Considerations and consequences of allowing DNA sequence data as types of fungal taxa
Considerations and consequences of allowing DNA sequence data as types of fungal taxa


1Museum of Evolution, Uppsala University, Norbyvägen 16, 75236 Uppsala, Sweden; corresponding author e-mail: jczamora@em.uu.se
2National Central University, Taoyuan City, Taiwan
3Universidad Rey Juan Carlos, 28933 Móstoles, Madrid, Spain
4Avenida Padre Claret 7, 05º G, 09400 Aranda de Duero, Burgos, Spain
5Naturalis Biodiversity Center, Vondelaan 55, 2332AA Leiden, The Netherlands
6Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden

*A first version of this text was prepared by the first eight authors and the last one. The other listed co-authors support the content, and their actual contributions varied from only support to additions that substantially improved the content. The full details of all co-authors, with their affiliations, are included in Supplementary Table 1 for reasons of clarity and space.

**Abstract:** Nomenclatural type definitions are one of the most important concepts in biological nomenclature. Being physical objects that can be re-studied by other researchers, types permanently link taxonomy (an artificial agreement to classify biological diversity) with nomenclature (an artificial agreement to name biological diversity).

Two proposals to amend the International Code of Nomenclature for algae, fungi, and plants (ICN), allowing DNA sequences alone (of any region and extent) to serve as types of taxon names for voucherless fungi (mainly putative taxa from environmental DNA sequences), have been submitted to be voted on at the 11th International Mycological Congress (Puerto Rico, July 2018). We consider various genetic processes affecting the distribution of alleles among taxa and find that alleles may not consistently and uniquely represent the species within which they are contained. Should the proposals be accepted, the meaning of nomenclatural types would change in a fundamental way from physical objects as sources of data to the data themselves. Such changes are conducive to irreproducible science, the potential typification on artefactual data, and massive creation of names with low information content, ultimately causing nomenclatural instability and unnecessary work for future researchers that would stall future explorations of fungal diversity. We conclude that the acceptance of DNA sequences alone as types of names of taxa, under the terms used in the current proposals, is unnecessary and would not solve the problem of naming putative taxa known only from DNA sequences in a scientifically defensible way. As an alternative, we highlight the use of formulas for naming putative taxa (candidate taxa) that do not require any modification of the ICN.

**Key words:** IMC11, nomenclature, speciation, taxonomy, typification, voucherless fungi

**Article info:** Submitted: 14 May 2018; Accepted: 21 May 2018; Published: 24 May 2018.

**INTRODUCTION**

Hawksworth et al. (2016) recently submitted a set of proposals to modify the *International Code of Nomenclature for algae, fungi, and plants* (ICN), aimed at allowing DNA sequences without vouchered specimens to serve as types for fungal taxon names. These proposals were first rejected by the Nomenclature Committee for Fungi (see Turland & Wiersema 2017) and subsequently by the XIX International Botanical Congress (IBC) in Shenzhen, China, in 2017. At the same time, a Special-purpose Committee on DNA sequences as types was proposed to explore and carefully discuss this issue, paving the way for further debate during the next IBC in Rio de Janeiro in 2023 (Turland *et al.* 2017).

However, apparently because of a perceived urgency in the establishment of a system for naming putative new taxa known only from DNA sequences, the same proposals were recently re-published (Hawksworth *et al.* 2018) with the intent that they be discussed and voted on at the forthcoming 11th International Mycological Congress (IMC11) in Puerto Rico in July 2018. The proposals aim at allowing the formal naming of fungal taxa only known by DNA sequences (the “dark matter fungi” of Grossart *et al.* 2016), by authorizing the DNA sequence itself to be the type of a taxonomic name in the absence of a specimen.

The ICN attempts to create “the provision of a stable method of naming taxonomic groups, avoiding and rejecting the use of names that may cause error and ambiguity or throw science into confusion” (Preamble 1). This provision relies on the use of the nomenclatural type, “the face — the desiccated, flattened face to be sure, but still the face — that is attached to the name of a species” (Daston 2004).

In our opinion, the fungal-specific amendments proposed to the ICN by Hawksworth *et al.* (2018) should be rejected on
the grounds that they would have major negative implications for fungal nomenclature and systematics, or more specifically, violate Preamble 1, promote irreproducible science, and fundamentally change the meaning of the type concept compared to how it has been applied during the last century. An informed debate is needed to avoid any unwanted effects of a rushed decision.

THE PROPOSALS

The proposals of Hawksworth et al. (2018) intend to insert a single article, Art. F.4.2, through proposal (F-005), followed by three recommendations, Rec. F.4A.1-3, through proposal (F-006). As only Art. F.4.2 would be mandatory, it is crucial to evaluate proposal (F-005) in particular detail: “(F-005) Insert a new paragraph after Art. F.4.1 as follows: F.4.2. In fungi, when DNA sequence data corresponding to a new taxon have been detected, but no physical specimen has been found to serve as the type of the name of the new taxon (Art. 8.1–8.4), the type may be composed of DNA sequence data deposited in a public repository.’’

The recommendations that follow suggest, in summary, that “the new taxon should be described with reference to a published phylogenetic analysis” (Rec. F.4A.1), that the new taxon “should be represented by multiple sequences obtained in independent studies” (Rec. F.4A.2), and that the sequence should derive from “the molecular regions that are appropriate for delimiting species” (Rec. F.4A.3). These are merely recommendations, however, and need not be followed (as emphasized by Turland & Wiersema 2017).

SPECIES VERSUS DNA SEQUENCES

It has been argued that “the Code serves only to regulate the valid publication of names, not to pass judgment on the scientific hypotheses embodied in names” (Herr et al. 2015). Although nomenclature can be seen as a “remarkable act of applied metaphysics” (Daston 2004), the circumscription of the taxa being named is a fundamentally scientific process. The proposal recommends that a new taxon “be described with reference to a published phylogenetic analysis” (Rec. F.4A.1 of proposal F-006). This wording implies that it is possible to first circumscribe a new taxon by phylogenetic analysis, then name the new taxon using a DNA sequence type that can be unequivocally associated with the new taxon. For the reasons outlined below, this may not readily be the case at the level of species in recombining organisms, which we suspect is where Art. F.4.2 is most frequently going to be applied.

Assuming that species are understood as somehow separately evolving units (e.g. de Queiroz 1998, 2005, 2007, Hey 2006), they can, sooner or later after formation, be detected using a variety of methods (often misleadingly termed ‘species concepts’; Hey 2006), e.g. reproductive isolation (the ‘biological species concept’), morphology, or genealogical monophyly with or without auxiliary criteria like concordance among genes (corresponding to the genetic versions of ‘phylogenetic species concept’). During a simple divergence of one ancestral species into two daughter species, (nearly) neutral loci will inherit random samples of alleles from the ancestral species, some of which are likely to be shared across the daughter species (ancestral polymorphisms). Given time, ancestral alleles will go extinct randomly and new alleles will arise, in the most likely case causing species to appear non-monophyletic on the gene trees. Finally, species will achieve reciprocal monophyly on the gene trees. This process has been known and described in the literature for decades (e.g. Tajima 1983, Takahata & Nei 1985, Neigel & Avise 1986, Nei 1987, Pamilo & Nei 1988, Takahata 1989, Avise & Ball 1990, Hudson et al. 1992, Hey 1994, Harrison 1998, Avise 2000, Hudson & Coyne 2002, Rosenberg 2003, Coyne & Orr 2004, Naciri & Linder 2015) and has been elegantly explained and illustrated by, for example, Leliaert et al. (2014). The lag time from lineage divergence until reciprocal monophyly in neutral loci will depend on the effective population size, generation time, and population structure (Hudson 1990, Wakeley 2000) and its duration will vary stochastically between nuclear loci in recombining organisms (Hudson & Truelli 2003). Obviously, any species recognition protocol requiring reciprocal monophyly will only be able to detect the species long after they diverged (Hudson & Coyne 2002). Positive selection can substantially shorten the time it takes to remove ancestral polymorphisms and finally reach reciprocal monophyly. The proportion of the genome undergoing positive selection during and after speciation appears to be small, however, probably reaching at most a few per cent (e.g., 1.1 and 1.7 % of the genes in humans and chimpanzee, respectively; Bakewell et al. 2007). As an aside, the stochastic process finally leading to reciprocal monophyly in the individual genes also means that there cannot exist a universal divergence threshold for delimiting fungal (or other) species using DNA sequences, not for the very widely used internal transcribed spacer (ITS) region in fungi (e.g. Nilsson et al. 2008, Badotti et al. 2017), nor any other DNA region in any organism group (e.g. Meier et al. 2006 concerning metazoans).

Gene histories, a standard product in applied phylogenetics, cannot automatically be equated with the species history (e.g. Tajima 1983, Pamilo & Nei 1988, Maddison 1997, Knowles & Carstens 2007). There is no reason to think that any DNA region or any organism group is free of mechanisms that create a discordance between the gene and species histories. Such mechanisms have been found to be widespread across the tree of life (e.g. Sota & Vogler 2001, Rautenberg et al. 2008, Bianco-Pastor et al. 2012, Kutschera et al. 2014, Lamichhaney et al. 2015, Garrido et al. 2017, Kudryavtseva & Gladkikh 2017, Meyer et al. 2017, Parks et al. 2017, Peyrégne et al. 2017, Vd’Aenry 2017). Incongruence between gene histories, demonstrating that at least some of them must be different from the history of the species, has indeed also been demonstrated to occur in the fungi (e.g. O’Donnell & Cigelnik 1997, Sung et al. 2007, Harder et al. 2013, Altermann et al. 2014, Saag et al. 2014, Stewart et al. 2014). A conflict between the gene histories and species history is not only caused by the randomness of genetic drift described above. Other mechanisms, all observed also in fungi, obscure relationships among taxa.
and some (the first three) have the potential to cause non-
identifiability of a single DNA sequence: the exchange
of entire nuclei between heterospecific fungal synctia,
horizontal gene transfer, hybridization (sometimes followed
by introgression or allopolyploidy), gene duplication
(including also pseudogene and numt formation), and
intra-individual variability in the ribosomal DNA repeat
caused by limits to concerted evolution (Dean et al. 2005,
Ruths & Naklhe 2005, Jeffroy et al. 2006, Neafsey et al.
2010, Ellison et al. 2011, Lindner & Banik 2011, Roper et
2013, Gladieux et al. 2014, Som 2014, Naciri & Lindner 2015,
Shapiro et al. 2016, Thiéry et al. 2016, Fourie et al. 2017,
Obviously, species delineations generated from a single
marker cannot be evaluated using data from the same
marker, because that would make the argument circular.

We conclude that a DNA sequence of an allele cannot
be seen as “corresponding to” any taxon (the wording of the
proposal), but represents the diversity of alleles of the gene
from which it was derived. An allele cannot be expected to
be unique to the species from which it was derived and we
cannot know whether or not alleles are unique to a species
when sequence data are only available from a single or a
limited number of markers and individuals (e.g. the popular
ITS barcode in fungi; Schoch et al. 2012, Badotti et al.
2017). “If species membership is contingent for organisms
in general, it ought to be contingent for those chosen as the
type specimens for their species” (Levine 2001). Having said
that, some of these pitfalls are more easily detected and
remedied when the number of markers is high and methods
designed to handle them (including but not limited to versions
of the ‘phylogenetic analysis’ prescribed by Rec. F.4A.1) are

IMPACT ON NOMENCLATURAL TYPES
(SPECIMENS VERSUS DNA SEQUENCES)

An acceptance of the proposal would fundamentally alter
the meaning of nomenclatural types. This is because instead of
using a physical object as the type of a name, we would just
use information from a character of the organism as the type.
Indeed, the parallel to the designation of a DNA sequence as
a type would be the designation of information extracted from
organisms (specimens) as types, not with the designation of
specimens as types. In other words, this would be akin to
designating a sample of spore measurements as the type of
an organism. It should be noted that the possibility to select
a description as a type existed before the publication of the
Berlin Code in 1988. However, this option was eventually
rejected by the scientific community, and removed from the
Berlin Code with this note in the Preface: “The provision
that existed for a type to be a description under certain
circumstances — something that many felt amounted to a
repudiation of the type method — has been deleted from the

Names of taxa are applied to organisms, not to characters
of those organisms. Therefore, a physical object should
preferably serve as the type of a name, rather than the
characteristics of that object. By allowing already extracted
data, such as a DNA sequence, to serve as type instead of
the source of the data, new information cannot be obtained
when this is required (see below). In addition, we suspect that
bypassing the current concept of a type is often unnecessary,
because techniques exist to visualize fungal DNA with high
specificity (Baschien et al. 2001, Behrens et al. 2003, Inácio
et al. 2016). Although not yet standard parts of the mycological
toolbox, such techniques can with relative ease be applied to
locate physical specimens even for taxa that cannot currently
be cultivated.

According to the ICN, a nomenclatural type is “that
element to which the name of a taxon is permanently
attached, whether as the correct name or as a synonym”
(Art. 7.2). For species-level taxa and intraspecific taxa,
which are the basic units in taxonomy, a type is “either a
single specimen conserved in one herbarium or other
collection or institution, or an illustration” (Art. 8.1). Why
have researchers agreed to keep these definitions for such
a long time? The answer is straightforward: because types
are an almost never-ending source of information, as they
can be analyzed by different people using different methods
and thus provide new answers. Every time a type specimen
is re-examined, there is an opportunity to extract new
information, which may be useful for solving problems that
are constantly arising as our knowledge increases. Most
types are specimens (especially nowadays) because a
specimen of any living organism is such a complex entity that
it is hard to imagine us being able to extract all the possible
information contained in it. These properties have already
been considered in an editorial of IMA Fungus written by
the President of the International Mycological Association
(Seifert 2017). Therefore, even though the problem of non-
unique characters used for diagnosis is not restricted to
sequence data, the crucial distinction from morphological
descriptions of biological type specimens is that having a
DNA sequence as type virtually precludes the obtaining of
any new information to resolve any taxonomic problems.
In contrast, even illustrations, which are now accepted as
types only in very specific situations (see Art. 40.5 for the
current use of these) and increasingly falling into disuse,
may be a source of overlooked information.

Epitype selection may be seen as a possible solution
in the expected cases when the DNA sequence alone is
insufficient for the precise application of the name of a taxon
(Ryberg & Nilsson 2018). Epitypification was conceived as
a practical solution in cases when the type of a name turns
out to be ambiguous (ICN, Art. 9.8). Epitypes are frequently
designated for old names, and they are not free of undesired
problems affecting nomenclatural stability (Rindi et al. 2017).
Epitypifications have to be based on an existing type, and
are often being made because our knowledge or the present
technology are the limits for extracting the needed information
from the type that already exists. Those limitations may be
overcome by other researchers or by new technologies in the
future. For DNA sequence data, the type itself would always
be the limiting bottleneck, regardless of the researcher’s
skills or the progress of science.
IMPACT ON NAMES OF TAXA AND FUTURE TAXONOMIC STUDIES

The main argument used by Hawksworth et al. (2016), to justify the urgency of allowing DNA sequences as types, is that taxa only known from DNA sequences "require scientific names in order to facilitate communication about them". While researchers indeed need names of taxa to communicate among colleagues and with the general public, those names are linked to information that makes them useful, like biology, distribution, ecology, morphology, physiology, pathology, etc. (Crous et al. 2015). In other words, we are using scientific names because they are meaningful to a wide range of people.

In addition, taxonomists are aware that an increased number of validly published names will not necessarily facilitate communication. On the contrary, in the not uncommon situation in which the same taxon has been named on several occasions, much confusion may arise until the identity of those names is finally settled. Indeed, taxa based solely on DNA sequences not precisely matching any of those present in public repositories have already been described and fallen into more or less immediate synonymy, because the necessary comparisons with previously described taxa were not undertaken (Gams 2016). The proposals would promote such bad practice.

An undesired side-effect that should also be considered is that, in practice, few researchers will be devoted to re-describing (or actually describing) species that have been previously named based on just a DNA sequence. This has several causes, but among them, there is an important bias in research journals disfavoring the publication of re-descriptions of already known taxa, versus the description of new taxa. Another reason is time constraints, since it is not uncommon that specialists do not have the time to properly describe all of the numerous undescribed species they are aware of. This makes them focus on those that are more likely to be published as new species and not on those that have been already described, even if previous descriptions are faulty or defective. Anyhow, having numerous names only based on DNA sequences and few descriptions of the actual organisms would create an enormous number of validly published names applied to taxa for which virtually no information exists.

RELIABILITY AND EXTENT OF DATA

The proposed Art. F.4.2. effectively means that any DNA sequence of any region and extent, generated by any procedure or taken from a public repository, could serve as the type of a name of a taxon somehow indicated to be new. In practice, the sequence selected as the type could range from an oligonucleotide to the entire genome. The proposal provides very little guidance, except for the recommendations that the type sequence should be represented by "multiple sequences" and that the selected marker should be "appropriate for delimiting species" (proposed Rec. F.4A.2, F.4A.3). It is not clear what 'multiple' means or how a marker is established as universally 'appropriate'. One can infer, however, that the 'appropriate' marker will, in most applications, be the ITS region, which has been dubbed as the primary barcode marker in fungi (Schoch et al. 2012).

A major concern is the reliability of the DNA sequence data (Bridge et al. 2003, Nilsson et al. 2006). PCR or cloning errors (including the introduction of chimeras), DNA degradation, and post-processing of chromatograms, have been shown to be a source of sequence variation in at least some groups (Haas et al. 2011, Sandoval-Sierra et al. 2014, Hughes et al. 2015, Strid et al. 2015, Aas et al. 2017, Nilsson et al. 2017, Thielecke et al. 2017, Bieker & Martin 2018). Such DNA sequences are not real and cannot be checked or corrected without access to a physical specimen or, as a minimum, access to the raw sequence reads (Tripp & Lendemer 2014). If accepted as types, this means mycology would embrace irreproducible science.

The concerns outlined here, in combination with the risk of comparing non-orthologous sequences or incompletely concerted copies of the ribosomal DNA, are really about scientific quality and not nomenclature per se. However, nomenclature assumes that taxa are first delineated, then named. The proposal, if implemented, would risk opening the floodgates to poor data and questionable scientific practice being translated into formally named taxa that will throw fungal taxonomy into paralysis and disrepute.

CANDIDATE NAMES

If we really want to strive for a comprehensive code of nomenclature able to cover all living organisms, it is necessary to consider the rules of the other existing codes of nomenclature. For our purposes, these are mainly the International Code of Zoological Nomenclature (ICZN; Ride et al. 1999) and the International Code of Nomenclature of Prokaryotes (ICNP; Parker et al. 2015). Also, it is important to consider the use of nomenclature by specialists in different taxonomic groups. In general, we think it is better to strive for standardization of rules instead of sharpening the differences between Codes. The goal should be to create a solid code of nomenclature that, some day, may perhaps cover all living organisms with all their peculiarities (e.g. the BioCode initiative; Greuter et al. 2011, http://www.bionomenclature.net/biocode2011.html).

An interesting formula concerning taxa that cannot be properly described under the rules of a code of nomenclature is the use of the term "Candidatus". Originally, this working term was proposed by Murray & Schiefer (1994), and soon after improved by Murray & Stackebrandt (1995) for "describing prokaryotic entities for which more than a mere sequence is available but for which characteristics required for description according to the Code are lacking". It was proposed because, under the rules of the ICNP, a prokaryotic organism can only be validly described if the type, which in this case is a living strain, can be conserved as an axenic culture. There are of course thousands of prokaryotic taxa that are not cultivable in such a way. Many of them can, however, be studied with regard to morphology, ecology, metabolism, DNA data, etc. For fungi, having such additional information for a particular cluster of DNA sequences (never a single one), or several
DNA regions from the same organism (ultimately and ideally, a complete genome), would be essential to ensure that a true taxon is being provisionally named, and to comply with basic scientific standards.

The Candidatus working term has proved to be a good solution for microbiologists who want to respect the rules of the ICNP as well as to apply useful names to certain taxa. Being aware that important information (e.g. a proper living strain as type) is lacking to allow a formal description, such taxa can be validated when the requirements of the ICNP are fulfilled. The best example of how well this alternative nomenclature works is the Candidate Phyla Radiation, a huge, well-known and well-communicated group of Bacteria that was proposed based on the combined information of hundreds of genomes, obtained from single cells as well as metagenomics (Hug et al. 2016, Danczak et al. 2017).

The alternative of using preliminary names for taxa only known from DNA data has already been proposed by Öpik et al. (2009) as “virtual taxa”, by Taylor (2011) as “ENAS fungi”, by Kõljalg et al. (2013) as “species hypothesis”, and indeed also by Hibbett et al. (2011) as “candidate species”. We think this is an interesting idea that should be further explored and discussed in the future. Such candidate names can be re-evaluated and possibly formally described in the future when enough information has become available to provide a good taxon description (see also Seifert 2017). Finally, they could be used with some freedom, since no specific rules within the codes of nomenclature apply for invalidly published names. If a major concern about fungi only known from DNA sequences is that “they do not enter names-based taxonomic databases” (see Herr et al. 2015), a reasonably easy solution would be to allow the registration of candidate or putative names in those databases, in the process making it clear that those names have not yet been validly published because one or more of the requirements for valid publication are lacking (e.g. http://www.bacterio. net/-candidatus.html for candidate names of prokaryotic taxa).

CONCLUSIONS

We consider the proposals by Hawksworth et al. (2018) highly problematic for the following reasons:

- DNA sequence types will have a very low information content; subsequent extraction of additional data or verification of the already extracted data will not be possible.
- Two different taxa may share identical DNA sequences at a given locus, even for already tested barcoding markers. Conversely, not all members of a species can be assumed to share the same DNA sequence at a specific locus.
- Intraspecific (or even intraindividual) differences in the DNA sequence of a marker may be comparable to or exceed interspecific differences.
- Some DNA sequences generated through different sequencing techniques may be artifacts and consequently not represent reality. The proposal does not say anything about data validation other than a recommendation that the DNA sequence should be represented by ‘multiple sequences’.
- The proposal promotes the mechanical production of taxon names based on minor sequence divergence, without taking any other data (such as genetic variability or already described taxa) into account. Much downstream time will have to be spent by future mycologists gathering additional information.
- As taxa with DNA sequence types accumulate, the description of a new species will be increasingly difficult without DNA sequence data. Describing new species based on the morphology of unsequenced material will in practice not be feasible if the possibility exists that this species has been described based on a DNA sequence.
- Since the proposals allow any part of the genome to be used as a DNA type, situations in which different taxa may have been described using different parts of the genome will force researchers to sequence a variety of loci to establish whether an earlier name already exists. Likewise, a single taxon may be described as novel several times using different genomic regions as type. This will be impossible to detect without a specimen from which different genomic regions can be sequenced and may contribute to the description of unnecessary new names.

FINAL REMARKS

As discussed above, there are alternative ways of communicating the existence of taxa only known from DNA data, which do not require modifications to the ICNP. Instead of allowing DNA data as types for taxon names, database registration of candidate names can be used for putative new taxa, when their existence has been made plausible based on various sources of information (including but not limited to DNA sequences). A functional system for environmental sequences under the Candidatus or species hypotheses approach could result from a carefully selected set of requirements to ensure high-quality data and reproducibility.

We submit that proposals F-005 and F-006, for the reasons outlined here, will not solve the problems they are intended to solve, disregard knowledge acquired through decades of research in the genetics of speciation, and will instead create confusion and substantial extra work for contemporary and future mycologists. We all have the responsibility to maintain the scientific standards of reproducibility as well as to provide well-considered rules for coming generations, so they can improve on our work and take appropriate, well-informed taxonomic decisions using all available information.

ACKNOWLEDGEMENTS

We want to express our gratitude to all people who enriched both the text and our views by generously sharing their opinions and scientific experience.
REFERENCES


