Ten reasons why a sequence-based nomenclature is not useful for fungi anytime soon

Marco Thines1,2, Pedro W. Crous3, M. Catherine Aime4, Takayuki Aoki5, Lei Cai6, Kevin D. Hyde7, Andrew N. Miller8, Ning Zhang9, Marc Stadler10

1Goethe University, Department of Biological Sciences, Institute of Ecology, Evolution and Diversity, Max-von-Lauve-Str. 13, D-60483 Frankfurt am Main, Germany; corresponding author e-mail: m.thines@thines-lab.eu
2Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany
3Westerdijk Fungal Biodiversity Institute, Uppsalaalaan 8, 3584 CT Utrecht, The Netherlands
4Department of Botany and Plant Pathology, Purdue University, 915 W. State Street, West Lafayette, IN 47907, USA
5Genetic Resources Center, National Agriculture and Food Research Organization (NARO), 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan
6State Key Laboratory of Mycology, Institute of Microbiology, Chinese Academy of Sciences, NO.1 Beichen West Road, Chaoyang District, Beijing 100101, China
7Center of Excellence in Fungal Research, Mae Fah Luang University, Chiang Rai 57100, Thailand
8Illinois Natural History Survey, University of Illinois, 1816 South Oak Street, Champaign, IL 61820, USA
9Department of Plant Biology, Rutgers University, 59 Dudley Road, Foran Hall 201, New Brunswick, New Jersey 08901, USA
10Department of Microbial Drugs, Helmholtz-Zentrum für Infektionsforschung, Inhoffenstrasse 7, D-38124 Braunschweig, Germany

Abstract: The large number of species still to be discovered in fungi, together with an exponentially growing number of environmental sequences that cannot be linked to known taxa, has fuelled the idea that it might be necessary to formally name fungi on the basis of sequence data only. Here we object to this idea due to several shortcomings of the approach, ranging from concerns regarding reproducibility and the violation of general scientific principles to ethical issues. We come to the conclusion that sequence-based nomenclature is potentially harmful for mycology as a discipline. Additionally, a classification based on sequences as types is not within reach anytime soon, because there is a lack of consensus regarding common standards due to the fast pace at which sequencing technologies develop.

Key words: Biodiversity nomenclature systematics voucherless taxa

INTRODUCTION

Fungi are highly diverse, with an estimated number of up to 6 Million species (Taylor et al. 2014), of which far less than 5 % have been described to date. While nothing is known about most of these missing species, environmental sequencing studies have revealed many sequences that currently cannot be associated with any described species. As a consequence, there has been the temptation to describe these species on grounds of only their sequences, which have been proposed to serve as substitutes for type specimens of new taxa (Hawksworth et al. 2016). In this manuscript we outline ten reasons why we feel that a nomenclature based on sequences is not useful and applicable to fungi anytime soon, emphasizing potential pitfalls and unintended detrimental effects of such an approach.

TEN REASONS

1. The resolution of barcoding loci, especially ITS, varies among different groups

The idea of using sequence similarity as a measure of defining taxa is tempting, and due to the lack of other readily available characteristics, bacteriologists have embraced this concept for the delimitation of bacterial taxa (Stackebrandt et al. 2002), although, importantly, there are several additional requirements needed for formally naming bacterial taxa according to the latest version of the International Code of Nomenclature of Prokaryotes (Parker et al. 2015). To be useful in the discrimination of species throughout all the diversity of a given organismic group, sequence divergence in DNA barcoding loci needs to be strongly correlated to the genetic diversity that is needed to provide effective barriers for gene-flow. However, while this does not even hold up for bacteria (Fraser et al. 2009), it certainly does not for fungi. The universal barcode for fungi are the internal transcribed spacers (ITS), regulatory, non-structural RNA transcripts with a common core of secondary structure (Schultz et al. 2005, Schoch et al. 2012). The ITS regions are rather
conserved in many species groups, in particular within the Sordariomycetes and other classes of Ascomycota (Stadler et al. 2014b). However, they may vary strongly in other groups, such as some groups of downy mildews (Thines 2007), rust fungi (Aime et al. 2017) and the Fusarium fujikuroi complex, in which species have two divergent ITS2 types (O’Donnell & Cigelnik 1997). This can lead to two potential types of error. As exemplified by the genus Dalbina (Stadler et al. 2014b), entire species groups that are very different in terms of ecology, morphology, and biochemistry but share very similar ITS sequences would be lumped together into a single species if a unique ITS sequence were already acceptable as a type. Conversely, in other species there are little constraints to variation in some loop regions of the ITS, leading to different sequence types that could be erroneously interpreted as separate species.

2. There is a high risk of introducing artefacts as new species

Most complete ITS sequences are still produced by conventional dideoxy sequencing (“Sanger” sequencing), but given the routine nature of barcoding fungi, little effort is often put into quality control of the sequences by visual editing or by sequencing the complementary strand (Janda & Abbott 2007), as evidenced by an increase of variant bases in sequences deposited in public databases towards either end of the sequences, where low quality base-calls are usually present. As some variants are more likely to occur, e.g. homopolymer errors or wrong base-calls after GC-rich stretches, these might look like actual sequence types. In addition, most widely used polymerases, such as the Taq DNA polymerase have a high rate of incorporating wrong nucleotides, which is usually no problem in direct sequencing, but problematic when sequencing clones or when using high throughput sequencing, which exposes these errors (Oliver et al. 2015).

As the vast majority of previously uncharacterised species is prone to be sequenced in screens from cultures derived from environmental samples (Glynou et al. 2016). These usually do not focus on taxonomy, but just on a very rough classification, and consequently, quality control is not necessarily focussed on the DNA sequences generated due to the large amount of data. In addition, PCR can produce chimeric sequences (Hughes et al. 1999), especially when DNA derived from multiple species is used (e.g. environmental DNA), by the attachment of non-homologous, incompletely synthesised PCR products to each other. There are approaches to detect such chimeras (Edgar et al. 2011, Nilsson et al. 2015), but especially when sequence divergence is only moderate, filtering is difficult. When high-throughput sequencing is used for barcoding, additional problems arise, e.g. additional chimera formation during bridge-PCR in Illumina sequencing (Coissac et al. 2012, Schnell et al. 2015). The situation is further complicated by the potential presence of multiple divergent ITS copies within genomes of one species, or, as shown by Peršoh et al. (2009) and Stadler et al. (2014a), even among single spore isolates from the same perithecium in a heterokaryotic setting. Such variations may be due to degenerate copies (Wor et al. 2015, Harpke & Peterson 2008), failure to converge into a single canonical ITS version (Li et al. 2013), potentially with multiple polymorphic positions, or the maintenance of multiple rDNA cistrons (Ko & Jung 2002, Wörheide et al. 2004, Lindner & Banik 2010, Harrington et al. 2014, Kijpornyongpan & Aime 2016). All of these issues are prone to produce artefact “shadow taxa” if a barcode sequence were sufficient to serve as the type for a species.

3. There is no consensus regarding the data type or amount needed for species delimitation

As least some of the issues mentioned thus far, especially those pertaining to ITS sequences, could probably be addressed by using high quality sequences of additional loci, but currently there is little consensus regarding how much sequence data are needed to reliably identify and delimit a species or a corresponding OTU and how it should be treated (Creer et al. 2016, Hibbett 2016a, b, Köljalg et al. 2016). While sometimes even fractions of ITS1 or ITS2 might be sufficient for resolution at the species level (Miller et al. 2016), often it will be necessary to sequence multiple loci for proper species delimitation (Stadler et al. 2014a, Choi et al. 2015). While multigene genotyping has become standard in some groups (Choi et al. 2015, Choi & Thines 2015, Kruse et al. 2017a, Wendt et al. 2018), others still rely mostly on ITS because of difficulties in generating primers for other loci (Kruse et al. 2017b). Also the kind of loci that can be amplified by universal primers differs largely among groups. While in some, actin might amplify well (Voigt & Wöstemeyer 2000), others only work for some ribosome-associated proteins (Matheny et al. 2002, Stielow et al. 2015). This makes a consensus with respect to which loci to use difficult. Even a recommendation with respect to how many nucleotides should be sequenced cannot be made, as mutation rates differ between loci and organism groups. If it were argued that a single nucleotide difference would be enough to delimit species, there would be a high risk of introducing artificial shadow taxa on the basis of artefacts. However, in order to find ten or 20 different nucleotides, thousands of nucleotides would need to be sequenced in some groups (Choi et al. 2015). But, as this would most likely require a specimen, this would challenge the whole idea of sequence-based types, as species based on these are meant to be introduced in the absence of a specimen (Hawksworth et al. 2016). As genomes are becoming more widespread, they might even become commonplace when new taxa are introduced in the future, similar to recommendations in bacteriology (Rosello-Móra & Amann 2014). However, due to the repeat nature of the ribosomal DNA cistrons, the regions currently used for barcoding are often not well-assembled or even masked during repeat masking steps so that they are seemingly absent from annotated genomes.

In addition, there is also no consensus regarding the type of sequence data that should be acceptable. It could be argued that high quality short fragments of a few hundred base-pairs are sufficient, e.g. such as those produced by current short-read sequencers, but also long reads from single molecule sequencing could be seen as acceptable if they contain enough high quality base-calls, despite intermittent low quality stretches. In addition, there are also several derived sequence data types (assembled or clustered reads), which have their own complexities (see point 7) but are seen as acceptable sequence data for species discovery and naming by some authors (Hawksworth et al. 2016, Jagielski et al. 2016).
4. Voucherless data are not reproducible
Reproducibility and testability are essential in science (Popper 1968, Cassey & Blackburn 2006). The value of a physical specimen, which is a requirement for valid publication of preserving fungi and organisms treated as such since 2007, is that it can be assessed by other researchers for testing the species hypothesis (Bradley et al. 2014). In other words, a voucher specimen serves as the embodiment of a species hypothesis, and contains a suite of characters that can be tested, evaluated, and reinterpreted by future researchers, including characters (such as DNA sequences themselves) that may not have been recognized at the time of typification, yet may become crucial in future taxonomic evaluations. An important concern with respect to sequence-only types is that they are not reproducible and it would be impossible to generate additional data for other characters or loci. However, this might be needed if there are competing species hypotheses or it would be later determined that the deposited sequence is insufficient to allow differentiation in a species complex. All of these concerns can only be addressed if a vouchered specimen is deposited. If such a specimen is present, the designation of sequence data as type becomes obsolete. It could be proposed that in the case of sequence-based species hypotheses from environmental sequencing a preparation of the environmental sample could serve as a specimen. However, such specimens would still not guarantee reproducibility as: (1) the organism from which the actual sequence was derived might not be in the preparation; (2) the sequence might still be an artefact (see also Points 3 and 7); (3) the sequence might have been derived from free environmental DNA so that no identifiable parts of the organisms are within the sample; and (4) it has been shown that there is often no full overlap between two independent assessments of the same sample, and that sequence composition strongly varies with the PCR annealing temperature used (Schmidt et al. 2013).

5. Sequence-based types cannot be verified
As discussed in Point 4, any scientific hypothesis needs to be testable (Popper 1968). In order to be testable, the information related to the hypothesis needs to be verifiable. However, voucherless sequence-based types cannot be verified or reproduced – they have to be taken as absolutes. This also means that the species hypothesis they support cannot be tested, rendering systematic mycology a pseudo-science. Testability of taxonomic hypotheses due to the possibility to assess physical type specimens has been one of the greatest advances in systematic biology, which has led to an increase in nomenclatural stability, has facilitated communication, and allowed the reassessment of concepts when new technologies became available (Singh et al. 2015). Allowing the requirement that taxonomic hypotheses for preservable species need to be backed up by a physical type to be abandoned would be a giant step backward.

6. Sequence-based types are not relatable
Related to Points 4 and 5, characters of specimens listed in diagnoses or descriptions can always be related to the specimens from which they have been derived. They do not stand isolated, but rather are a proxy for the description of the entire set of characters of the taxonomic hypothesis they relate to. Sequence data are just one of many characters of a species, even though they might be a good starting point for in-depth investigations (Kekkonen and Hebert 2014). If they alone were eligible as types, they would stand alone in the way a specimen would. But in contrast, no additional characters could be assessed and sequence data do not relate to any real-world object. Furthermore, species typified by a sequence can only be compared to other species sequenced at the same locus. They would no longer be comparable to species typified by single sequences at other loci, greatly limiting their taxonomic utility and again creating the potential for shadow taxa. Presently about 120 000 species are acknowledged, but there are more than 400 000 names (Dayarathne et al. 2016). Only a mere fraction of the 120 000 accepted species have DNA sequences deposited. If species were named based on environmental sequences, and they were given the same status as species with specimens, the risk would arise that all work done before the first DNA sequences were deposited in GenBank, in 1991, would be deliberately ignored. Thus, sequence-based naming of species is prone to prohibit careful research relating DNA data to existing names, and erecting numerous new and superfluous names that actually belong to species that have already been named, but not yet sequenced. Consequently, sequence-based types would be fragments of a parallel system to which no organismic entity could be related and which, as such, could not be used as a foundation for scientific knowledge.

7. Sequences of reported OTUs are derived, not actual sequences
The whole debate on allowing DNA-sequences as type has originated from the wish of molecular environmentalists to give ‘proper’ names to the numerous enigmatic OTUs they have found (Hibbett et al. 2016), which are only known from their (partial) ITS sequences, but cannot be associated with any known (and barcoded) species. However, there is a common misconception that sequences of an operational taxonomic unit (OTU) correspond to sequences of an actual organism, which is not the case (Ryberg 2015, Callahan et al. 2016, Selosse et al. 2016). This is because OTUs are usually derived from computational methods, such as clustering and thus do not represent primary data (Callahan et al. 2016, Selosse et al. 2016). In most studies dealing with fungi, either a 99 % or a 97 % threshold is assumed (Gweon et al. 2015, Vermeulen et al. 2016, Glynou et al. 2017a, 2018). This means that sequences sufficiently similar to meet the criteria are being clustered together and their consensus sequence is being calculated. In many fungal groups a similarity of 99 %, i.e. 5–6 different nucleotide positions in ITS regions, would encompass several species (Choi et al. 2015), while a similarity of 97 % could consequently encompass dozens of species. In either case, the generation of the consensus sequence is largely dependent on the amount and divergence of reads and the kind of sequences in the dataset that is used for clustering, but it is also influenced by the clustering approach used (Mahé et al. 2014). Thus, OTUs depend on the context in which they are embedded in terms of sampling, PCR, sequencing, and clustering methods and are not easily reproducible (Brown et al. 2015, Oliver et al. 2015, Meisel et al. 2016). In
any case, OTU sequences do not need to correspond to an actual sequence found in an organism, as they are derived sequences. Therefore, they cannot be used as a specific type. Even if the most prevalent individual read sequence were taken as the type for a specific OTU, all the problems attached to such sequences, e.g. the numerous potential artefacts during PCR and sequencing, remained unresolved. Also, it would be unclear where to draw boundaries between the different OTUs as there will always be the potential for overlap between OTUs if they are derived from rather similar sequences.

8. Sequence-based types favour well-funded large mycology labs and leave researchers in developing countries behind

Environmental sequencing can only be pursued by mycologists with access to laboratories with molecular biology equipment and computational infrastructure sufficient for the handling of large datasets. In addition, a large amount of specialised knowledge in molecular biology and computation is needed. Therefore, it is not surprising that the vast majority of environmental sequencing initiatives are run by laboratories in the richest countries of the world. Apart from all the issues mentioned so far, allowing DNA sequences as type would thus create an even larger gap between developing countries and developed countries, leaving the former behind when it comes to the discovery of new species. Even in richer countries, the specialists for certain taxonomic groups can nowadays only be found among amateur mycologists, who may likewise lack the financial resources for sequencing.

9. Allowing sequence-based types would be detrimental for mycology as a discipline

A major issue in mycology is species discovery, i.e. finding the millions of species predicted to exist (Nilsson et al. 2016). If the act of publishing a sequence could be seen as the formal act of introducing a new species, there is a high risk that interest in the actual discovery of the organism would diminish, as the discovery of the actual organism would become the equivalent of an epitypification, which would probably be done for only a few highly prevalent or interesting organisms (Nilsson et al. 2016). There is already a recent trend wherein many taxa are described only on the basis of a ‘new’ ITS sequence by researchers not aware of or neglecting the fact that the majority of fungal species already described have not been barcoded (De Beer et al. 2016). There is also the risk that in systems where quantity in research is valued higher than quality, massive amounts of names without detailed quality checks would be published, flooding fungal nomenclature with tens of thousands of meaningless names that would need to be sorted out in future decades or centuries. If it is possible to publish new species from the computer just on the basis of a DNA sequence, not only knowledge of the morphology, anatomy, chemistry, physiology, life history strategies and ecology of fungi would lose value, but researchers interested in organisal mycology might be discouraged to intensely study and characterise species right from the start, eroding the foundation on which fungal systematics is built. If all the ‘dark matter’ of the cryptic basal lineages of fungi (Grossart et al. 2016) would be formally named based on sequence data, this would probably also discourage the laborious search for these organisms by FISH and other microscopy techniques (Jones et al. 2011, Lazarus & James 2015, Lepère et al. 2016, Matsuyashashi et al. 2017). Another problematic issue is that if sequence data were accepted as type, specimens might be seen as obsolete and only cost-prohibitive museum objects, as they are more difficult to store, curate and preserve than sequence data. This could herald the end of fungaria and the decline of culture collections, even though these might hold the key for substances of unpredictable value for human welfare, such as antibiotics, therapeutically relevant metabolites, as well as platform chemicals and enzymes for biotechnology (Mcclusky et al. 2010, Boundy-Mills 2012, Sette et al. 2013). In groups such as Ascomycota that comprise numerous species that are rich producers of novel secondary metabolites (Helaly et al. 2018), the non-mycologists studying the chemistry of the species often tend to assign the species or genus name according to the most similar DNA sequence found in a BLAST search. This has led to manifold inaccuracies, which has prompted Raja et al. (2017) to encourage a more accurate treatment of the taxonomy of the species. A DNA based typification would send the wrong signal also to the scientists of other communities who, for a correct interpretation of their results, rely on mycologists providing sound species concepts using polyphasic methodology.

10. An introduction of sequence-based nomenclature is impossible at present due to the fast pace at which sequencing technologies develop

The field of high throughput DNA sequencing is a little older than a dozen years (Shokralla et al. 2012), and is still moving quickly, with new technologies evolving and others becoming obsolete (Goodwin et al. 2016, Valentini et al. 2016). The initially revolutionary 454 technology is now virtually obsolete, while long-read sequencing currently enables read lengths of dozens of kilobases, albeit currently with higher error rates (Kennedy et al. 2018). From the very beginning, high throughput sequencing has been used to characterise microbial and fungal communities on the basis of environmental DNA (Hamady et al. 2008, Buee et al. 2009, Jumpponen and Jones 2009). Initially, short barcodes were commonly used (Nilsson et al. 2011), with the recent chemistry on the Illumina MiSeq platform and some modifications, it is possible to obtain complete ITS sequences (Biro et al. 2013). Very recently complete rDNA regions have been sequenced at high quality using single molecule sequencing approaches, such as nanopore sequencing and PacBio sequencing (Wurzbacher et al. 2018). It is difficult to predict what will be possible in the near future, but whole genome sequencing from environmental samples seems to be within reach during the next decade. Right now, there is little agreement on best practices and techniques for sequencing and data handling, which is no wonder, given the fast turnover of sequencing technologies and software packages to deal with the huge amounts of data associated with high throughput sequencing. Thus, it seems premature to devise any rules on how to describe taxa based on sequence data alone. This might become a useful approach when whole genomes become available, even though many of the points mentioned above
would remain valid. At present, any such approaches are probably as useful as it had been to define communication standards for current mobile phones when the first portable telephones appeared in the late 80’s. When devising new rules for the various nomenclatural codes, the potential harm and benefit should always be carefully weighed. And while there is a huge potential for significant damage that would need to be sorted out by generations of future taxonomists, who would ask themselves why there was so little foresight at our time, it is hard to see any positive effects of DNA-based nomenclature for mycology as a discipline.

ACKNOWLEDGEMENTS

This manuscript was stimulated by discussions in the International Commission for the Taxonomy of Fungi (ICTF). Juan Carlos Zamora is gratefully acknowledged for helpful comments on the manuscript. M.T. has been funded by the LOEWE excellence initiative of the government of Hessen, in the framework of IFP (Integrative Fungal Research Cluster) and TBG (Translational Biodiversity Genomics Centre).

REFERENCES


Reasons against a sequence-based nomenclature


