MycoKeys 4: 37–63 (2012)
doi: 10.3897/mycokeys.4.3606
www.pensoft.net/journals/mycokeys

RESEARCH ARTICLE

A peer-reviewed open-access journal
MycoKeys
Launched to accelerate biodiversity research

Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences

R. Henrik Nilsson[1], Leho Tedersoo[2,3], Kessy Abarenkov[2], Martin Ryberg[4], Erik Kristiansson[5], Martin Hartmann[6,7], Conrad L. Schoch[8], Johan A. A. Nylander[9], Johannes Bergsten[10], Teresita M. Porter[11], Ari Jumpponen[12], Parag Vaishampayan[13], Otso Ovaskainen[14], Nils Hallenberg[1], Johan Bengtsson-Palme[15], K. Martin Eriksson[1], Karl-Henrik Larsson[16], Ellen Larsson[1], Urmas Kõljalg[2,3]

1 *Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden* **2** *Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia* **3** *Natural History Museum, University of Tartu, Tartu, Estonia* **4** *Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610, USA* **5** *Department of Mathematical Statistics, Chalmers University of Technology, 412 96 Göteborg, Sweden* **6** *Molecular Ecology, Agroscope Reckenholz-Tänikon Research Station ART, Zurich, Switzerland* **7** *Forest Soils and Biogeochemistry, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland* **8** *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892, USA* **9** *Department of Biodiversity Informatics, Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden* **10** *Department of Entomology, Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden* **11** *Department of Biology, McMaster University, Life Sciences Bldg., 1280 Main Street West, Hamilton, ON L8S 4K1, Canada* **12** *Division of Biology, Kansas State University, Manhattan, KS 66506, USA* **13** *Biotechnology and Planetary Protection Group, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA* **14** *Department of Biosciences, University of Helsinki, PO Box 65, FI-00014 Helsinki, Finland* **15** *Department of Neuroscience and Physiology, The Sahlgrenska Academy, University of Gothenburg, Box 434, 405 30 Göteborg, Sweden* **16** *Department of Research and Collections, University of Oslo, Natural History Museum, Postboks 1172, Blindern, 0318 Oslo, Norway*

Corresponding author: *R. Henrik Nilsson* (henrik.nilsson@bioenv.gu.se)

Academic editor: *Thorsten Lumbsch* | Received 29 June 2012 | Accepted 14 August 2012 | Published 5 September 2012

Citation: Nilsson RH, Tedersoo L, Abarenkov K, Ryberg M, Kristiansson E, Hartmann M, Schoch CL, Nylander JAA, Bergsten J, Porter TM, Jumpponen A, Vaishampayan P, Ovaskainen O, Hallenberg N, Bengtsson-Palme J, Eriksson KM, Larsson K-H, Larsson E, Kõljalg U (2012) Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. MycoKeys 4: 37–63. doi: 10.3897/mycokeys.4.3606

Abstract

Molecular data form an important research tool in most branches of mycology. A non-trivial proportion of the public fungal DNA sequences are, however, compromised in terms of quality and reliability, contributing noise and bias to sequence-borne inferences such as phylogenetic analysis, diversity assessment, and barcoding. In this paper we discuss various aspects and pitfalls of sequence quality assessment. Based on our observations, we provide a set of guidelines to assist in manual quality management of newly generated, near-full-length (Sanger-derived) fungal ITS sequences and to some extent also sequences of shorter read lengths, other genes or markers, and groups of organisms. The guidelines are intentionally non-technical and do not require substantial bioinformatics skills or significant computational power. Despite their simple nature, we feel they would have caught the vast majority of the severely compromised ITS sequences in the public corpus. Our guidelines are nevertheless not infallible, and common sense and intuition remain important elements in the pursuit of compromised sequence data. The guidelines focus on basic sequence authenticity and reliability of the newly generated sequences, and the user may want to consider additional resources and steps to accomplish the best possible quality control. A discussion on the technical resources for further sequence quality management is therefore provided in the supplementary material.

Key words

ITS, sequence reliability, sequence quality control, fungi, databases, barcoding

Introduction

The inconspicuous and largely subterranean or endophytic nature of much of fungal life presents a challenge to mycology. Many fungal lineages do not seem to produce tangible fruiting bodies, and for those that do, the factors promoting - and acting against - fruiting body formation are only partly understood. As a result, most sampling sites and habitats host a much greater fungal diversity than the above-ground view offered by fruiting bodies would lead the observer to believe (Porter et al. 2008; Hibbett et al. 2011). Furthermore, discriminatory yet easily assessed morphological characters are something of a rare commodity in mycology, and morphology alone often falls short of providing unequivocal species identification and delimitation. For these and other reasons, mycologists were quick to embrace molecular (DNA sequence) data as a research tool in the early 1990s (Horton and Bruns 2001; Anderson and Cairney 2004). Today, DNA sequences represent a key source of information in nearly all branches of mycology, including systematics, taxonomy, and ecology (Stajich et al. 2009), and the landmarks include the establishment of a phylogenetic backbone and a classification system for the fungal kingdom (Blackwell et al. 2006; James et al. 2006; Hibbett et al. 2007).

For all their advantages, molecular data do not solve all open research questions in mycology, and examples of where the misuse and misinterpretation of molecular data hampered mycological progress are easy to point out (Nilsson et al. 2006). Sequences of compromised technical quality or of incorrect taxonomic or ecological annotations are major contributors in this respect in that they may lead researchers to erroneous results and conclusions. When such entries are made publicly available through the international sequence databases, their compromised integrity becomes a problem not only for the researcher who generated them in the first place but for the entire mycological - indeed, scientific - community. Several studies have reported on the various shortcomings of the public DNA sequence corpus (e.g., Gilks et al. 2002; Harris 2003; Bidartondo et al. 2008), but none have succeeded in halting the continual submission of substandard entries to the databases. On the contrary, there are indications that the proportion of several classes of compromised sequences - such as chimeras and reverse complementary sequences - increases over time (Abarenkov et al. 2010b). While very experienced users may perhaps be able to look through such broken data, many others may not be in a position to do so, particularly not since a growing number of people from outside mycology - even outside the academia - now use fungal sequence data as a part of their work. The highly automated nature of many sequence analysis pipelines similarly makes software suites susceptible to several kinds of sequence errors - such as incorrect taxonomic annotations - since these automata are often built to accept certain classes of information at face value.

The most popular genetic marker for mycological research questions at and below the genus level is the nuclear ribosomal internal transcribed spacer (ITS) region, a ca. 450–650 base pair (bp.) region consisting of the two variable spacers ITS1 and ITS2 and the intercalary, highly conserved 5.8S gene (Begerow et al. 2010). In addition to being widely used for phylogenetic inference and in systematics, the ITS region is proposed as the formal fungal barcode and forms the primary choice for molecular identification of fungi from environmental samples (Vrålstad 2011; Schoch et al. 2012). Several of the present authors have spent significant time pursuing compromised ITS sequences in the International Nucleotide Sequence Databases (INSD: GenBank, ENA, and DDBJ; Karsch-Mizrachi et al. 2012) and UNITE (Abarenkov et al. 2010a; http://unite.ut.ee) or have worked with sequence reliability in other respects. Over time we have noticed several features that signal high-quality, as well as substandard, ITS entries. The most striking observation is probably that, in nearly all cases, severely compromised ITS sequences can be detected manually using just a few simple guidelines (Table 1), without the assistance of technical software packages or access to significant computational power. Many of these guidelines have been put in writing by us and others, but they are scattered across the literature and often mentioned just in passing. In addition, several of them are published in outlets rarely consulted by mycologists. The present publication aims to bring those guidelines and observations on how to establish basic authenticity and reliability of newly generated ITS sequences together in a single, easily digestible publication. The guidelines are simple and straightforward to apply; substantial bioinformatics expertise is not required, and only on-line resources of the paste-and-click type are used. Their simple nature notwithstanding, we believe that these guidelines would have caught the vast majority of the present severely compromised fungal ITS sequences in the public corpus, had they been available and applied at the time of data generation and accessioning.

We would like to stress that the guidelines described here focus on basic sequence authenticity and reliability; they are certainly no panacea for sequence quality management. Their purpose is to assist in pruning severely compromised entries from newly generated, nearly full-length (typically, but not exclusively, Sanger-derived) fungal ITS datasets before those sequences are put to scientific use. The target audience comprises researchers who have just started to use molecular tools (e.g., students) as well as those who otherwise would have taken little action in the direction of quality management.

Table 1. Overview of the five guidelines.

Target of guideline	Way of getting there
1. Establish that the sequences come from the intended gene or marker	Do a multiple alignment of the sequences and verify that they all feature some suitable, conserved sub-region (here the 5.8S gene)
2. Establish that all sequences are given in the correct (5' to 3') orientation	Examine the alignment for any sequences that do not align at all to the others; re-orient these; re-run the alignment step; and examine them again
3. Establish that there are no (bad cases of) chimeras in the dataset	Run the sequences through BLAST in INSD/UNITE and verify that the best match comprises more or less the full length of the query sequences
4. Establish that there are no other major technical errors in the sequences	Examine the BLAST results carefully, particularly the graphical overview and the pairwise alignment, for anomalies
5. Establish that any taxonomic annotations given to the sequences make sense	Examine the BLAST hit list to see that the species names produced make sense

For the user wishing to apply the most advanced and technical quality control solutions to a new dataset right from the start, we provide an account of the bioinformatics of ITS sequence quality control in Appendix. One is nevertheless mistaken to believe that sequence reliability is a matter of bioinformatics only; taxonomic knowledge and common sense are just as important, if much more difficult to algorithmize. What follows is an attempt at a joint treatment of these three aspects.

A word on the query and reference datasets

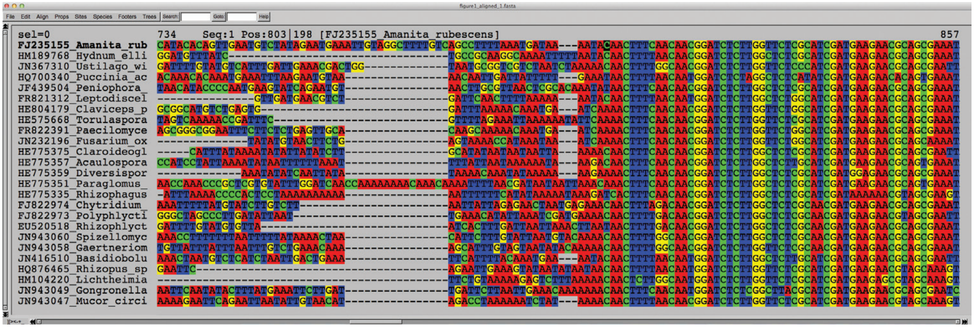
The sequences in INSD and UNITE are often used as reference datasets to which newly generated ("query") sequences are compared in pursuit of taxonomic and ecological annotation. Neither INSD nor UNITE seek to store full ITS sequence datasets generated by next-generation sequencing (NGS) technologies such as 454 pyrosequencing (Margulies et al. 2005), at least not as primary sequences. The sheer volume and the high frequency of platform-generated sequencing errors derived from NGS approaches necessitate extensive, elaborate quality control measures (Gilles et al. 2011; Quince et al. 2011), and the guidelines presented here should certainly not be used as a replacement for those. Indeed, the present paper primarily targets ITS sequences derived through traditional Sanger sequencing, that is, ITS sequences that usually cover more or less the full length of the ITS region (≥500 bp.). The guidelines thus apply first and foremost to research endeavours where full-length ITS sequences are used, including most ITS-borne studies in systematics, taxonomy, and ecology. Many data mining efforts also fall within the scope of the guidelines, as do the core ITS sequences of INSD/UNITE.

Much of the following will apply also to genes and markers other than the ITS region – particularly the neighbouring ribosomal small subunit (SSU) and large subunit (LSU) genes - and it will certainly apply to the ITS region in groups of organisms other than fungi. Nevertheless, for the sake of example, the user is assumed to have a newly generated fungal ITS dataset (with chromatograms), ideally of near-full-length sequences or at least sequences covering approximately the same part of the ITS region. A proportion of the sequences is assumed to be annotated to various hierarchical classification levels, such as "Uncultured chytridiomycete", "*Penicillium* sp.", and "*Amanita muscaria*". To avoid overly simplified examples, we will furthermore assume that the data offer some degree of taxonomic complexity and span several fungal phyla and multiple orders. If the dataset is small - say fewer than 50 sequences - the user should probably consider each sequence individually. For datasets up to a few hundred sequences, the user could use a clustering tool such as the BLASTclust implementation at http://toolkit.tuebingen.mpg.de/blastclust to reduce the dataset to one representative sequence per "species" or operational taxonomic unit (OTU; Blaxter et al. 2005). The BLASTclust settings of 97–98% similarity over at least 90% of the length of the shortest sequence in a pairwise alignment will do a reasonable job at approximating the species level in mixed-fungi datasets. For the remainder of this document, the user would then only have to consider one (representative) sequence per such OTU, bypassing the need to address large numbers of near-identical entries. For larger datasets still, the user could further reduce the BLASTclust settings to 85% clustering similarity or even somewhat lower, provided that the length criterion is kept at 90%. The clustering step is optional and only meant as a way to reduce the number of sequences in need of examination; the present paper does not seek to give advice on how to cluster sequences into OTUs for purposes of richness estimation or similar endeavours. While the clustering step removes the user one level from the actual sequence data, we have found the difference to be negligible in terms of basic sequence authenticity and reliability. If any of the clusters contain two or more sequences with full or partial taxonomic annotations, the user should take the opportunity to skim through these to verify that they make approximate sense, meaning that the sequences in the cluster are expected to be annotated as closely related taxa. A cluster with the confamilial ascomycete genera *Penicillium* and *Aspergillus* would probably make sense under the relaxed clustering settings discussed here; a cluster with *Penicillium* (*Ascomycota*) and *Amanita (Basidiomycota)* would not. In the latter case, one or more of the sequences are mislabelled or otherwise deficient, e.g., chimeric. The truly impatient user may now make use of the fact that severely compromised sequences tend to be unique in the nature of their misfortune and thus come out as singletons (clusters of only one sequence) in the clustering process (cf. Huse et al. 2010). However, we argue that checking singletons only is a low-resolution approach that should be reserved for the largest of datasets (more than ~5,000 sequences), and that each sequence or at least representative OTU sequence (preferably the most common sequence type, rather than the consensus sequence or the longest sequence, of each OTU) in smaller datasets should be individually scrutinized using the guidelines provided below.

Guideline 1. It is simple to check that all query sequences represent the ITS region

Upwards of five hundred public sequences are, or have previously been, annotated as ITS sequences when they in fact have been shown to represent other genes or markers or are noise (seemingly random nucleotide letters) throughout. The reasons could be many and range from primer matches to unexpected parts of the genome at hand to the mixing up of test tubes, files, or individual sequences. These sequences contribute significant noise to any data-mining effort targeting the fungal ITS sequence corpus by, e.g., inflating diversity estimates. For molecular identification of fungi, these sequences pose something of an indirect problem, since they are very unlikely to show up in ITS-based BLAST searches (Altschul et al. 1997; documentation at http://www.ncbi.nlm.nih.gov/books/NBK1762/). Nevertheless, a user - knowing that a particular species is present through an ITS sequence in the reference database - may want to confirm the hypothesized taxonomic affiliation of a newly generated ITS sequence, only to arrive at what seems to be a proof that the newly generated sequence does not belong to that very species. In other words, it is a matter of database integrity that genetic annotations really reflect the true marker in question.

An expedient way to ensure that all query sequences represent the ITS region is to compute a multiple sequence alignment in any of a number of on-line multiple alignment services, notably MAFFT (http://mafft.cbrc.jp/alignment/server/ ; Katoh and Toh 2010). Such quickly derived, manually unedited multiple alignments of the ITS region are of limited scientific usefulness save one aspect: the highly conserved, ca. 160 bp. 5.8S gene of the ITS region will form a firm anchor in the middle part of nearly any such alignment. Thus all sequences for which the 5.8S is aligned in this way must be ITS sequences; it is inconceivable that they would produce a good alignment to the 5.8S if they in fact represent a different gene or marker altogether or if they were composed of stochastic, artefactual nucleotide data. Figure 1 shows an alignment featuring five sequences each of the fungal phyla *Ascomycota*, *Basidiomycota, Glomeromycota, Chytridiomycota,* and *Zygomycota* s.l.*;* the reader will probably agree that the 5.8S is easy to spot, despite the disparate taxonomic scope of the sequences. The obvious conclusion is that all sequences in that alignment represent the ITS region. The user is recommended to have MAFFT order the sequences in the alignment by similarity (“Output order: Aligned”), which normally has the effect of forcing any deviant sequences to the bottom of the alignment (or to produce separate sequence blocks that do not align well together). The separation of non-deviant from deviant sequences makes the former much easier to look at and the latter much easier to spot in the first place. The MAFFT server usually returns even large alignment jobs within half an hour, and to scroll down the alignment along the characteristic 5’ (“left”) end of the 5.8S (cf. Figure 1 or Hibbett et al. 1995) in an alignment editor to check for alignment compliance should not take more than one minute. After that minute – if the 5.8S was found in all sequences - the user can be



```
sel=0                734    Seq:1 Pos:803|198 [FJ235155_Amanita_rubescens]                                    857
FJ235155_Amanita_rub CATACACAGTTGAATGTCTATAGAATGAAATTGTAGGCTTTTGTCAGCCTTTTAAATGATAA---AATACAACTTTCAACAACGGATCTCTTGGTTCTCGCATCGATGAAGAACGCAGCGAAAT
HM189768_Hydnum_elli GGATGTTTATC----------------------------------TGCCGCAAGGCAAAATTTTTAATACAACTTTTAACAACGGATCTCTTGGCTCTCGCATCGATGAAGAACGCAGCGAAAT
JN367310_Ustilago_wi GATTTTGTATGTCATTTGATTGAAACGACTGG-------------TAATGCGGTCGTCTAATCTAAAAACAACTTTTGGCAACGGATCTCTTGGTTCTCCCATCGATGAAGAACGCAGCGAATT
HQ700340_Puccinia_ac ACAAACACAAATGAAATTTAAGAATGTAA----------------AACAATTGATTATTTTT---GAAATAACTTTTAACAATGGATCTCTAGGCTCTCATATCGATGAAGAACACAGTGAAAT
JF439504_Peniophora_ TAACATACCCCAATGAAGTATCAGAATGT----------------AACTTGCGTTAACTCGCACAAATATAACTTTCAACAACGGATCTCTTGGCTCTCGCATCGATGAAGAACGCAGCGAAAT
FR821312_Leptodiscel ---------------GTTGATGAACGTCT----------------GATTCAACTTTTAAAAA---AATACAACTTTTAACAATGGATCTCTTGGCTCTGGCATCGATGAAGAACGCAGCGAAAT
HE804179_Claviceps_p GCGGCATGTCTGAGTG-----------------------------GATTTAAAAACAAATGA---ATCAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAAT
HE575668_Torulaspora TAGTCAAAAACCGATTTC---------------------------GTTTTAGAAATTAAAAAATATTCAAAACTTTCAACAACGGATCTCTTGGTTCTCGCATCGATGAAGAACGCAGCGAAAT
FR822391_Paecilomyce AGCGGGCGGAATTTCTTCTCTGAGTTGCA----------------CAAGCAAAAACAAATGA---ATCAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAAT
JN232196_Fusarium_ox --------------TATATGTAACTTCTG----------------AGTAAAACCATAAATAA---ATCAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCAAAAT
HE775375_Claroideogl -----CATTTATAAAATATATTATATCTT----------------GCATATAATAATAATTA---AAAACAACTTTCAACAACGGATCTCTTGGCTCTCGCATCGATGAAGAACGCAGCGAATT
HE775357_Acaulospora CCATCCTATTAAAATATAATTTTTTAAAT----------------TTTATTAATAAAAAATA---AAGACAACTTTCAACAACGGATCTCTTGGCTCTTGCATCGATGAAGAACGCAGTGAAAT
HE775359_Diversispor -------------AAATATATCAATTATA----------------TAAAACAAATATAAAAA---AAGAAAACTTTCAACAACGGATCTCTTGGCTCTCGCATCGATGGAGAACGCAGCGAAAT
HE775351_Paraglomus_ AACCAAACCCGCTCGTGTATTTGGATCAACCAAAAAAAACAAACAAAATTTTAACGATAATAATTAAACAAACTTTCAACAACGGATCTCTTGGCTCTCGCATCGATGAAGAACGCAGTGAAAT
HE775335_Rhizophagus -ATTTAAAACCCCACTCCTAAAAAAAAAA----------------AATTTTTTCATATAAAAATAAAGATCACTTTCAACAACGGATCTCTTGGCTCTCGCATCGATAAAAAACGTAGCGAAGT
FJ822974_Chytridium_ AAATTTTTATGTATCTTGTCTT-----------------------AATTATTAGAGAACTAATGAGAAACAACTTTAGACAACGGATCTCTTGGCTCTCGCAACGATGAAGAACGCAGCGAAAT
FJ822973_Polyphlycti GGGCTAGCCCTTGATATTAAT------------------------TGAAACATATTAAATCGATGAAAACAACTTTTGACAACGGATCTCTTGGCTCTCGCAACGATGAAGAACGTAGCGAATT
EU520518_Rhizophlyct GATTTTGTATGTGTTA-----------------------------ATCACTTTGATTAATTAAACTTAATAACTTTTGACAACGGATCTCTTGGCTCTCGCAACGATGAAGAACGCAGCGAAAT
JN943060_Spizellomyc AAACCTTTTTTTAATTTTTATAAAACTAA----------------CATTCTTTGTATTAATGTACAAAACAACTTTTGGCAACGGATCTCTTGGCTCTCGCAACGATGAAGAACGCAGCGAAAT
JN943058_Gaertneriom TGTTATTTATTTAATTTGTCTGAAACAAA----------------AGCATTTGTAGTAATATACAAAAACAACTTTTGGCAACGGATCTCTTGGTTCTCGCAACGATGAAGAACGCAGCGAAAT
JN416510_Basidiobolu AAACTAATGTCTCATCTAATTGACTGAAA----------------TTCATTTTACAAATGAA---AATACAACTTTTAACAATGGATCTCTTGGCTCTTGCATCGATGAAGAACGCAGCGAATT
HQ876465_Rhizopus_sp GAATTC---------------------------------------AGAATTGAAAGTATAATATAATAACAACTTTTAACAATGGATCTCTTGGTTCTCGCATCGATGAAGAACGTAGCAAAGT
HM104220_Lichtheimia ---------------------------------------------TTCTGTAAAAAGAGTCTTTAAAAACAACTCTTGGCAATGGATCTCTTGGTTCTCGCATCGATGAAGAGCGTAGCAAAGT
JN943049_Gongronella AATTCAATATACTTTATGAAATTCTTGAT----------------TGATTCTTAATTGAAACAAAAAAACAACTTTCAACAATGGATCTCTCGGCTTACGCATCGATGAAGAACGCAGCGAATC
JN943047_Mucor_circi AAAAGAATTCAGAATTAATATTGTAACAT----------------AGACCTAAAAAATCTAT---AAAACAACTTTTAACAACGGATCTCTTGGTTCTCGCATCGATGAAGAACGTAGCAAAGT
```

Figure 1. An ITS alignment featuring five random species each of the fungal phyla *Ascomycota, Basidiomycota, Glomeromycota, Chytridiomycota,* and *Zygomycota* s.l. The left half of the screen represents the ITS1 and the right half the 5.8S. Whereas the ITS1 alignment appears more or less chaotic, the 5.8S stands out as a very conserved element throughout these five phyla. The 5.8S starts at position 803 (indicated by the black cursor in the uppermost sequence). Seaview (Gouy et al. 2010) was used to display the alignment.

sure that all sequences in the alignment indeed are ITS sequences. (Strictly speaking they need not be fungal ITS sequences however; oomycete, metazoan, and plant ITS sequences are sometimes retrieved with so-called "fungus-specific" ITS primers (e.g., Tedersoo et al. 2010). The process of verifying hypothesized taxonomic affiliations is discussed in Guideline 5.)

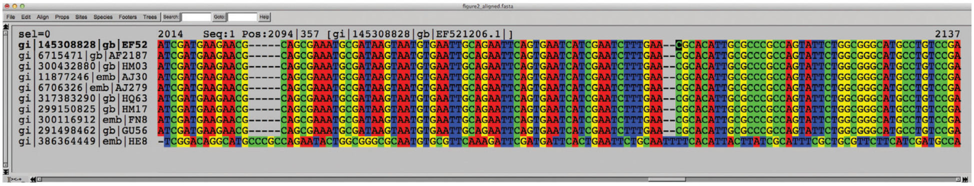
Sequences that do not produce any noteworthy similarity to the 5.8S region of the alignment are likely to belong to one of four categories: 1) they may be partial ITS sequences, containing nothing, or very little, of the 5.8S; 2) they may represent genes or markers other than the ITS (comprising, for example, the 3' SSU intron); 3) they may be of very low read quality or even feature random sequence data altogether; and 4) they may be reverse complementary. The case of reverse complementary sequences is handled separately below (Guideline 2); for the other three - and for the few fungi with truly divergent 5.8S/ITS region sequences, such as *Cantharellus* and *Tulasnella* (Feibelman et al. 1994; Taylor and McCormick 2008) - a simple manual NCBI-BLAST search in INSD is likely to reveal the nature of the complication. The user is advised to pay attention to any sequences for which the 5.8S cannot be located, and not to make scientific use of those sequences until their nature has been clarified.

As an alternative to the alignment-based approach, the user may choose to subject the query sequences - individually or, more likely, in batches - to BLAST searches in INSD. Whether or not a sequence is an ITS sequence can usually be inferred from the annotation of the top five matches alone. As a rule of thumb, a high-quality fungal ITS sequence that features the full 5.8S gene will always produce at least 100 ITS-related BLAST (blastn) matches of a bitscore of about 200 or greater (if only to the 5.8S itself) in INSD under default settings. A sequence that, in contrast, produces just a handful of matches most certainly requires further scrutiny and is, in our experience, very unlikely to qualify as a high-quality ITS sequence in the end.

Guideline 2. A single alignment step can assess the orientation of the query sequences

While it perhaps would seem natural to assume that all newly generated sequences come in the correct (5' to 3') orientation, this is in practice not always the case. A study by Nilsson et al. (2011b) showed that about 1% of the fungal ITS sequences in INSD in fact were given backwards and with all purines and pyrimidines transposed (e.g., ...TAGC... instead of the correct ...GCTA...), that is, they are reverse complementary. Whereas some software tools account for the presence of reverse complementary entries - notably the sequence similarity search engine BLAST - most tools for, e.g., multiple alignment and sequence clustering do not, at least not by default. Reverse complementary sequences can become a tangible problem when sequences are downloaded from sequence databases for use in, e.g., phylogenetic inference or diversity assessments. If the user recognizes the disparate nature of these entries - which the user is likely to do when viewing a multiple alignment but not when working with sequence clustering - the problem is easy to fix through any of a number of web services for sequence reorientation (e.g., http://www.bioinformatics.org/sms/rev_comp.html). However, if the user does not recognize these entries as problematic, they are certain to introduce significant noise into the study.

It would seem likely that most reverse complementary sequences are produced during the contig assembly, a semi-to-fully-automated step where the sequence data produced by each primer employed are brought together to form the full sequence – a contig (cf. Miller and Powell 1994). Whereas the assembly software - such as Sequencher (GeneCodes Corp., Ann Arbor, MI, USA) - usually get sequence orientation and general assembly right, the user sometimes has to step in and provide assistance. Failure of man or machine to account for the read direction of the individual primers may lead to sequence data in the reverse complementary orientation (suggesting that it may be a good idea to add the name of the primer to the name of each primer read to facilitate manual identification of mistakes). Fortunately, the process of establishing read orientation for a set of newly generated ITS sequences is straightforward. A multiple alignment of all query sequences as outlined under Guideline 1, preferably ordered by sequence similarity, is normally enough. By locating the 5.8S gene in that alignment, the user will quickly find any entries that do not seem to contain the 5.8S (Figure 2). By reorienting those seemingly anomalous entries and re-running the multiple alignment step, the user will find out whether any of the sequences in fact were reverse complementary initially. In locating the 5.8S, the user should make sure to check for the characteristic 5' end of the gene (CAACTTTC... or various minor variations thereof in nearly all fungi; see Figure 1 or Hibbett et al. (1995)). Verifying the presence of the 5' end is a necessary precaution against the (unlikely) case that most or all sequences in the alignment in fact are reverse complementary (in the former case, the correctly oriented sequences would be in the minority and appear "anomalous" at the end of the alignment). Excluding the time it takes for the server to compute the multiple alignment, the time consumption of this step is very small - even for large datasets it should be less than five minutes.



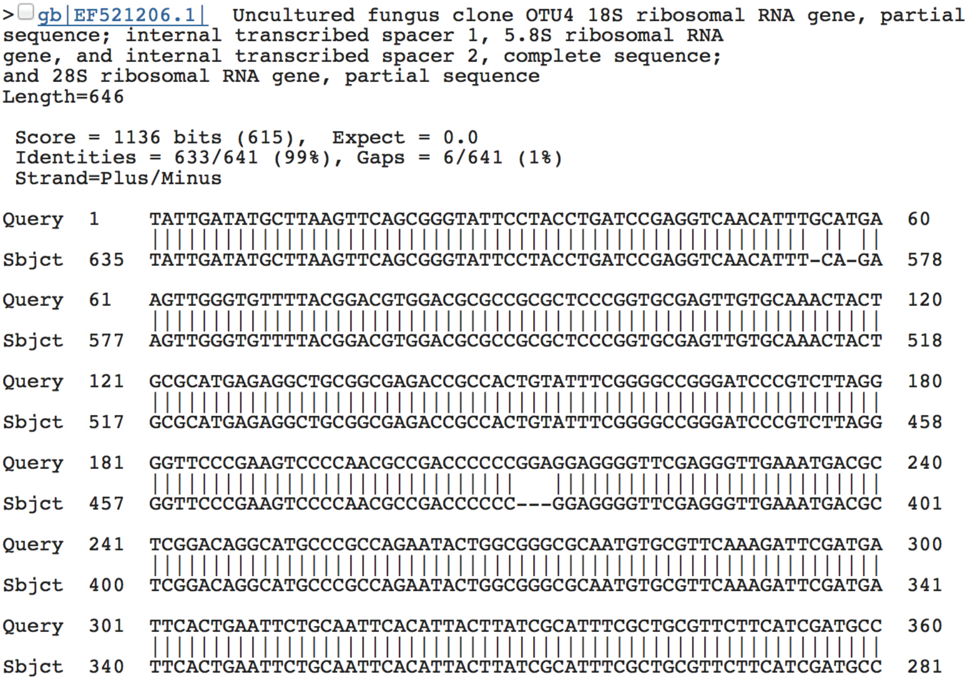
```
sel=0                  2014   Seq:1 Pos:2094|357 [gi|145308828|gb|EF521206.1|]                                             2137
gi|145308828|gb|EF52   ATCGATGAAGAACG-----CAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAA--CGCACATTGCGCCCGCCAGTATTCTGGCGGGCATGCCTGTCCGA
gi|6715471|gb|AF2187   ATCGATGAAGAACG-----CAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAA--CGCACATTGCGCCCGCCAGTATTCTGGCGGGCATGCCTGTCCGA
gi|300432880|gb|HM03   ATCGATGAAGAACG-----CAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAA--CGCACATTGCGCCCGCCAGTATTCTGGCGGGCATGCCTGTCCGA
gi|11877246|emb|AJ30   ATCGATGAAGAACG-----CAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAA--CGCACATTGCGCCCGCCAGTATTCTGGCGGGCATGCCTGTCCGA
gi|6706326|emb|AJ279   ATCGATGAAGAACG-----CAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAA--CGCACATTGCGCCCGCCAGTATTCTGGCGGGCATGCCTGTCCGA
gi|317383290|gb|HQ63   ATCGATGAAGAACG-----CAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAA--CGCACATTGCGCCCGCCAGTATTCTGGCGGGCATGCCTGTCCGA
gi|299150825|gb|HM17   ATCGATGAAGAACG-----CAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAA--CGCACATTGCGCCCGCCAGTATTCTGGCGGGCATGCCTGTCCGA
gi|300116912|emb|FN8   ATCGATGAAGAACG-----CAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAA--CGCACATTGCGCCCGCCAGTATTCTGGCGGGCATGCCTGTCCGA
gi|291498462|gb|GU56   ATCGATGAAGAACG-----CAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAA--CGCACATTGCGCCCGCCAGTATTCTGGCGGGCATGCCTGTCCGA
gi|386364449|emb|HE8   -TCGGACAGGCATGCCCGCCAGAATACTGGCGGGCGCAATGTGCGTTCAAAGATTCGATGATTCACTGAATTCTGCAATTTTCACATTACTTATCGCATTTCGCTGCGTTCTTCATCGATGCCA
```

Figure 2. A reverse complementary sequence (bottom) aligned to its nine best BLAST matches, all of which were nearly identical to the query sequence based on BLAST scores, and all of which were given in the correct orientation by their respective authors.

An alternative, and perhaps less advisable, approach to reverse complementary control involves BLAST in INSD. By default, BLAST offers native support for reverse complementary queries (as well as reference sequences) and makes very little noise if a reverse complementary sequence is found. In fact, the user has to scroll down several pages of BLAST output - to the actual alignment produced by BLAST - to get an idea of whether a query sequence is reverse complementary or not. Here, the item “Strand=Plus/Plus” indicates that both the query and the reference sequence are in the same read direction. If the five to ten best matches are all “Strand=Plus/Plus” (and particularly if they come from two or more different studies), the user can be reasonably certain that the query sequence is given in the correct orientation. Similarly, several consecutive “Strand=Plus/Minus“ suggest that the query sequence is reverse complementary (Figure 3). Problematically, but logically, a reverse complementary sequence in INSD will produce a “Strand=Plus/Plus” BLAST result to a reverse complementary query, with the second match hopefully showing “Strand=Plus/Minus“. In other words, based on the BLAST output alone it is not always easy to conclude which sequence is reverse complementary and which is given in the regular orientation. Indeed, the hypothetical existence of large batches of reverse complementary INSD sequences for some particular species would interfere with the above observations, suggesting that the best way to approach reverse complementary control is by looking at the actual sequence data in a multiple alignment. A special case of reverse complementary sequences - the reverse complementary chimera - is treated under Guideline 4 below.

Guideline 3. PCR chimeras tend to lack full counterparts in the sequence databases and are therefore usually easy to spot through BLAST

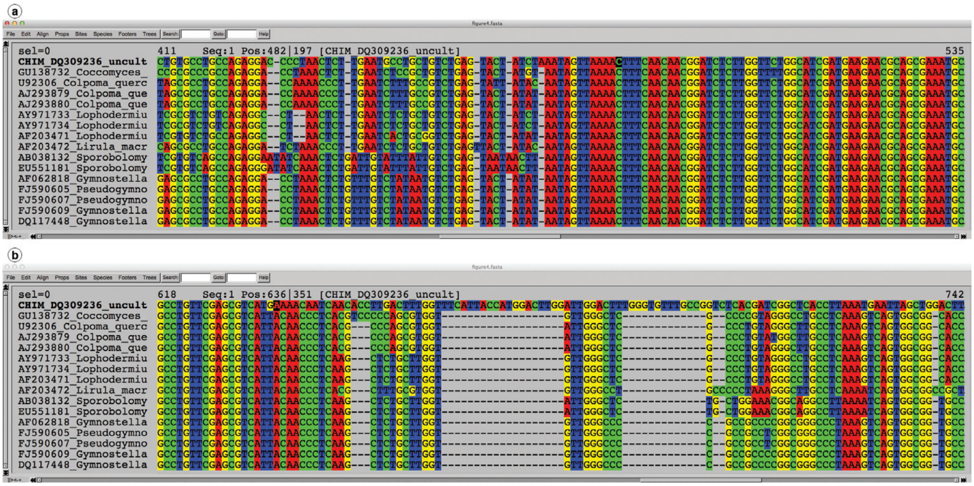
The traditional view of a PCR chimera is an artificial sequence resulting from the joining of two (or occasionally more) sequence fragments that do not originate from the same species (see Guideline 4 for a wider definition). In a typical fungal ITS chimera, either the ITS1 comes from one species and 5.8S plus ITS2 come from another, or ITS1 plus 5.8S come from one species and ITS2 from another (Figure 4). In other words, the chimeric breakpoint often seems to be located in the first – and more conserved - part of the 5.8S. These traditional chimeras can unintentionally be



```
>□gb|EF521206.1|  Uncultured fungus clone OTU4 18S ribosomal RNA gene, partial
sequence; internal transcribed spacer 1, 5.8S ribosomal RNA
gene, and internal transcribed spacer 2, complete sequence;
and 28S ribosomal RNA gene, partial sequence
Length=646

 Score = 1136 bits (615),  Expect = 0.0
 Identities = 633/641 (99%), Gaps = 6/641 (1%)
 Strand=Plus/Minus

Query  1    TATTGATATGCTTAAGTTCAGCGGGTATTCCTACCTGATCCGAGGTCAACATTTGCATGA  60
            |||||||||||||||||||||||||||||||||||||||||||||||||||||| || ||
Sbjct  635  TATTGATATGCTTAAGTTCAGCGGGTATTCCTACCTGATCCGAGGTCAACATTT-CA-GA  578

Query  61   AGTTGGGTGTTTTACGGACGTGGACGCGCCGCGCTCCCGGTGCGAGTTGTGCAAACTACT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  577  AGTTGGGTGTTTTACGGACGTGGACGCGCCGCGCTCCCGGTGCGAGTTGTGCAAACTACT  518

Query  121  GCGCATGAGAGGCTGCGGCGAGACCGCCACTGTATTTCGGGGCCGGGATCCCGTCTTAGG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  517  GCGCATGAGAGGCTGCGGCGAGACCGCCACTGTATTTCGGGGCCGGGATCCCGTCTTAGG  458

Query  181  GGTTCCCGAAGTCCCCAACGCCGACCCCCCGGAGGAGGGGTTCGAGGGTTGAAATGACGC  240
            ||||||||||||||||||||||||||||||   |||||||||||||||||||||||||||
Sbjct  457  GGTTCCCGAAGTCCCCAACGCCGACCCCCC---GGAGGGGTTCGAGGGTTGAAATGACGC  401

Query  241  TCGGACAGGCATGCCCGCCAGAATACTGGCGGGCGCAATGTGCGTTCAAAGATTCGATGA  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  400  TCGGACAGGCATGCCCGCCAGAATACTGGCGGGCGCAATGTGCGTTCAAAGATTCGATGA  341

Query  301  TTCACTGAATTCTGCAATTCACATTACTTATCGCATTTCGCTGCGTTCTTCATCGATGCC  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  340  TTCACTGAATTCTGCAATTCACATTACTTATCGCATTTCGCTGCGTTCTTCATCGATGCC  281
```

Figure 3. "Strand=Plus/Minus" indicates that the query and reference sequence come in opposing read directions. Another hint comes from the observation that the alignment starts at the first base (1) in the query sequence and progresses upwards to base 60 in the first alignment line; however, for the reference sequence, the alignment starts at base 635 and progresses downwards to base 578.

produced in the PCR step when the DNA of two or more species are present and when the gene or marker in question features a highly conserved segment (here the 5.8S; cf. Fonseca et al. 2012). If the conserved segment in the extending strand is similar enough to the corresponding segment in the contaminant species, this strand can re-anneal to the contaminant DNA instead, with a chimeric sequence as the result. (The risk of producing chimeras in mixed-template PCRs can be reduced by optimizing the PCR protocol, see Wang and Wang 1997 and Qiu et al. 2001.) Chimeras form a particularly treacherous class of compromised sequences, because at a cursory glance they often seem like perfectly fine ITS sequences: all of ITS1, 5.8S, and ITS2 are typically present in their full length and in the expected order. One of the two underlying species dominates the sequence by comprising the ITS1+5.8S or 5.8S+ITS2, and it is the dominant species that tends to prevail in BLAST searches. The scientific (Latin) name given to a chimeric sequence is wrong by definition, but the name is particularly troublesome in cases where the dominant species formed the contaminant (non-targeted) species initially. Such sequences invite BLAST-based misannotations, often spanning fungal orders or even phyla (cf. Hugenholtz and Huber 2003). Chimeric sequences without species names (e.g., "Unidentified fungus")


a
figure4.fasta
File Edit Align Props Sites Species Footers Trees Search: Goto: Help
b
figure4.fasta
File Edit Align Props Sites Species Footers Trees Search: Goto: Help

```
sel=0                 411     Seq:1 Pos:482|197 [CHIM_DQ309236_uncult]                535
CHIM_DQ309236_uncult  CTGTGCCTGCCAGAGGAC-CCCTAACTCT-TGAATGCCTGCTGTCTGAG-TACT-ATCTAAATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
GU138732_Coccomyces_  CCGCGCCCGCCAGAGGA--CCTAAACTCT-TGAATCTCCGCTGTCTGAG-TACT-ATGT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTTTGGCATCGATGAAGAACGCAGCGAAATGC
U92306_Colpoma_querc  TAGCGCCTGCCAGAGGA--CCAAAACCCT-TGAATCTTTGCCGTCTGAG-TATT-ATAC-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
AJ293879_Colpoma_que  TAGCGCCTGCCAGAGGA--CCAAAACCCT-TGAATCTTTGCCGTCTGAG-TACT-ATAT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
AJ293880_Colpoma_que  TAGCGCCTGCCAGAGGA--CCAAAACCCT-TGAATCTTTGCCGTCTGAG-TACT-ATAT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
AY971733_Lophodermiu  TCGCGTCTGTCAGAGGC--CT--AACTCT-TGAATCTCTGCTGTCTGAG-TACT-ATCT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
AY971734_Lophodermiu  TCGCGTCTGTCAGAGGC--CT--AACTCT-TGAATCTCTGCTGTCTGAG-TACT-ATCT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
AF203471_Lophodermiu  TCGTGTCTGCCAGAGGC--CT--AACTCT-TGAATCACTGCGGTCTGAG-TACT-ATAT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
AF203472_Lirula_macr  CAGCGCCTGCCAGAGGA--TCTAAACCCT-TGAATCTCTGCTGTCTGAGTTACT-ATAC-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
AB038132_Sporobolomy  TCGTGTCAGCCAGAGGAATATCAAACTCTGATTGTATTTATTGTCTGAG-TAATAACTT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
EU551181_Sporobolomy  TCGTGTCAGCCAGAGGGATATCAAACTCTGATTGTATTTATTGTCTGAG-TAATAACTT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
AF062818_Gymnostella  GAGCGCCTGCCAGAGGA--CCTAAACTCTGTTTGTCTATAATGTCTGAG-TACT-ATAT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
FJ590605_Pseudogymno  GAGCGCCTGCCAGAGGA--CCTAAACTCTGTTTGTCTATAATGTCTGAG-TACT-ATAT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
FJ590607_Pseudogymno  GAGCGCCTGCCAGAGGA--CCTAAACTCTGTTTGTCTATAATGTCTGAG-TACT-ATAT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
FJ590609_Gymnostella  GAGCGCCTGCCAGAGGA--CCTAAACTCTGTTTGTCTATAATGTCTGAG-TACT-ATAT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
DQ117448_Gymnostella  GAGCGCCTGCCAGAGGA--CCTAAACTCTGTTTGTCTATAATGTCTGAG-TACT-ATAT-AATAGTTAAAACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGC
```

```
sel=0                 618     Seq:1 Pos:636|351 [CHIM_DQ309236_uncult]                742
CHIM_DQ309236_uncult  GCCTGTTCGAGCGTCATGAAAACAATCAACACCTTGACTTTGGTTTCATTACCATGGACTTGGATTGGACTTTGGGTGTTTGCCGGTCTCACGATCGGCTCACCTTAAATGAATTAGCTGGACTT
GU138732_Coccomyces_  GCCTGTTCGAGCGTCATTACAACCCTCACGTCCCCCAGCGTGGT-------------------GTTGGGCTC-------------G--CCCCGTAGGGCCTGCCTCAAAGTCAGTGGCGG-CACC
U92306_Colpoma_querc  GCCTGTTCGAGCGTCATTACAACCCTCACG---CCCAGCGTGGT-------------------ATTGGGCTC-------------G--CCCTGTAGGGCTTGCCTCAAAGTCAGTGGCGG-CACC
AJ293879_Colpoma_que  GCCTGTTCGAGCGTCATTACAACCCTCACG---CCCAGCGTGGT-------------------ATTGGGCTC-------------G--CCCTGTATGGCTTGCCTCAAAGTCAGTGGCGG-CACC
AJ293880_Colpoma_que  GCCTGTTCGAGCGTCATTACAACCCTCACG---CCCAGCGTGGT-------------------ATTGGGCTC-------------G--CCCTGTAGGGCTTGCCTCAAAGTCAGTGGCGG-CACC
AY971733_Lophodermiu  GCCTGTTCGAGCGTCATTACAACCCTCAAG---CTCTGCTTGGT-------------------GTTGGGCTC-------------G--CCCTGTAGGGCCTGCCTCAAAGTCAGTGGCGG-CACC
AY971734_Lophodermiu  GCCTGTTCGAGCGTCATTACAACCCTCAAG---CTCTGCTTGGT-------------------GTTGGGCTC-------------G--CCCTGTAGGGCCTGCCTCAAAGTCAGTGGCGG-CACC
AF203471_Lophodermiu  GCCTGTTCGAGCGTCATTACAACCCTCAAG---CTCTGCTTGGT-------------------GTTGGGCTC-------------G--CCCTGTAGGGCCTGCCTCAAAGTCAGTGGCGG-CACC
AF203472_Lirula_macr  GCCTGTTCGAGCGTCATTACAACCCTCACG---CTTTGCGTGGT-------------------GTTGGGCCT-------------GCCCCCTAAAGGGCTTGCCTCAAAATCAGTGGCGGCCGCT
AB038132_Sporobolomy  GCCTGTTCGAGCGTCATTACAACCCTCAAG---CTCTGCTTGGT-------------------ATTGGGCTC-------------TG-CTGGAAACGGCAGGCCTTAAAATCAGTGGCGG-TGCC
EU551181_Sporobolomy  GCCTGTTCGAGCGTCATTACAACCCTCAAG---CTCTGCTTGGT-------------------ATTGGGCTC-------------TG-CTGGAAACGGCAGGCCTTAAAATCAGTGGCGG-TGCC
AF062818_Gymnostella  GCCTGTTCGAGCGTCATTACAACCCTCAAG---CTCTGCTTGGT-------------------GTTGGGCCC-------------C--GCCGCCCCGGCGGGCCCTAAAGTCAGTGGCGG-TGCC
FJ590605_Pseudogymno  GCCTGTTCGAGCGTCATTACAACCCTCAAG---CTCTGCTTGGT-------------------GTTGGGCCC-------------C--GCCGCCTCGGCGGGCCCTAAAGTCAGTGGCGG-TGCC
FJ590607_Pseudogymno  GCCTGTTCGAGCGTCATTACAACCCTCAAG---CTCTGCTTGGT-------------------GTTGGGCCC-------------C--GCCGCCTCGGCGGGCCCTAAAGTCAGTGGCGG-TGCC
FJ590609_Gymnostella  GCCTGTTCGAGCGTCATTACAACCCTCAAG---CTCTGCTTGGT-------------------GTTGGGCCC-------------C--GCCGCCCCGGCGGGCCCTAAAGTCAGTGGCGG-TGCC
DQ117448_Gymnostella  GCCTGTTCGAGCGTCATTACAACCCTCAAG---CTCTGCTTGGT-------------------GTTGGGCCC-------------C--GCCGCCCCGGCGGGCCCTAAAGTCAGTGGCGG-TGCC
```

Figure 4. A multiple alignment where the topmost sequence is chimeric and the remaining sequences represent its best BLAST matches. The alignment is fine in ITS1 and 5.8S (**a**; the 5.8S starts at position 479), but the alignment in ITS2 (**b**; position 637 and on) falls far short of scientific rigour. Alignments like these bespeak chimeric unions.

are perhaps less of a problem to molecular species identification, but like all chimeras they inflate diversity assessments such as sequence/OTU richness, estimated richness, and phylogenetic diversity measures (in the latter case for the reason that chimeric sequences tend to form long branches; cf. Tedersoo et al. 2011). Chimeras may however also be detrimental to endeavours other than diversity assessment, for example through skewing multiple alignments.

UNITE has a record of about 1,000 chimeric fungal ITS sequences in the public corpus, corresponding to 0.4% of the number of such sequences. The real number of chimeras is probably significantly higher, since chimeras between closely related species are much more difficult to find than chimeras between distantly related ones. The vast majority of the 1,000 known chimeras are of the "distantly related" type; the chimera in Figure 4 is such an example. Cloning of PCR amplicons is a component in many studies in which chimeras were subsequently reported, suggesting that studies employing cloning should be particularly vigilant against chimeric unions. Fortunately, finding at least bad cases of chimeras in newly generated datasets is fairly straightforward. The solution draws from the observation that chimeric sequences tend to be unique in datasets of small to moderate sizes, i.e., that any given illegitimate union of sequence fragments happened only once in the study. This somewhat rough approximation means that the user can cluster the query dataset at approximately the species level (97-98% similarity, 90% sequence coverage; see above) and then focus on the singletons (or all small-sized OTUs) only. By subjecting the singleton sequences to BLAST searches and keeping an eye on the graphical summary of the BLAST hits provided by NCBI-BLAST (http://blast.ncbi.nlm.nih.gov/), the user will be able to

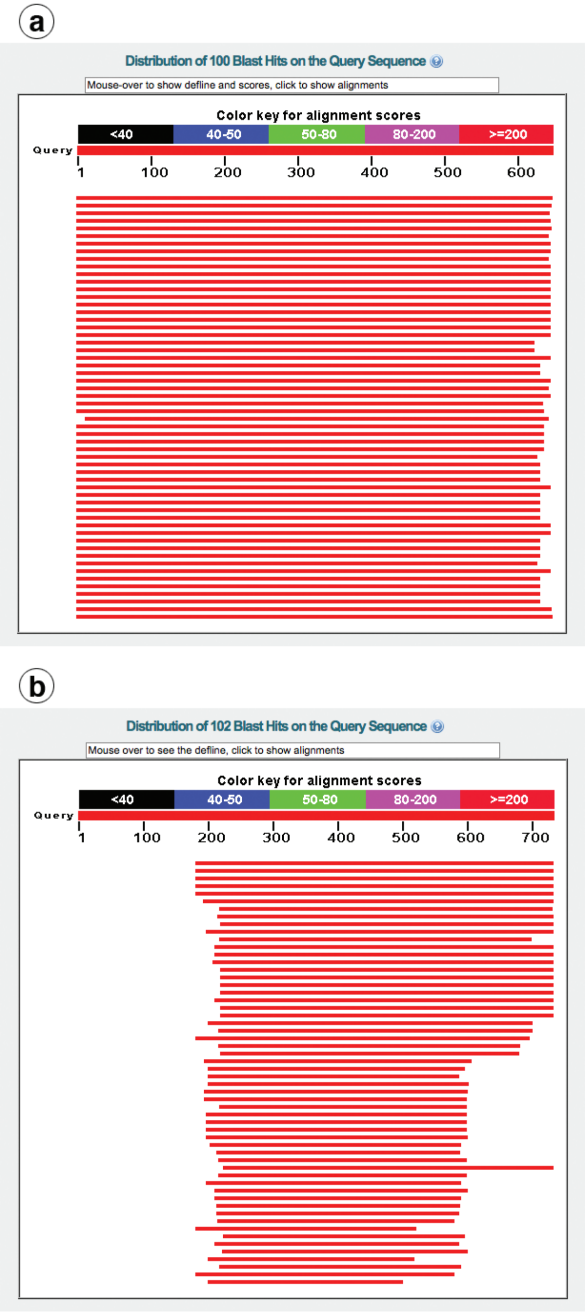

a
Distribution of 100 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
100
200
300
400
500
600
b
Distribution of 102 Blast Hits on the Query Sequence
Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
100
200
300
400
500
600
700

Figure 5. a Graphical overview of the BLAST results of a regular sequence **b** BLAST results of a chimeric sequence where the ITS1 comes from another species, such that the ITS1 is not involved in the alignment featuring the 5.8S+ITS2 (hence the lack of a match for the first ca. 180 bp.). Obviously, a severely compromised sequence that is already in INSD will always find a perfect match through BLAST in INSD: itself. In that case, the presence of a 100% similar reference sequence cannot be used as a testimony to the authenticity of the query sequence.

identify sequences in need of further scrutiny. Figure 5a shows a BLAST run where a query sequence was well matched across its full length by the topmost hits. Figure 5b, in contrast, shows a chimeric sequence where the 5.8S and ITS2 were well matched by the topmost hit, whereas the ITS1 could not be aligned at all to it. This corresponds to the case where the ITS1 comes from a distantly related species with respect to the remainder of the sequence. All cases where ITS1+5.8S - or 5.8S+ITS2 - produce nearly perfect matches, whereas ITS2 or ITS1, respectively, produces an unexpectedly poor match, call for closer scrutiny.

In the case of Figure 5b, it is the ITS1 that does not harmonize with the remainder of the sequence. Doing a BLAST search based on ITS1 alone shows that it is a polypore (100% similarity); the 5.8S+ITS2 BLAST, in contrast, shows that those parts belong to an agaric (100% similarity). By doing separate BLAST searches like this, the user will come fairly close to practical proof that the sequence in question is chimeric. Such sequences should be pruned from the query dataset, and they should similarly not be submitted to the sequence databases. However, the user should keep in mind that legitimate query sequences - particularly long ones - can also produce BLAST results similar to that in Figure 5b for the reason that the most similar reference sequences were much shorter due to, e.g., primer choice. The BLAST alignment indicates at what base in the query and the reference sequence the alignment starts. For example, if the alignment start is “1” in the reference sequence but “350” in the query sequence, then the seemingly odd BLAST results simply reflect the absence of reference data. Introns such as the one at the 3’ end of the SSU may produce similar results. However, also in these situations, subjecting the non-matching part of the query sequence to a BLAST search is likely to reveal the nature of the problem.

Problematically, not all cases of chimera detection will be as straightforward as the example in Figure 5b, and the user will sometimes face difficult decisions. After all, ITS sequence data are available for a mere 1% of the hypothesized 1.5 million extant species of fungi (Hawksworth 2001; Hibbett et al. 2011), and some newly generated sequences will be singletons, and perhaps look odd, for the reason that they have not been sequenced before, such that no fit objects of comparison are available. To routinely exclude sequences that differ from known sequences would obviously not be a good way to expand our knowledge of the fungal kingdom. The user is probably best advised to delete the sequences she feels sure are chimeric and leave the rest of the sequences in the dataset; it would still be a major improvement over not checking the dataset for chimeras at all. If these dubious sequences are of particular relevance to the study, and if there is fungal material left from which to regenerate those sequences, then the user would have the opportunity to verify the biological, or artefactual, origin of those sequences through another round of sequencing. A further complication is that in studies with great sequencing depth, more or less identical chimeras between the most common OTUs may occur more than once in the dataset. A solution to this problem could be to check a representative sequence also from OTUs that are not singletons (focusing, as needed, on all OTUs with few constituent sequences).

Guideline 4. Sequences can be broken in other, puzzling ways; BLAST, again, will tell

BLAST also has the capacity to indicate several other classes of compromised entries. Figure 6 shows an assembly chimera, which is the product of incorrect assembly of two or more sequence fragments (primer reads) into a single contig. The dotted vertical line in the reference sequences indicates a break in the alignment between these and the query sequence. The user will have to scroll down to the BLAST alignment to learn of the exact nature of the break. Often one finds that such sequences were assembled with the ITS1 and the ITS2 in the wrong order. The resulting BLAST alignment will be divided into sections, and the user might find that, e.g., base 285 to 614 in the query sequence are matched by bases 1 to 330 in the reference sequence. Bases 1-284 in the query are, however, best matched by bases 331-614 in the reference sequence; although it may not always be straightforward to see exactly what the problem is, the non-contiguous nature of these alignment segments at least makes it easy to see that there indeed seems to be a problem to begin with. If all alignment sections are in the Strand=Plus/Plus orientation, and the next few reference sequences similarly produced such sectioned alignments with respect to the query, then the user can be certain that the query sequence is an assembly chimera. It is easy to see that assembly chimeras may follow as a result of minimal overlap between the fragments under assembly and the subsequent failure of the contig software – under the settings applied - to pick the correct ends for merger. If there is no overlap at all between the fragments - such that there should have been additional sequence data between two fragments that are now joined - the corresponding BLAST results will look something like Figure 6. Such bridged sequences may also be produced inadvertently in, e.g., the phylogenetic analysis package PAUP (Swofford 2003) when the user excludes certain alignment regions from the analysis due to, e.g., poor alignability using the generic "EXCLUDE" command. If the user then exports the alignment analysed for INSD or TreeBase (Sanderson et al. 1994) deposition, the individual sequences will lack the parts excluded from the analysis and therefore qualify as chimeric. Alternatively, if an extraneous sequence segment was assembled into a position where it should not have been, such as in the middle of the 5.8S, the BLAST results tend to look similar to those shown in Figure 7. Finally, reverse complementary chimeras are produced when a sequence is assembled to contain one or more fragments in the regular orientation and one or more fragments in the reverse complementary orientation (cf. Hartmann et al. 2011). The BLAST results of such sequences often look like Figure 6, and the BLAST alignment will indicate that one or more of the sections are in the opposite direction, "Strand=Plus/Minus".

The distal (5' and 3') ends of newly generated sequences are typically of lower read quality than the interior parts of the sequence. It is the job of the contig assembly software to highlight poorly read bases clearly enough that the user can address them before the final sequence is produced from the contig. Untrimmed sequences tend to look like the one in Figure 8 when run through BLAST; note that the match does not include the first ca. 20, and the last ca. 30, base-pairs. Unless all of the reference

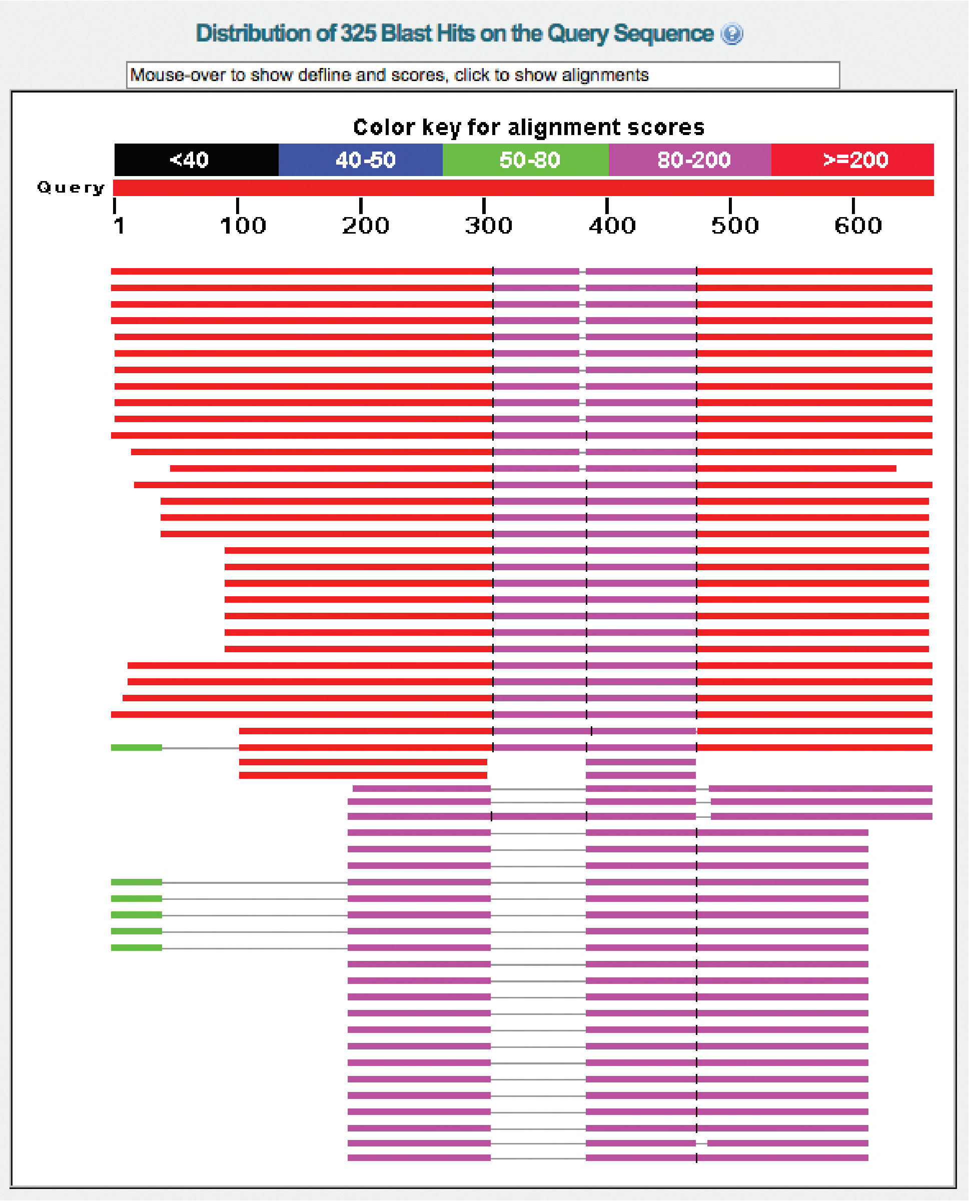

Distribution of 325 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
100
200
300
400
500
600

Figure 6. An assembly chimera. The black dashed lines indicate breaks in the BLAST alignment and should always be taken to mean that manual examination is needed.

sequences are in fact shorter than the query sequence, the user should probably re-check the chromatograms in the distal parts of the sequence - and consider trimming regions of poor quality - at this stage. Many public ITS sequences, in turn, are poorly trimmed, sometimes leaving the process of telling whether it is the query or the reference sequence that features the low-quality bases all but intractable. This speaks to the

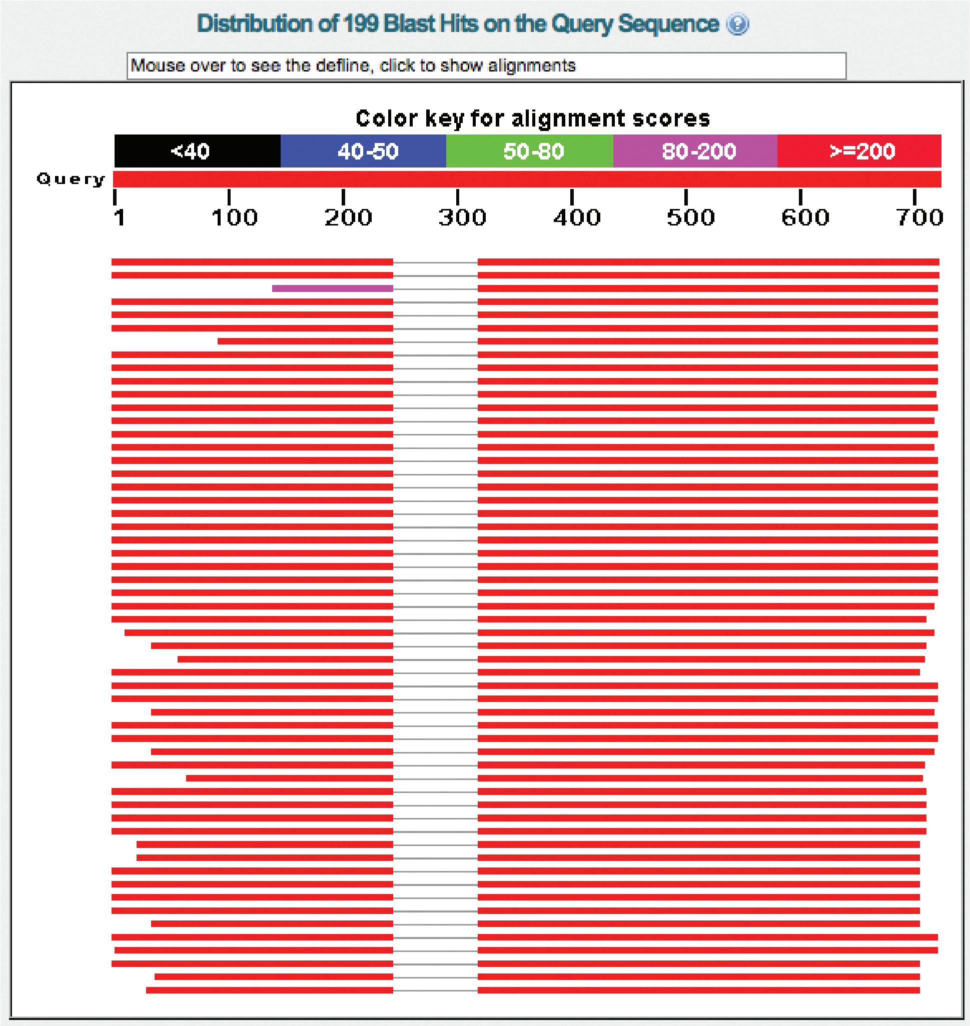

Distribution of 199 Blast Hits on the Query Sequence
Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
100
200
300
400
500
600
700

Figure 7. An assembly chimera. An extraneous sequence segment was assembled into a position where it should not have been, such as in the middle of the 5.8S. The white area in the reference sequences indicates the absence of sequence data for this particular part of the query sequence. Manual examination is always needed in cases like this.

importance of always taking the sequence assembly step seriously and of paying special attention to any region where the chromatograms appear substandard. Other newly generated sequences are of reduced read quality throughout. One obvious sign is that they may feature IUPAC DNA ambiguity symbols (e.g., N and S; Cornish-Bowden 1985). If these are scattered along the full length of the sequence, our experience is that the sequence should be discarded altogether. If they, on the other hand, are clustered in some single region of the sequence - typically at either distal end - and the chromatograms look satisfactory in the remaining regions of the sequence, then the sequence

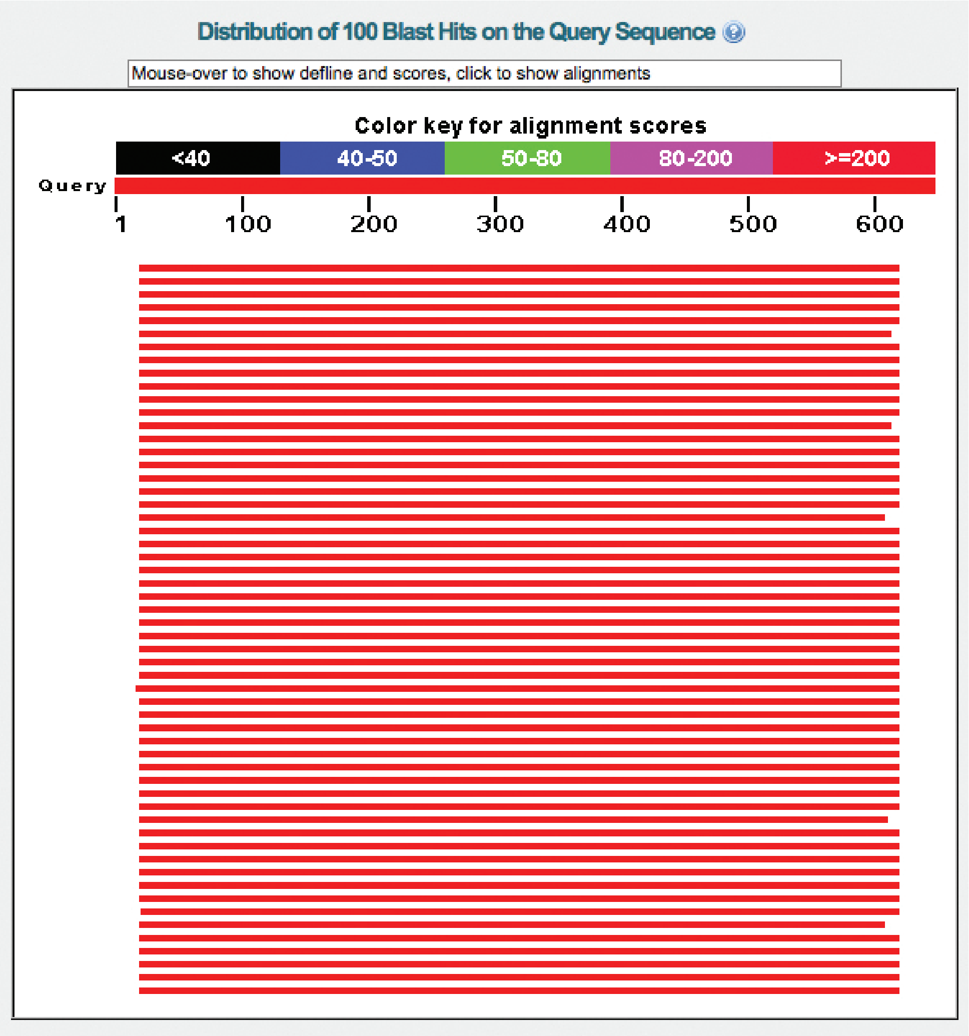

Distribution of 100 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
100
200
300
400
500
600

Figure 8. Untrimmed sequences tend to look like this when run through BLAST. Note how the first ca. 20 bp., and the last ca. 30 bp., of the query sequence (represented by the red bar with scale marks every 100 bp.) do not align to any of the BLAST hits. The use of different but closely situated primers may give a similar pattern, however, pointing at the need to also look at the BLAST alignments for start and end positions of the reference sequences.

is probably reliable (although in need of distal trimming). Another tell-tale sign may be suspiciously large homopolymer regions (e.g., ...AAAAAAAAAA...); again the user should go back to the chromatograms to scrutinize these regions. A complication is that the underlying fungal individual may have alleles of different lengths in these regions, making exact base-calling hard. Of particular difficulty are those sequences in which neither ambiguity symbols nor suspicious homopolymer regions are present, but that still are very distant from the closest BLAST hit. The BLAST alignment may offer some tentative clues here. If the mismatches are scattered more or less evenly across the full length of the query sequence, it is likely that the general sequence quality is substandard, such that the sequence should be discarded. If, on the other hand, there are no - or significantly fewer - mismatches in the region corresponding to the 5.8S in the BLAST alignment, this would suggest that the sequence is authentic, if very deviant from everything else. Indeed, several large groups of previously unknown fungi have been described in recent years (e.g., Jones et al. 2011; Rosling et al. 2011).

Guideline 5. Taxonomic annotations should be verified before the sequences are used

About half of the 250,000 public, full-length fungal ITS sequences are annotated to the level of species (Hibbett et al. 2011). Several studies have, however, shown that the taxonomic reliability of the entries in the public sequence databases has yet to reach perfection, and more than 10% of the public fungal ITS sequences that carry a species name may in fact carry an incorrect species name (e.g., Nilsson et al. 2006). It is easy to see that morphology-based species identification procedures sometimes go wrong among closely related or otherwise highly similar species, and these misidentifications would then carry over to the taxonomic annotation of the sequence generated from the specimen. Many of the misidentifications we have come across, however, span orders, classes, and frequently also phyla of the fungal tree of life. Indeed, more than 20 fungus-related cases of misidentification at the kingdom level are indexed in UNITE. This suggests that taxonomic competence is only one of several processes leading to incorrect taxonomic annotations of public sequences. Unintended sequencing of epifungal - or intrasporocarp - parasites, mutualists, or commensalists appears common, for example. PCR contaminations and the mixing up of test tubes, computer files, and labels stand out as other major sources of error. Incorrectly identified or contaminated cultures – even in the major international culture collections – form an additional, serious concern. The conclusion is obvious: nobody - regardless of degree of taxonomic competence - should by default assume that their taxonomic annotations are correct and not in need of verification.

We take the position that all sequences in a newly generated dataset should be verified for taxonomic affiliation, even if they are annotated only to kingdom level (e.g., "Uncultured fungus"). The process of verifying a hypothesized taxonomic annotation - or at least ruling out the possibility that the annotation is way off - is usually trivial and amounts to a simple BLAST run. A sequence annotated as *Penicillium* is expected to hit other *Penicillium* sequences (usually in a chaotic list of anamorphic and teleomorphic names, species complexes, and numerous environmental sequences; a visit to Index Fungorum (http://www.indexfungorum.org/) or MycoBank (http://www.mycobank.org/) may be needed to establish the relations of the names obtained). A quick check of some degree of consistency among the top ten matches is normally enough to confirm the basic authenticity of the taxonomic affiliation, particularly if the top ten matches stem from two or more different studies. The INSD keyword "BARCODE" (specified in the description of the entry) indicates that a sequence complies with a number of quality criteria (http://barcoding.si.edu/pdf/dwg_data_standards-final.pdf) and so should be weighted as a more reliable reference sequence. However, looking at BLAST hit lists is often more difficult than one might think. The following five basic principles may be good to keep in mind. **a)** BLAST is sensitive to the length and level of sequence conservation of the query and reference sequences, and the user is advised to prune any large parts of the SSU and LSU from the ITS sequences before doing BLAST searches (cf. Kang et al. 2010). It sometimes pays off to use only the ITS1 or ITS2 for the searches. **b)** BLAST does local alignment and so will base its core statistics on the part of the query sequence it managed to align rather than the full length of the query sequence. Thus, even if a match says "100% similar", it will typically not apply to the full length of the query sequence, and confirming the proportion of the sequence aligned requires examination of the coverage statistics reported in the BLAST searches. If the user is concerned with the absolute similarity of the query sequence to the best match, a second alignment step (in, e.g., MAFFT) and a pocket calculator may be needed. **c)** In the case of identical BLAST bitscores (matches), the order of the hits is for all practical purposes uninformative. This cautions against looking only at the very topmost match; if there are several equally good matches, they are all equally relevant. **d)** The degree to which the ITS region is species specific differs among fungal lineages, as does the average distance to the closest species for any given species (Nilsson et al. 2008). It is a good idea to refrain from oversimplified approaches to species identification and sequence annotation, such as enforcing a strict 97% similarity criterion at all times. Indeed, BLAST reports on similar sequences rather than species names. **e)** The taxon sampling of fungi is still very much incomplete (Brock et al. 2009; Nilsson et al. 2011a). Thus, even if some particular query sequence does not hit any of the species the user had expected - but more remotely related ones instead - it does not have to mean that anything is wrong; it could just be a case of thin taxon sampling. The GenBank Nucleotide (http://www.ncbi.nlm.nih.gov/nuccore/) query string "Amanita[ORGN] AND 5.8S[TITL]" will show whether there are any ITS sequences annotated as belonging to the genus *Amanita* in GenBank. Such simple queries permit the user to examine and establish whether the expected species are present among the available ITS sequences in the database. Ross et al. (2008) and Ovaskainen et al. (2010) provide interesting statistics on the performance of BLAST under varying conditions, including incomplete database coverage.

It is typically simple to establish basic authenticity of the taxonomic annotations for a set of query sequences. The process described above will often take the user to the genus level or even the species level in some cases, at which stage one can rule out severe misannotation. Going all the way to actually verifying the species-level annotation is a trickier objective, and one that will not always be possible based on BLAST and the public sequence databases alone. A phylogenetic analysis of the query sequence and the 20-30 best BLAST matches (or as many as alignability allows) is a good starting point for a more robust examination of the taxonomic affiliation of the query sequence (cf. Taylor et al. 2000). The alignment/phylogenetic analysis combination may also be helpful in locating otherwise anomalous sequence data; (single) sequences that are found on unusually long branches or that do not find well-supported positions may be worth looking closer at.

Taking action on bad sequences in INSD / UNITE

Anyone using the public sequence databases to pursue low-quality entries in a newly generated dataset will sooner or later find low-quality entries also in these databases. When skimming through BLAST hit lists, for instance, one regularly sees entries whose taxonomic annotation simply has to be wrong for one reason or the other - a single *Betula* (birch) in a list of *Amanita* (fly agaric), for instance. It is easy to feel that some mistakes are so far off and absurd as to be harmless. In reality they are harmless only to a limited number of people, namely those with a relevant taxonomic background; with a reasonable insight into how BLAST operates; and with enough time on their hands to interpret their sequence similarity searches manually. Everyone else may be in harm's way. We did an informal evaluation of 20 fungal ITS sequences whose taxonomic annotation was off at the ordinal or class level by simply running the accession numbers through Google. Three of the sequences (15 %) had been used under their original (incorrect) name in at least one other scientific publication than the one through which they were released. Even taxonomic experts would be hard put to spot many such derived mistakes since they are published one level removed from the original data, suggesting a route through which errors and mistakes can be cited and re-cited enough to eventually be accepted as truths. There is thus every reason to take some form of action when one comes across a public DNA sequence associated with significant error.

Hartmann et al. (in press) discuss several ways to take action on compromised public sequences. We will assume here that the user is very pressed for time and unwilling to spend more than a minute on the matter; we also assume that the nature of the complication is severe enough to be beyond questioning or interpretation. A quick, friendly email to the original sequence authors is in fact likely to solve the problem altogether, because few scientists would presumably like their names to be associated with persistent, broken data. Sequence authors have considerable say over their entries in INSD, and a request from them to the INSD staff (e.g., http://www.ncbi.nlm.nih.gov/About/glance/contact_info.html) is unlikely to go unheeded. It is however notoriously difficult to find people and their present contact information over the web (cf. Wren 2008). Another path to take action on broken sequences is therefore to write an email to the INSD staff directly. We have found the INSD staff to be very friendly and service-minded in these matters. Several options are open to the INSD staff to deal with misidentified sequences. One recent example is to add an UNVERIFIED keyword to highly problematic sequences and exclude the sequence from BLAST, although the sequence will still be archived in INSD. Finally, it is possible to use the third-party sequence annotation feature of UNITE/PlutoF (Abarenkov et al. 2010b) to simply replace the incorrect species name with the correct one, or to mark the entry as chimeric, or to take whatever other action appropriate. Third-party annotations of sequences via PlutoF are visible to users in the European Nucleotide Archive of the INSD through a link-out function. We feel that the exact way in which the user chooses to take action is less important compared to whether or not the user chooses to take action in the first place, and we hope that the mycological community will be able to set a high standard here.

Concluding remarks

The present document brings together a set of guidelines, recommendations, and observations towards identifying severely compromised sequences before they are put to scientific use. While they were written with the non-bioinformatician in mind and aim to be non-technical and straightforward to apply, we still believe they are powerful enough to have prevented the deposition of the vast majority of severely compromised fungal ITS sequences in the public sequence databases, had they been applied at the time of sequence accessioning. Importantly, however, these guidelines would not have caught all cases of badly damaged sequence data. Thus, the application of the principles presented here will not guarantee - but rather just increase the chance - that the dataset at hand will be of reasonable standard after processing. Furthermore we would like to stress that these guidelines offer little in way of fine-grained authenticity and reliability. Misidentification among closely related species, somewhat reduced levels of general sequence read quality, and base-inflation in homopolymer regions are all examples of problems that are only partly addressed by this document. We certainly do not want our guidelines to be used as replacements for more advanced, technical solutions; we rather hope that they will be used by those who, for one reason or the other, do not have access to or would not consider running any advanced, technical solutions in the first place (e.g., Appendix).

Our guidelines come with no other software requirement than a web browser. They still require something else of the user too: a critical, inquisitive, and perhaps imaginative mind. It would seem impossible to lay down firm rules to which all high-quality sequences would comply and that all low-quality sequences would violate. Rather the user should expect to find herself in situations where the user herself is the best arbiter of what is correct and what isn't. Although such a situation would not be unfamiliar to anyone in systematics or taxonomy, we would still like to point out the importance of common sense in pursuing broken sequence data. The present authors spent considerable time trying to make this document as rich and multi-faceted as possible, but it goes without saying that additional, relevant observations and advice are to be found among the remaining members of the scientific community. We hope that anyone in the position to improve or add to the present set of guidelines will take the time and opportunity to do so. The potential outlets are many and range from the "Add comments" feature of the present journal to separate publications in this or any other journal. The ever-increasing weight assigned to molecular data in mycology - and the life sciences as a whole - suggests that any such move may have positive ramifications extending far beyond the datasets of each individual user.

Acknowledgements

RHN acknowledges financial support from Swedish Research Council of Environment, Agricultural Sciences, and Spatial Planning (FORMAS, 215-2011-498). LT and UK received financial support from ETF grants 8235, 9286, and FIBIR. EK acknowledges support from the Life Science Area of Advance at Chalmers University of Technology, the Swedish Research Council (VR), and FORMAS. CLS was supported in part by the Intramural Research Program of the NIH, National Library of Medicine. TMP acknowledges financial support from the Government of Canada through Genome Canada and the Ontario Genomics Institute through the Biomonitoring 2.0 project (OGI-050). AJ acknowledges financial support from Department of Energy's Biological and Environmental Research Program (Award #ER65000). Yann Bertrand and Chris Quince are acknowledged for valuable input on the manuscript. Support from the Gothenburg Bioinformatics Network is gratefully acknowledged.

References

Abarenkov K, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Ursing BM, Vrålstad T, Liimatainen K, Peintner U, Kõljalg U (2010a) The UNITE database for molecular identification of fungi - recent updates and future perspectives. New Phytologist 186(2): 281–285. doi: 10.1111/j.1469-8137.2009.03160.x

Abarenkov K, Tedersoo L, Nilsson RH, Vellak K, Saar I, Veldre V, Parmasto E, Prous M, Aan A, Ots M, Kurina O, Ostonen I, Jõgeva J, Halapuu S, Põldmaa K, Toots M, Truu J, Larsson K-H, Kõljalg U (2010b) PlutoF - a web-based workbench for ecological and taxonomical research, with an online implementation for fungal ITS sequences. Evolutionary Bioinformatics 6: 189–196. doi: 10.4137/EBO.S6271

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25(17): 3389–3402. doi: 10.1093/nar/25.17.3389

Anderson IC, Cairney JWG (2004) Diversity and ecology of soil fungal communities: increased understanding through the application of molecular techniques. Environmental Microbiology 6(8): 769–779. doi: 10.1111/j.1462-2920.2004.00675.x

Begerow D, Nilsson RH, Unterseher M, Maier W (2010) Current state and perspectives of fungal DNA barcoding and rapid identification procedures. Applied Microbiology and Biotechnology 87(1): 99–108. doi: 10.1007/s00253-010-2585-4

Bidartondo M, Bruns TD, Blackwell M et al. (2008) Preserving accuracy in GenBank. Science 319(5870): 1616. doi: 10.1126/science.319.5870.1616a

Blackwell M, Hibbett DS, Taylor JW, Spatafora JW (2006) Research Coordination Networks: a phylogeny for kingdom Fungi (Deep Hypha). Mycologia 98(6): 829–837. doi: 10.3852/mycologia.98.6.829

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005) Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society of London Series B – Biological Sciences 360(1462): 1935–1943. doi: 10.1098/rstb.2005.1725

Brock PM, Döring H, Bidartondo MI (2009) How to know unknown fungi: the role of a herbarium. New Phytologist 181(3): 719–724. doi: 10.1111/j.1469-8137.2008.02703.x

Cornish-Bowden A (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Research 13(9): 3021–3030. doi: 10.1093/nar/13.9.3021

Feibelman T, Bayman P, Cibula WG (1994) Length variation in the internal transcribed spacer of ribosomal DNA in chanterelles. Mycological Research 98(6): 614–618. doi: 10.1016/S0953-7562(09)80407-3

Fonseca VG, Nichols B, Lallias D, Quince C, Carvalho GR, Power DM, Creer S (2012) Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. Nucleic Acids Research 40(9): e66. doi: 10.1093/nar/gks002

Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA (2002) Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics 18(12): 1641–1649. doi: 10.1093/bioinformatics/18.12.1641

Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin J-F (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12: 245. doi: 10.1186/1471-2164-12-245

Gouy M, Guindon S, Gascuel O (2010) SeaView Version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27(2): 221–224. doi: 10.1093/molbev/msp259

Harris JD (2003) Can you bank on GenBank? Trends in Ecology and Evolution 18(7): 317–319. doi: 10.1016/S0169-5347(03)00150-2

Hartmann M, Howes CG, Veldre V, Schneider S, Vaishampayan PA, Yannarell AC, Quince C, Johansson P, Johanna Björkroth K, Abarenkov K, Hallam SJ, Mohn WW, Nilsson RH (2011) V-RevComp: Automated high-throughput detection of reverse complementary 16S ribosomal RNA gene sequences in large environmental and taxonomic datasets. FEMS Microbiology Letters 319(2): 140–145. doi: 10.1111/j.1574-6968.2011.02274.x

Hartmann M, Bengtsson J, Nilsson RH (in press) High-throughput quality control of next-generation sequencing datasets from environmental studies using hidden Markov models.

Hawksworth DL (2001) The magnitude of fungal diversity: the 1.5 million species estimate revisited. Mycological Research 105(12): 1422–1432. doi: 10.1017/S0953756201004725

Hibbett DS, Tsuneda A, Fukumasa-Nakai Y, Donoghue MJ (1995) Phylogenetic diversity in shiitake inferred from nuclear ribosomal DNA sequences. Mycologia 87(5): 618–638. doi: 10.2307/3760806

Hibbett DS, Bindera J, Bischoffb JF, Blackwellc M, Cannond PF, Erikssone OE, Huhndorff S, Jamesg T, Kirkd PM, Lückingf R, Lumbschf HT, Lutzonig F, Mathenya PB, McLaughlinh DJ, Powelli MJ, Redheadj S, Schochk CL, Spataforak JW, Stalpersl JA, Vilgalysg R, Aimem MC, Aptrootn A, Bauero R, Begerowp D, Bennyq GL, Castleburym LA, Crousl PW, Dair Y-C, Gamsl W, Geisers DM, Griffitht GW, Gueidang C, Hawksworthu DL, Hestmarkv G, Hosakaw K, Humberx RA, Hydey KD, Ironsidet J-E, Kõljalgz U, Kurtzmanaa CP, Larssonab K-H, Lichtwardtac R, Longcoread J, Miądlikowskag J, Millerae A, Moncalvoaf J-M, Mozley-Standridgeag S, Oberwinklero F, Parmastoah E, Reebg V, Rogersai JD, Rouxaj C, Ryvardenak L, Sampaioal JP, Schüßleram A, Sugiyamaan J, Thornao RG, Tibellap L, Untereineraq WA, Walkerar C, Wanga Z, Weiras A, Weisso M, Whiteat MM, Winkae K, Yaoau Y-J, Zhang N (2007) A higher-level phylogenetic classification of the Fungi. Mycological Research 111(5): 509–547. doi: 10.1016/j.mycres.2007.03.004

Hibbett DS, Ohman A, Glotzer D, Nuhn M, Kirk P, Nilsson RH (2011) Progress in molecular and morphological taxon discovery in Fungi and options for formal classification of environmental sequences. Fungal Biology Reviews 25(1): 38–47. doi: 10.1016/j.fbr.2011.01.001

Horton TR, Bruns TD (2001) The molecular revolution in ectomycorrhizal ecology: peeking into the black-box. Molecular Ecology 10(8): 1855–1871. doi: 10.1046/j.0962-1083.2001.01333.x

Hugenholtz P, Huber T (2003) Chimeric 16S sequences of diverse origin are accumulating in the public databases. International Journal of Systematic and Evolutionary Microbiology 53(1): 289–293. doi: 10.1099/ijs.0.02441-0

Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environmental Microbiology 12(7): 1889–1898. doi: 10.1111/j.1462-2920.2010.02193.x

James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, Lumbsch HT, Rauhut A, Reeb V, Arnold AE, Amtoft A, Stajich JE, Hosaka K, Sung G-H, Johnson D, O'Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schüßler A, Longcore JE, O'Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies RD, Humber RA, Morton JB, Sugiyama J, Rossman AY, Rogers JD, Pfister DH, Hewitt D, Hansen K, Hambleton S, Shoemaker RA, Kohlmeyer J, Volkmann-Kohlmeyer B, Spotts RA, Serdani M, Crous PW, Hughes KW, Matsuura K, Langer E, Langer G, Untereiner WA, Lücking R, Büdel B, Geiser DM, Aptroot A, Diederich P, Schmitt I, Schultz M, Yahr R, Hibbett DS, Lutzoni F, McLaughlin DJ, Spatafora JW, Vilgalys R (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature 443(7113): 818–822. doi: 10.1038/nature05110

Jones MDM, Forn I, Gadelha C, Egan MJ, Bass D, Massana R, Richards TA (2011) Discovery of novel intermediate forms redefines the fungal tree of life. Nature 474(7350): 200–203. doi: 10.1038/nature09984

Kang S, Mansfield MAM, Park B, Geiser DM, Ivors KL, Coffey MD, Grünwald NJ, Martin FN, Lévesque CA, Blair JE (2010) The promise and pitfalls of sequence-based identification of plant pathogenic fungi and oomycetes. Phytopathology 100(8): 732–737. doi: 10.1094/PHYTO-100-8-0732

Karsch-Mizrachi I, Nakamura Y, Cochrane G (2012) The International Nucleotide Sequence Database Collaboration. Nucleic Acids Research 40(D1): D33–D37. doi: 10.1093/nar/gkr1006

Katoh K, Toh H (2010) Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics 26(15): 1899–1900. doi: 10.1093/bioinformatics/btq224

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk JP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu HL, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057): 376-380. doi: 10.1038/nature03959

Miller MJ, Powell JI (1994) A quantitative comparison of DNA sequence assembly programs. Journal of Computational Biology 1(4): 257–269. doi: 10.1089/cmb.1994.1.257

Nilsson RH, Ryberg M, Kristiansson E, Abarenkov K, Larsson K-H, Kõljalg U (2006) Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. PLoS ONE 1: e59. doi: 10.1371/journal.pone.0000059

Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson K-H (2008) Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. Evolutionary Bioinformatics 4: 193–201. doi: 10.4137/EBO.S653

Nilsson RH, Ryberg M, Sjökvist E, Abarenkov K (2011a) Rethinking taxon sampling in the light of environmental sequencing. Cladistics 27(2): 197-203. doi: 10.1111/j.1096-0031.2010.00336.x

Nilsson RH, Veldre V, Wang Z et al. (2011b) A note on the incidence of reverse complementary fungal ITS sequences in the public sequence databases and a software tool for their detection and reorientation. Mycoscience 52(4): 278–282. doi: 10.1007/s10267-010-0086-z

Ovaskainen O, Nokso-Koivistoa J, Hottolaa J, Rajalab T, Pennanenb T, Ali-Koveroa H, Miettinenc O, Oinonenc P, Auvinend P, Paulind L, Larssone K-H, Mäkipää R (2010) Identifying wood-inhabiting fungi with 454 sequencing – what is the probability that BLAST gives the correct species? Fungal Ecology 3(4): 274–283. doi: 10.1016/j.funeco.2010.01.001

Porter TM, Skillman JE, Moncalvo J-M (2008) Fruiting body and soil rDNA sampling detects complementary assemblage of Agaricomycotina (Basidiomycota, Fungi) in a hemlock-dominated forest plot in southern Ontario. Molecular Ecology 17(13): 3037–3050. doi: 10.1111/j.1365-294X.2008.03813.x

Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV, Tiedje JM, Zhou J (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. Applied and Environmental Microbiology 67(2): 880-887. doi: 10.1128/AEM.67.2.880-887.2001

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12: 38. doi: 10.1186/1471-2105-12-38

Rosling A, Cox F, Cruz-Martinez K, Ihrmark K, Grelet G-A, Lindahl BD, Menkis A, James TY (2011) Archaeorhizomycetes: Unearthing an ancient class of ubiquitous soil fungi. Science 333(6044): 876–879. doi: 10.1126/science.1206958

Ross HA, Murugan S, Li WLS (2008) Testing the reliability of genetic methods of species identification via simulation. Systematic Biology 57(2): 216–230. doi: 10.1080/10635150802032990

Sanderson MJ, Donoghue MJ, Piel W, Eriksson T (1994) TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. American Journal of Botany 81(6): 183.

Schocha CL, Seifertb KA, Huhndorfc A, Robertd V, Spougea JL, Levesqueb CA, Chenb W, Fungal Barcoding Consortium (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for *Fungi*. Proceedings of the National Academy of Sciences USA 109(16): 6241–6246. doi: 10.1073/pnas.1117018109

Stajich JE, Berbee ML, Blackwell M, Hibbett DS, James TY, Spatafora JW, Taylor JW (2009) The fungi. Current Biology 19(18): R840–855. doi: 10.1016/j.cub.2009.07.004

Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Taylor JW, Jacobson DJ, Kroken S, Kasuga T, Geiser DM, Hibbett DS, Fisher MC (2000) Phylogenetic species recognition and species concepts in fungi. Fungal Genetics and Biology 31(1): 21–32. doi: 10.1006/fgbi.2000.1228

Taylor DL, McCormick MK (2008) Internal transcribed spacer primers and sequences for improved characterization of basidiomycetous orchid mycorrhizas. New Phytologist 177(4): 1020–1033. doi: 10.1111/j.1469-8137.2007.02320.x

Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G, Kõljalg U (2010) 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. New Phytologist 188(1): 291–301. doi: 10.1111/j.1469-8137.2010.03373.x

Tedersoo L, Abarenkov K, Nilsson RH, Arthur Schüssler, Grelet G-A, Kohout P, Oja J, Bonito GM, Veldre V, Jairus T, Ryberg M, Larsson K-H, Kõljalg U (2011) Tidying up International Nucleotide Sequence Databases: ecological, geographical, and sequence quality annotation of ITS sequences of mycorrhizal fungi. PLoS ONE 6: e24940. doi: 10.1371/journal.pone.0024940

Vrålstad T (2011) ITS, OTUs and beyond—fungal hyperdiversity calls for supplementary solutions. Molecular Ecology 20(14): 2873–2875. doi: 10.1111/j.1365-294X.2011.05149.x

Wang GC, Wang Y (1996) The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. Microbiology 142(5): 1107–1114. doi: 10.1099/13500872-142-5-1107

Wren JD (2008) URL decay in MEDLINE—a 4-year follow-up study. Bioinformatics 24(11): 1381–1385. doi: 10.1093/bioinformatics/btn127

Appendix

Technical considerations. (doi: 10.3897/mycokeys.4.3606.app) File format: PDF.

Explanation note: Discussion on sequence quality and reliability assessment for the more technically inclined user.

Citation: Nilsson RH, Tedersoo L, Abarenkov K, Ryberg M, Kristiansson E, Hartmann M, Schoch CL, Nylander JAA, Bergsten J, Porter TM, Jumpponen A, Vaishampayan P, Ovaskainen O, Hallenberg N, Bengtsson-Palme J, Eriksson KM, Larsson K-H, Larsson E, Kõljalg U (2012) Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. MycoKeys 4: 37–63. doi: 10.3897/mycokeys.4.3606.app