<td>2.9 %</td>
</tr>
<tr>
<td>10 000 000</td>
<td>0.4 %</td>
<td>0.9 %</td>
<td>0.6 %</td>
<td>98.2 %</td>
<td>1.5 %</td>
</tr>
</tbody>
</table>

Box 5. Introducing voucherless names in a known genus

For a new species based on a sequence type, without a physical voucher specimen, to be validly established in a genus where the type species are generally specimens: (A) All available names in the containing genus should ideally have been linked to a phylogenetically defined and named clade different from that with the new species (to be shown in the phylogenetic analysis accompanying the description), or must have been established as valid species or synonyms in other genera. (B) Names based on sequence types are not given priority over names based on physical types (specimen or culture), unless later epitypified with a matching specimen or culture.

There is also the option of names based on sequence types alone carrying a specific prefix (e.g. “DNA”, “SEQ”, “ENAS”, Candidatus) denoting absence of priority and that prefix would be removed upon epitypification by a physical type. This procedure would, however, be confusing to non-systematists familiar with binomial nomenclature, especially ecologists likely to want to refer to such taxa, and so we do not commend such an approach here.
irreconcilable, parallel classifications if, for instance, one study described new, broadly defined species of *Fusarium* using ITS, whereas another study found more narrowly defined species based on TEF1. In such a case, there would be no way of knowing which of the TEF1-based clades correspond to which of the ITS-based species, although this could be resolved by epitypification. Therefore, unless there is community-wide agreement that in particular taxa, another marker could be consistently used instead of, not in addition to, ITS, an approach using markers of choice is not feasible.

While resolution and accuracy of a barcoding marker is crucial to resolve species, this issue is less important in sequence-based nomenclature of vouchercless *Fungi*. First, there are no phenotype characters that could result in conflict with phylogenetically defined species. Second, resolving difficult species complexes is not the objective of this endeavour (see box above). With further advancements of NGS technologies (Koren et al. 2013), it might eventually be possible to generate more than one marker or entire genomes from a single template and the limitation to a single marker could be removed.

Even when using ITS as a single marker, the problem of parallel classifications goes further. The approximately 1 billion fungal ITS reads in the SRA have an average length of 353 bases, which mostly corresponds to either the ITS1 or the ITS2 region. As a consequence, reads that correspond only to the ITS1 or ITS2 region cannot be used in parallel to establish species-level clades. Instead, besides using complete ITS sequences from Sanger sequencing and newer NGS technologies, there would have to be an agreement with regard to short reads whether to use either ITS1 or ITS2 only (Bazzicalupo et al. 2013). Conceptually, this does not impose a limitation on resolution; ITS1 and ITS2 separately are mostly congruent with full-length ITS data (Blaalid et al. 2013), although an eukaryote-wide study suggests that ITS1 is generally superior to ITS2 as barcode marker, particularly in the Ascomycota (Wang et al. 2015). Again, as outlined above, this issue is not relevant to the purpose of sequence-based nomenclature.

**Simultaneous description of new species**

In a traditional context, the description of new taxa depends on access to material, including types for comparative studies, and taxonomic expertise. It is therefore uncommon that the same species is described simultaneously with the corresponding authors being unaware of each other’s work. In the case of describing new species based on environmental sequence data, such a situation is much more likely because there is universal, unrestricted and simultaneous worldwide access to data including type sequences (whereas a physical type can only be studied at one place at a given time) and the required expertise of phylogenetic analysis including species recognition methods is more widely dispersed and not limited to taxonomic experts of a group. Therefore, there is a greater possibility of different workers simultaneously studying the same data and describing the same taxon under different names. The principle of priority would take care of this as it does for names based on physical types, but it would be unfortunate to unnecessarily duplicate work.

There are several mechanisms that could be introduced to prevent this from happening or reduce the possibility:

- A network in which ongoing studies are announced and defined.
- Immediate release of type sequences of taxa described as new so that similar or identical sequences can be immediately detected.
- Free accessibility to registered new taxa in manuscript stage prior to publication.
- Peer review by experts that have an overview of the field.

This would, however, require changes in the procedures in the Code (Turland et al. 2018: Art. F.5) for the current mandatory system for the registration of names of new taxa in the approved repositories (Fungal Names. Index Fungorum, or MycoBank). It is not recommended that new taxa are registered prior to a paper being accepted for publication (Rec. F.5A.1), as changes are often made during the peer review process and there are many names in the repositories that have never been validly published. Further names are not released by the repositories until they have been effectively published.

**SPECIES DELIMITATION**

Environmental sequencing studies yield tens to hundreds of thousands of reads each. With 20,879 experiments (NGS runs) containing 1 222 062 203 fungal ITS reads currently in the SRA (see above), the average number of reads per sequencing run is 58,531. Analysing sequences from the SRA representing a particular clade of interest could potentially retrieve millions of reads. Such huge amounts of data can only be classified by fast methods such as blasting and clustering (Li & Godzik 2006, Schloss et al. 2009, Edgar 2010, 2013, Caporaso et al. 2010, Huang et al. 2010, Huse et al. 2010, Kumar et al. 2011, Nilsson et al. 2011). Unfortunately, clustering is inferior to alignment-based phylogenetic methods, resulting in overestimations of taxonomic diversity (Quince et al. 2009, Engelbrektson et al. 2010, Kunin et al. 2010, Porter & Golding 2011, Powell et al. 2011, Unterseher et al. 2011, Zhou et al. 2011). Estimates of global species richness based on such approaches may lead to exaggerated numbers. For instance, O’Brien et al. (2005) estimated the number of fungal species at 5.1 million (Blackwell 2011, Hawksworth 2012), and a recent study by Locey & Lennon (2016) predicted a trillion(!) species on Earth, many of these ecologically cryptic *Fungi* and other microorganisms detected through environmental sequencing.

While the problem of overestimating taxonomic diversity based on clustering is well-documented, clustering continues to be the method of choice for analysing large amounts of NGS data. Clustering works fast and capable of sorting large amounts of data based on pairwise alignment. In pairwise...
Description of sequenced-based voucherless *Fungi*

Alignment, sequencing errors such as CAFIE are interpreted as substitutions (Gazis et al. 2011, Lücking et al. 2014), unless the gap penalty is substantially lowered which, however, may lead to false interpretation of true substitutions as indels. Therefore, sequences of the same species containing errors are parsed out into different clusters, inflating taxonomic diversity (Fig. 3). This problem does not occur in multiple alignment-based phylogeny, since multiple alignments of closely related sequences place erroneous indels in gapped columns, where they have practically no effect on the resulting topology (Fig. 4, Lücking et al. 2014).

Another problem of clustering is the requirement of a fixed threshold value which, depending on the study, is usually set between 95 % and 99 % (O’Brien et al. 2005, Smith et al. 2007b, Morris et al. 2008, Ryberg et al. 2008, Walker et al. 2008). Such fixed thresholds do not exist in nature (Bruns et al. 2007, Nilsson et al. 2008, Hughes et al. 2009), since intraspecific and interspecific sequence divergence is a function of time, population size and geographic distribution. Fixed thresholds have also taken into consideration potential sequencing errors, at rates between 0.2 % and 1.5 %, in additive fashion, whereas in reality, effects of sequencing errors are augmented by their random positions relative to genuine substitutions and, due to the nature of pairwise alignment in clustering methods, can be multiplicative rather than additive (Lücking et al. 2014). Therefore, a fixed threshold cannot prevent sequencing errors to affect the outcome in a clustering approach.

As a consequence, description of new fungal species based on voucherless sequences must be based on approaches that employ rigorous, multiple alignment-based phylogenetic analysis and, in addition, should use quantitative species-delimitation methods such as GMYC or...
PTP (Fujisawa & Barraclough 2013, Zhang et al. 2013). An idealized protocol is outlined in Box 6.

**Box 6. A protocol for describing new voucherless species of fungi**

1. Assemble an initial data set of sequences of interest, e.g. through mining the SRA using a template sequence and retrieving all similar sequences setting a blast threshold; identify the sequences by study sample and read number.
2. Add related sequences from GenBank.
3. Digest the data set through clustering into smaller portions that can be handled by multiple alignment methods such as MAFFT, which currently allows alignments of up to 30,000 sequences but works best with 10,000 sequences or less.
4. Generate multiple alignments for data subsets using, e.g. Guidance and MAFFT.
5. Run phylogenetic analysis, e.g. maximum likelihood using RAxML.
6. Inspect resulting trees for potential anomalies, e.g. long branches with low support indicating potential chimeras; also, run ITS1 and ITS2 partitions separately and test for conflict.
7. Annotate and remove potentially problematic sequences.
8. Run final phylogenetic analysis, ideally both maximum likelihood and Bayesian.
9. Run various quantitative species delimitation methods, e.g. GMYC and PTP.
10. Identify clades consistently resolved as species that fulfil the requirement of the included sequences resulting from at least N independent study samples.
11. Combine all identified species-level clades into a single alignment and rerun (5) and (9).

**BACKBONE PHYLOGENY AND HIGHER CLASSIFICATION**

ITS is generally not fully alignable across a broader taxon set above species level. Therefore, employing the fungal barcoding marker as principal locus to delimit and formally describe new species of voucherless Fungi, without the possibility of using concatenated data sets with more conserved loci, may generate problems when attempting to establish higher-level phylogenies for these new taxa, particularly if they represent novel lineages at the genus, family, order or class level (Hibbett et al. 2016, Nilsson et al. 2016, Tedersoo et al. 2017). In addition, voucherless fungal classification makes it impossible to rank hierarchically structured clades based on phenotype features. However, there are options to deal with these shortcomings. For instance, Wang et al. (2011) successfully employed a simultaneous alignment and tree building approach to delimit genera and species in Geoglossomycetes based on (largely environmental) ITS data only.

ITS sequence reads can be placed within a broad, multilocus phylogenetic framework generated from known Fungi using the evolutionary placement algorithm (EPA) implemented in RAxML (Stamatakis et al. 2010, Berger et al. 2011, Zhang et al. 2013, Stamatakis 2014), an ideal tool for environmental sequencing studies (e.g. Sunagawa et al. 2013). While a stand-alone, full alignment of ITS sequences across a broad taxonomic range is challenging, one alternative is adding new ITS reads to a fixed, multi-locus alignment of reference taxa, as implemented in tools such as PPlacer (Matsen et al. 2010), ML TreeMap (Stark et al. 2010), PaPaRa (Berger & Stamatakis 2011), MAFFT (Katoh & Frith 2012), or T-BAS (Carbone et al. 2017). An initial fixed ITS alignment could be elaborated from reference taxa by means of a combined alignment and tree building method, such as BAi-Phy or SAtE (Suchard & Redelings 2006, Liu et al. 2009, 2012, Wang et al. 2011). Alternatively, ambiguously aligned regions can be recoded using PICS-Ord in a de-novo alignment, which as been shown to work effectively across broad taxon sets and large alignments of hundreds or thousands of sequences (Lücking et al. 2011).

A more reliable approach is de-novo alignment of the ITS across reference and query taxa using Guidance HoT scores for alignment confidence, which only retain columns aligned with high confidence (Penn et al. 2010a). Arguably, if used across an entire class or phylum, this approach would largely retain the conserved 5.8S region only, which is presumed to not contain sufficient resolution for a backbone phylogeny, but has been shown to work remarkably well in plants and Fungi (Hershkovitz & Lewis 1996). We tested this by analysing 210 ITS sequences of the genera Tremella (Tremellales), Auricularia (Auriculariales), Albatrellus, Peniophora, Russula (Russulales), Athelia (Atheliales), and Boletus, Coniophora, and Suillus (Boletales). The complete alignment for these taxa using MAFFT results in a length of 1354 columns, many of which are ambiguously aligned across the entire set. Running the sequences through the Guidance web server (Penn et al. 2010b) returns 407 columns aligned with a confidence of 95 % and higher, of which 174 columns represent a compact block present cross all taxa (Suppl. File S3). Analysing this alignment using RAxML (Stamatakis 2014), the resulting topology (Fig. 5) resolved the underlying
phylogeny remarkably well (except for the position of Auricularia), with the two orders Boletales and Russulales and most genera monophyletic except the collective genus Athelia (Rosenthal et al. 2017) and the genus Coniophora (resolved as paraphyletic grade), with moderate bootstrap support across genera (76 ± 17). Allowing a higher number of alignment columns by reducing the confidence limit to 70 % yields the same topology but strongly increases support across genera (95 ± 5). Thus, a much reduced ITS retaining only columns aligned with good to high confidence is not only capable of reconstructing the backbone phylogeny to a large extent but underlines the usefulness for the application of the EPA, with the added advantage that the entire process can be automated using a Guidance-MAFFT-RAxML pipeline.

There are several objective methods to hierarchically rank ITS backbone phylogenies in a consistent way. One approach is to “hijack” species delimitation methods such as GMYC, haplowebs, and PTP (Fujisawa & Barraclough 2013, Zhang et al. 2013, Dellicour & Flot 2015). Once a given set of sequences has been phylogenetically analysed and species-level clades have been identified, one sequence per species is retained. Applying the species delimitation method again will then denote higher level clades. Another approach is to run the Guidance HoT score analysis over a data set. The more closely related the included sequences, the lower the rank they represent as a whole, and the higher the number of columns that can be retained with confidence. Data of known taxa at various hierarchical levels can be used to establish correlations and thresholds. In the above example, aligning across Agaricomycotina (subphylum level) resulted in 30 % (407 of 1354) of all alignment columns retained at 95 % confidence, whereas for the genus Russula alone, 52 % (420 of 812) of the columns were retained. If these thresholds are consistent across taxa, an ITS data set of unidentified sequences retaining 30 % of columns with high confidence is likely to represent a class or subphylum, whereas 50 % point to a genus. Finally, temporal banding allows the definition of ranks based on divergence times obtained from an ultrametric or molecular clock tree, as recently suggested for Ascomycota, Sordariomycetes, Lecanoromycetes, and Parmeliaceae (Divakar et al. 2017, Hyde et al. 2017, Liu et al. 2017).

CONCLUSIONS

Voucherless, sequence-based nomenclature poses numerous challenges, but there appears to be no practicable alternative to formally naming the numerous novel fungal lineages now being detected in environmental sequencing studies. We showed that even if increased by an order of magnitude, specimen- and culture-based inventories will not be capable to formally classify a substantial portion of the predicted unknown fungal diversity within a reasonable time frame. The challenges of sequence-based nomenclature are manageable and there are numerous methods to classify voucherless Fungi, using a single marker such as the ITS, both at the species level and at higher taxonomic ranks. There have been arguments that voucherless, sequence-based nomenclature may threaten support to other branches of mycology, such as culture collections and their research, or on the contrary may favour large laboratories in North America or Europe and leave researchers in other countries behind. These arguments have no grounds, on
the contrary. Funding for fungal research is mostly based on the importance of Fungi for ecosystem services and their potential applications. These can only be studied based on specimens and cultures, but not based on voucherless, sequence-based taxa. Therefore, sequence-based nomenclature will not diminish funding to other branches of mycology, but can be expected to generate additional funding in areas of computational biology related to sequence read placement, an area that is already now one of the hot spots of phylogenetic tools. Also, sequence-based nomenclature does not require any laboratory equipment but is entirely computational and hence accessible to virtually anybody, since both data and software are freely available and servers allow access to computational clusters to perform large scale analyses. Therefore, if anything, mycologists in any area of the world have equal access to this approach. As a whole, voucherless, sequence-based nomenclature is not a threat to specimen-based mycology, but rather a complement to substantially speed up cataloguing global fungal diversity in those lineages that are rarely detected using specimen-based methods. If considered desirable, simple and straightforward provisions in the Code or a Code of Practice developed by a body such as the International Commission on the Taxonomy of Fungi (ICTF) can help avoid the description of artificial taxa or species for which names might already exist. Voucherless, sequence-based fungal taxonomy is universally accessible but is by no means “fast track” mycology, as this approach requires extremely careful work and high skill-levels comparable to those of specimen-based mycologists. However, control mechanisms and effective peer-review by the mycological community are crucial for a successful implementation of this approach, as in all other areas of research.

The time is right for the mycological community as a whole to consider and answer the following questions:

- Do we recognize the potential of environmental sequences as a substantial source of fungal diversity information that cannot be addressed similarly by other means?

- If we recognize that potential, do we want to allow formal nomenclature to be based on types other than those currently allowed by the Code (i.e. dried specimens, microscopic preparations, illustrations, metabolically inactive cultures), to capture this diversity?

- If we agree to adjust formal nomenclature, what alternative types would be allowable (the underlying environmental sample or ‘bag type’; the underlying DNA extract or ‘DNA type’; a graphic illustration of the type sequence; or the sequence itself or ‘sequence type’)?

- If we permit alternative types, what if any limitations on the formal establishment of sequence-based taxa do we want to hard-wire into the Code and what limitations do we want to trust to peer-review and scientific integrity?

Most importantly, we should all recognize that established practices need to change to facilitate our science and should not be a hindrance to its progress. Mycologists have an enviable record amongst nomenclaturalists in showing willingness to adopt new ways of working, after due debate. Examples include the acceptability of metabolically inactive, permanently preserved cultures as name-bearing types, adoption of a single starting point date for the naming of fungi, the requirement to register new scientific names for them to be valid, the ability to propose lists of names for protection, and ending the separate naming of morphs of the same species. All these changes followed much debate at mycological meetings and exchanges in the literature, and in the end consensus was achieved and the rules that govern these changed. In some of these cases this process took many decades, and in the interim some authors chose to ignore the rules then in force leading to conflicting treatments. This is already starting to happen in the area of voucherless types, and we feel that the community needs to agree on an acceptable solution as a matter of urgency, as with advancing technology environmental sequencing is now accelerating exponentially.

ACKNOWLEDGEMENTS

We are indebted to Paul M. Kirk and Conrad L. Schoch for providing data from Index Fungorum and GenBank, respectively. Travis Adkins kindly supplied fungal culture numbers for the NRRL. Daniel Lindner is thanked for providing raw data from the study on intragenomic ITS variation in Laetiporus sulphureus. The members of the International Commission on the Taxonomy of Fungi (ICTF), including Catherine Aime, Takayuki Aoki, Lei Cai, Pedro Crous, Wilhelm DeBeer, David Geiser, Peter Johnston, Tom May, Andrew Miller, Conrad Schoch, Marco Thines, Keith Seifert, Marc Stadler, and Ning Zhang, engaged in fruitful discussions on the subject (referred to in the text or the boxes). Conrad Schoch and two anonymous reviewers provided valuable comments that helped to improve this manuscript.

REFERENCES


Dupuis JR, Roe AD, Sperling FA (2012) Multi-locus species
delimitation in closely related animals and Fungi: one marker is not enough. Molecular Ecology 21: 4422–4436.
Hawksworth DL (2017) DNA sequences as types: a potential loophole in the rules discovered. IMA Fungus 8: (4).


NOTE: the Supplementary files S1–S3 can be found on the IMA Fungus website, www.imafungus.org.