

## **Rules and Format for Taxonomic Nucleotide Sequence Annotation for Fungi: a proposal**

### **The need for third-party sequence annotation**

Taxonomic names attached to nucleotide sequences occasionally need to be changed for three main reasons: 1) the original biological material from which the sequence was derived may have been misidentified or incompletely identified, 2) taxonomic concepts may have changed since the sequence was deposited, or 3) sequence or PCR artifacts are discovered (see Hugenholtz & Huber 2003).

Naming errors proliferate and lead to a loss in utility of the sequence database. Many published reports have now demonstrated that this problem is widespread and growing in our public sequence databases (Nilsson *et al.* 2006). This problem is particularly important for ecological studies in which sequences retrieved from natural substrates such as wood or soil are used for identification of the unseen organisms within these environments. Although the original authors of a sequence submission can make taxonomic changes, this is not an efficient or a sustainable system. It is usually a third-party user of the sequence that first identifies the need for a change and is motivated to make it. In contrast, the original submitter(s) may have left science, or may not be interested enough in the problem to devote time and energy to correct it. For these reasons, 257 scientists signed an open letter calling for third-party annotation of the public sequence database (Bidartondo *et al.* 2008), but the details of how third-party annotation would be done were not specified in the letter.

The goal of this white paper is to develop a consensus on how such changes should be made. It would be ideal if this consensus resulted in the adoption of third-party annotation within a reference collection at NCBI. However, even if NCBI acceptance or implementation does not occur in a timely way, reaching consensus would provide a system that a third-party database could adapt immediately to provide a temporary home for a reference set of sequences. These sequences could then be used for ecological applications and other situations where sequence data are used as the primary tool for identification, and they could ultimately provide a tool for automated classification of unknown sequences (Nilsson *et al.* 2009).

### **An ideal third-party annotation system would have the following characteristics:**

- 1) It would preserve the initial accession data.
- 2) It would allow for updating of the taxonomy of individual accessions as new information or concepts become available.

- 3) It would prevent excessive or repetitive changes, while allowing necessary changes to occur efficiently.
- 4) It would produce a record of annotations that includes date of annotation, reasons for annotation, and name and address information of annotators.
- 5) The annotation record should be easily queried and displayed.
- 6) The format of annotations needs to be compatible with current NCBI file formats, but adaptable enough so that it does not inhibit the development of more sophisticated data structures.
- 7) It should be an easy system that could be accessed by multiple qualified users.

### **Proposal - an accumulation of opinions**

The need for updating taxonomic names is not unique to nucleotide sequence databases. Indeed it has existed for centuries in herbaria where it has been dealt with by a strikingly simple system: the original name remains associated with the specimen, but additional annotation labels are appended to express new opinions and knowledge. If this basic model were used for nucleotide sequences it could achieve our first two goals: preservation of the original record while providing the ability to append new information to the record. By specifying a few additional rules for who can annotate, and how such annotations are done, the next four goals can also be achieved. The final goal of ease and access will require a software interface to be created or modified.

### **Who can annotate?**

Ideally we want annotation to be a scientific community-wide project, because it is the community that will find errors and is motivated to correct them during the course of systematic and ecological work. In addition, if annotation is left to a small number of professionals it is likely to occur much too slowly and it may not reflect the views or needs of the community at large. However, opening up the annotation process to the broad community creates the potential for abuse by a small number of people.

As a compromise, we propose a vetted community where anyone interested and qualified could request and be granted the ability to annotate, but this privilege could be withdrawn if the activities of the individual were deemed inappropriate. Evidence of qualifications might include, prior publication record, an advanced degree in an appropriate field, or current graduate studies toward such a degree. Applicants that have these basic qualifications would be granted annotation privileges and would retain them unless they abuse the privilege. A committee assembled from the active curators would approve new curators and determine whether annotation privileges needed to be revoked. In addition, the rules for annotation, which are described below, are designed to limit abuse.

### **When can an annotation be made?**

Taxonomic annotation can be made when new evidence or analyses provide a compelling case that an existing name and/or its annotations are incorrect or insufficient. This evidence should be published, but the publication need not be focused on systematics and the annotator need not be an author on the publication.

### **What sequences can be taxonomically annotated?**

The primary reason for taxonomic third-party annotation is to make the data more useful for identification of unknowns. For this reason, annotation would initially be restricted to a small subset of loci used for identification. In fungi this would currently be ribosomal genes and spacers, notably the internal transcribed spacers (ITS), but the list could be expanded to include selected nuclear and mitochondrial protein-coding loci.

Of all the deposited sequences for these loci, we would likely select a subset to be included in a reference collection and be subject to taxonomic curation. This subset would include all sequences derived from barcoding, and other voucher-based efforts, but would also include all other sequences that meet the basic requirements of sequence quality. These requirements might include minimum length of the target locus, low frequency of ambiguity codes, and non-chimeric origins. The broad inclusion of all sequences is crucial for an identification tool, because many groups are currently known only from their sequences (Weiss *et al.* 2004, Porter *et al.* 2008).

### **Limits to the number of annotations**

No limit will be set on the total number of taxonomic annotations to a particular sequence, but each annotation must be unique. Because of this restriction the number of annotations per accession is likely to be small and the goal of avoiding excessive or repetitive changes will be met.

### **The format**

The format for taxonomic annotations would be a short comment in the "**Source**" section of the "**Features**" portion of the sequence accession. The comment would be tagged specifically as a taxonomic annotation and following this tag it would give all the needed information in an ordered series separated by semicolons. Standardizing the format in this way would allow this information to be searched for and displayed in any way a user might desire. Here is an example of a generic template for such a comment:

```
/Taxonomic annotation = "rank: new name here; names of annotator(s);  
address; date; brief reason; publication citation"
```

**Qualifiers to names** (e.g. *aff.*, *cf.*, *group*, or *sensu*) that indicate affinities, doubt,

informal groupings, or particular concepts of the taxon are occasionally useful and can be added in the same ways that they are on typical specimen labels. The use of *aff.* ("with an affinity to") should be used for sequences close to, but likely distinct from a particular taxon, while *cf.* ("compare with") should be used for designations that are tentative and need further comparison to the indicated taxon. "Group" can be used in the case of species complexes, and numbers can be applied to indicate distinct clades within a complex when current names are insufficient.

**Some examples:**

The sequence depositor, Holly Gardener, calls the sequence "Uncultured soil fungus". This remains the main tag line for the sequence unless she changes it. The first person to annotate the sequence decides it is an *Inocybe* species and places the following annotation:

/Taxonomic annotation 1 = "genus: *Inocybe*; Joe Ecologist; Harvard University; 05-01-2009; phylogenetic analysis; *Molecular Ecology* 18: 596-97"

The next person adds a species name:

/Taxonomic annotation 2 = "Species: *Inocybe subochracea*; Jane Mycologist; Kew Gardens; 05-01-2010; phylogenetic analysis; *Mycological Research* 114:900-915"

If Jane later decides the name isn't correct she could annotation again:

/Taxonomic annotation 3 = "Species: *Inocybe olympiana*; Jane Mycologist; Kew Gardens; 09-06-2016; phylogenetic analysis AND Nomenclatural correction; *Mycological Research* 114:900-915 AND *Mycotaxon* 110:50-51"

Note that in this case the nucleotide sequence would have four names: the original "Uncultured soil fungus" would be the primary name for the sequence, unless it was changed by Holly; "*Inocybe* sp." by Joe would be the first annotation, and "*Inocybe subochracea*" and "*Inocybe olympiana*" by Jane are the third and fourth. In this way, all of the data of the original accession and the opinions of each annotators would be preserved. One could display any or all of these names when the sequence is retrieved. One could also select only the most recent name to display, or only the original, or all of those except Joe's (if the user disagrees with his annotations), or one could display the accession names by any other complex criteria of the user's choice. In any way one chooses to display the names, the record would be flagged to indicate there are differences of opinion on the name.

What should be done if the original submitter subsequently agrees with one of the annotations and decides to change the original name on the accession? For example, Holly wants to change her accession's name from "unknown Basidiomycota" to "*Inocybe olympiana*". She has the right to do so, but all annotations would be retained, and to preserve all of the original data her first name applied to the sequence would now become a special taxonomic annotation with a slightly different format:

/Original Taxonomic Name= "Uncultured soil fungus"

What should be done if the material upon which the sequence was based was incorrect? For example if the sequence is based on some mold that was growing on the specimen or a culture from a specimen turns out to be a contaminant instead of the intended species. In both cases the sequence was likely to be mislabeled because it would be linked to the specimen from which it was thought to be derived. These cases would be dealt with exactly as outlined above, but in the "reason" field of the annotation one might add something like "contaminant within *Glomus* spore, phylogenetic analysis".

A special case of annotation arises from chimeras a common type of PCR artifact. These can sometimes be detected by software designed for that purpose (Huber *et al.* 2004) or by phylogenetic analysis of the separate ends of a sequence. We should try to exclude these when the reference collection is assembled, but if a sequence slips through it should be labeled as a chimera directly:

/Taxonomic annotation 1 = Chimera; Frank Lee Blunt; USDA Peoria; 04-01-2010; phylogenetic analysis; *Appl. Env. Micro* 170: 1596-97"

Other sequence based errors or quality issues might affect the value of the sequence for taxonomy. For example there may be a high level of base ambiguity indicated as N, Y, R, etc., coupled with unlikely insertions. Although these problems may be troublesome, the issue is not a taxonomic one if the name applied to it is correct. However, it could become a taxonomic issue if the sequence quality is so poor that when it is compared to many other sequences from the taxon it stands out on an abnormally long branch. In this case one might apply a *cf.* tag to the taxon name and provide a reason such as "long branch appears to be due to sequence quality issues". However, the situation might be better dealt with by excluding that particular sequence from the reference collection.

The final goal of ease and access should be designed into a website for this purpose. If NCBI were to adopt this method tomorrow, it could be done without such an interface in the non-automated way that current changes are made: one

would email the proposed change to a curator. This provides a gate-keeping function that insures some quality control. However, a better way would be to provide web access for curation, but have all changes approved by a curator before they are appended to the record. This would speed up the process and preserve the gate-keeping utility of the current system. It could insure that the format for annotations was followed making retrieval and analysis of the data efficient.

### References cited

- Bidartondo M, al. e. 2008. Preserving the Accuracy in GenBank - an open letter. *Science* 319:1616.
- Huber T, Faulkner G, Hugenholtz P. 2004. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* 20:2317-2319.
- Hugenholtz P, Huber T. 2003. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *International Journal of Systematic and Evolutionary Microbiology* 53:289-293.
- Nilsson HR, Bok G, Ryberg M, Kristiansson E, Hallenberg N. 2009. A software pipeline for processing and identification of fungal ITS sequences. *Biology and Medicine* 4:(in press).
- Nilsson RH, Ryberg M, Kristiansson E, Abarenkov K, Larsson K-H, Koljalg U. 2006. Taxonomic Reliability of DNA Sequences in Public Sequence Databases: A Fungal Perspective. *Plos Biology* 1:e59.
- Porter TM, Schadt CW, Rizvi L, Martin AP, Schmidt SK, Scott-Denton L, Vilgalys R, Moncalvo JM. 2008. Widespread occurrence and phylogenetic placement of a soil clone group adds a prominent new branch to the fungal tree of life. *Molecular Phylogenetics and Evolution* 46:635-644.
- Weiss M, Selosse MA, Rexer KH, Urban A, Oberwinkler F. 2004. Sebaciales: a hitherto overlooked cosm of heterobasidiomycetes with a broad mycorrhizal potential. *Mycological Research* 108:1003-1010.