MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2010) **10**, 1076–1081 doi: 10.1111/j.1755-0998.2010.02850.x

COMPUTER PROGRAM NOTE

An open source chimera checker for the fungal ITS region

R. H. NILSSON,*†[1] KESSY ABARENKOV,†[1] VILMAR VELDRE,† STEPHAN NYLINDER,* PIERRE DE WIT,‡ SARA BROSCHÉ,* JOHAN F. ALFREDSSON,§ MARTIN RYBERG¶ and ERIK KRISTIANSSON‡**
*Department of Plant and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden, †Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai St., 51005 Tartu, Estonia, ‡Department of Zoology, University of Gothenburg, Box 463, 405 30 Göteborg, Sweden, §Oepir Consulting, Vasagatan 48:1, 411 37 Göteborg, Sweden, ¶Department of Ecology and Evolutionary Biology, University of Tennessee at Knoxville, TN 37996-1610, USA, **The Sahlgrenska Academy at the University of Gothenburg, Department of Neuroscience and Physiology, Box 434, 405 30 Göteborg, Sweden

Abstract

The internal transcribed spacer (ITS) region of the nuclear ribosomal repeat unit holds a central position in the pursuit of the taxonomic affiliation of fungi recovered through environmental sampling. Newly generated fungal ITS sequences are typically compared against the International Nucleotide Sequence Databases for a species or genus name using the sequence similarity software suite BLAST. Such searches are not without complications however, and one of them is the presence of chimeric entries among the query or reference sequences. Chimeras are artificial sequences, generated unintentionally during the polymerase chain reaction step, that feature sequence data from two (or possibly more) distinct species. Available software solutions for chimera control do not readily target the fungal ITS region, but the present study introduces a BLAST-based open source software package (available at http://www.emerencia.org/chimerachecker.html) to examine newly generated fungal ITS sequences for the presence of potentially chimeric elements in batch mode. We used the software package on a random set of 12 300 environmental fungal ITS sequences in the public sequence databases and found 1.5% of the entries to be chimeric at the ordinal level after manual verification of the results. The proportion of chimeras in the sequence databases can be hypothesized to increase as emerging sequencing technologies drawing from pooled DNA samples are becoming important tools in molecular ecology research.

Keywords: chimeric sequences, environmental sampling, fungi, internal transcribed spacer

Received 1 September 2009; revision received 4 February 2010; accepted 10 February 2010

Fungi form a ubiquitous group of primarily terrestrial organisms whose study is rendered difficult by the inconspicuous nature of fungal life. The correspondence between above-ground fruiting bodies and the full diversity of the actual fungal community below ground (or otherwise inside or associated with the substrate) is known to be poor, and a significant number of fungi do not appear to form fruiting bodies at all (O'Brien *et al.* 2005; Porter *et al.* 2008). Furthermore, reliable morphological characters for species identification and delimitation among fungi are relatively rare, and morphologically similar but phylogenetically distinct species are common throughout the fungal kingdom (Taylor *et al.* 2000; Hibbett 2007). DNA sequence data, particularly from the internal transcribed spacer (ITS) region of the nuclear ribosomal repeat unit, have thus proved an indispensable information source in the pursuit of the diversity of fungi at various scales and locations in time and space (Horton & Bruns 2001; Kõljalg *et al.* 2005; Vialle *et al.* 2009). With fully identified ITS sequences available for <1% of the estimated 1.5 million extant species of fungi, however, the process of assigning newly generated sequences to species level often proves a difficult and laborious task (Bueé *et al.* 2009; Ryberg *et al.* 2009; Seifert 2009). In addition, approximately 48%

Correspondence: R. Henrik Nilsson, Fax: 46-31-782 2650;
E-mail: henrik.nilsson@dpes.gu.se
[1]Equal contribution.

of the full-length fungal ITS sequences in the International Nucleotide Sequence Databases (INSD; Benson *et al.* 2008) lack a full species name, and as much as 20% of the fungal ITS sequences that indeed feature a species name may in fact be misidentified to species level or otherwise compromised (Nilsson *et al.* 2006; Bidartondo *et al.* 2008).

Identification procedures based on similarity searches are further complicated by an as yet unquantified proportion of chimeric fungal ITS sequences, i.e., artificial sequences generated unintentionally during the polymerase chain reaction (PCR) step and that typically feature sequence data from two distinct species (Wang & Wang 1997). Various studies have reported the discovery of the odd chimeric fungal ITS sequence in the public sequence databases (Ryberg *et al.* 2008; Taylor *et al.* 2008; Mullineux & Hausner 2009), but no estimate of their frequency exists, and the choice of the ITS – and its limited length of ~500 base pairs (bp) – as target region all but precludes any purposeful use of chimera check programs such as Bellerophon and Pintail (both of which were primarily designed with the considerably larger and length- and variability-wise much more homogeneous small subunit (16S) in mind; Huber *et al.* 2004; Ashelford *et al.* 2005). The proliferation of environmental sequencing efforts based on pooled DNA samples furthermore suggests that chimeric fungal sequences will be a problem of increasing concern over the next few years (cf. Christen 2008; Hibbett *et al.* 2009). The present study attempts a remedy in the form of an open source command-line software package to examine newly generated fungal ITS sequences for the presence of chimeric or otherwise artificial or anomalous elements. The package is written in Perl, makes use of only freely available auxiliary software and is available for download at http://www.emerencia.org/chimerachecker.html (Data S1) for UNIX-type operating systems, including MacOS X. Neither Internet access nor overly large amounts of computer memory (200 MB) are needed to run the software, which processes 1000 ITS sequences per 45 min on an average dual core computer. The command-line nature of the program suggests that some degree of familiarity with UNIX-type operating systems on behalf of the user is advantageous, although detailed instructions are available.

The package draws from the fact that the ITS region is composed of three subregions [the highly variable spacers ITS1 and ITS2 (~180 and ~170 bp, respectively) and the intercalary and very conserved gene encoding for the 5.8S (~160 bp; Nilsson *et al.* 2008; Fig. 1)] and the observation that the conserved 5.8S gene is by far the most likely place for any chimeric breakpoint (Ashelford *et al.* 2005). At least 100 bp (default) of both ITS1 and ITS2 – as well as the 5.8S gene in full – are required for a query (input) sequence to be processed. There is no limit on the number of query sequences, which are processed sequentially (Fig. 2). For each query sequence, the ITS1 and ITS2 are extracted *in silico* using HMMER 2 (Eddy 1998) and the Hidden Markov models of Nilsson *et al.* (2008). These two subregions, as well as the entire query sequence, are compared for similarity against a local (bundled) copy of the 75 000 fully identified fungal ITS sequences (as defined in Nilsson *et al.* 2005) in INSD (as of October 2009) using NCBI-BLAST (Altschul *et al.* 1997). The topmost entries in the BLAST hit list for both ITS1 and ITS2 are queried for their hierarchical INSD taxonomic affiliation (recently updated to reflect the new classification of fungi (Hibbett *et al.* 2007)). If the closest BLAST match of the ITS1 is annotated as belonging to a different taxonomic order (default level) than that of the ITS2, the entry is flagged as potentially chimeric and in need of further scrutiny. If the first and second closest BLAST match to the ITS1 region of the query sequence are annotated as belonging to different orders, the entry is marked as in potential need of further examination for the reason that taxonomic misidentification in INSD may obfuscate automated attempts at chimera discovery, and similarly for the ITS2 of the query sequence. Finally, to provide the user with a data-centric overview of the entries marked as potentially chimeric, all such entries are aligned jointly with their 15 (default) best BLAST matches using any of Clustal W (Thompson *et al.* 1997), DIALIGN-TX (Subramanian *et al.* 2008) or MAFFT (Katoh & Toh 2008). These alignments are viewable in alignment editors such as Seaview (Gouy *et al.* 2010) and Jalview (Waterhouse *et al.* 2009) and form a context in which the artificial nature of a chimeric sequence tends to stand out starkly.

Fig. 1 Schematic overview of the fungal ITS region. The very conserved 5.8S gene is located between the highly variable spacers ITS1 and ITS2, whose respective taxonomic signal are contrasted by the present software in the pursuit of potentially chimeric sequences.

The output of the package consists of three sets of files: (i) a comprehensible tab-separated file containing, for each query sequence found to be potentially chimeric, the BLAST results for each of ITS1, ITS2 and the full query sequence, together with additional information and statistics pertaining to the BLAST matches and the chimera check functions; (ii) a multiple alignment in the FASTA (Pearson & Lipman 1988) or Clustal W format for the entries flagged as potentially chimeric and their 15 best BLAST matches; and (iii) brief text format lists of all entries that were not flagged as potentially chimeric; entries that were flagged as potentially chimeric; and entries that

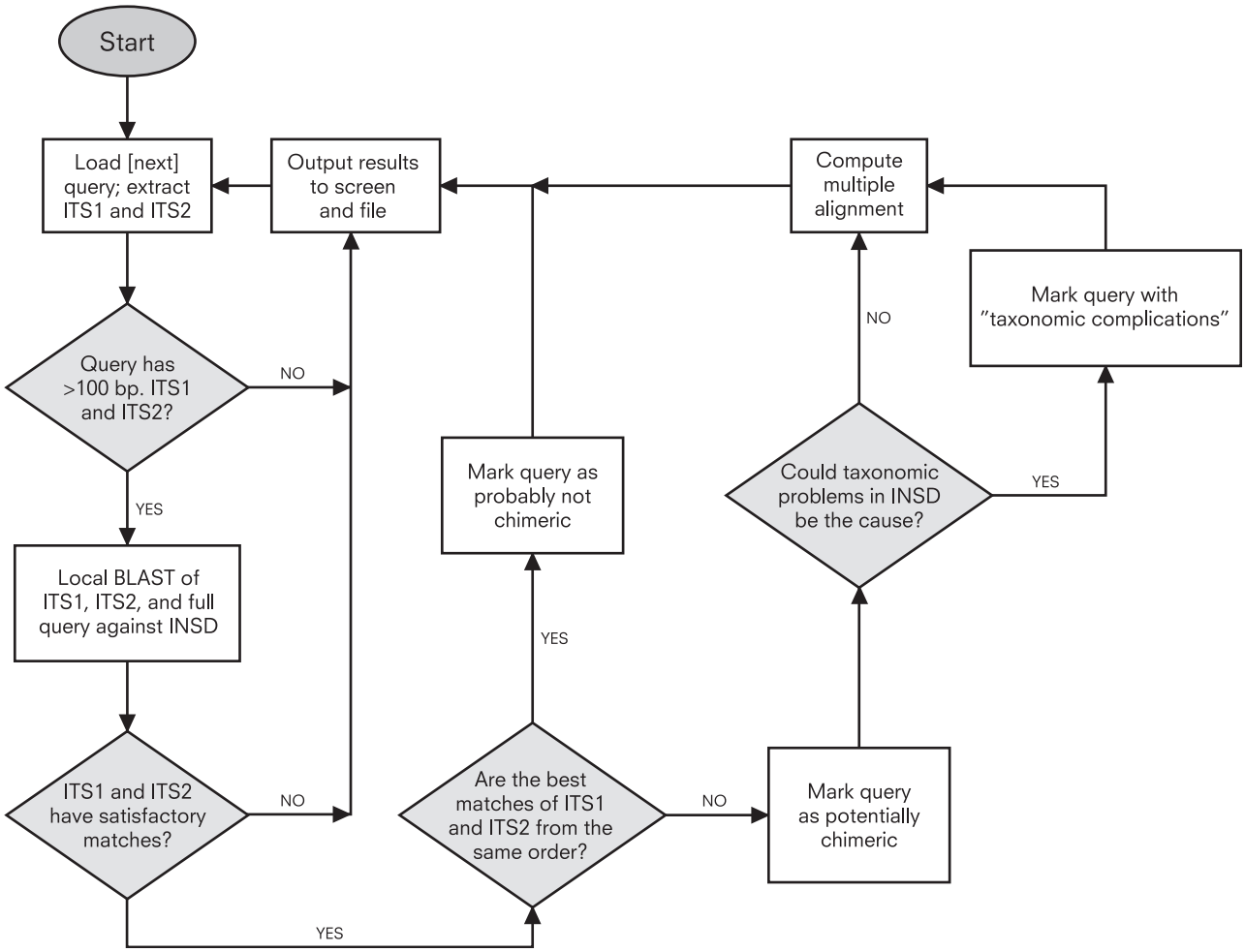

Fig. 2 Simplified flowchart representing the analysis pipeline of the software. An average nonchimeric sequence is processed in <15 s, but a chimeric sequence may take upwards of a minute to process, largely because of the multiple alignment step involved.

could not be processed for the lack of at least 100 bp of either or both of ITS1 and ITS2. Jointly, this set of files allows the user to get a detailed understanding of the reasons why a query sequence was flagged as potentially chimeric and provides an opportunity to view and compare the results in various ways.

Much like chimera check programs for other genes and groups of organisms, the present software package is in no position to *prove* that any query sequence is or is not chimeric. It is however designed to find sequences that are *potentially* chimeric and, in addition, to generate data and information pertinent to their evaluation. Even so, there is at present no way around manual interpretation of any potentially chimeric entries found. In an attempt at estimating how important the step of manual interpretation is, we randomly selected 15 000 insufficiently identified fungal ITS sequences (>450 bp) from INSD and used these as queries. A total of 2693 entries lacked either ITS1 or ITS2 and were excluded from the analysis. Of the remaining 12 307 entries, 1038 (8%) were flagged as potentially chimeric. These cases were examined manually using the alignment and data files generated by the software, and a total of 18% (182 of 1038) of these sequences were deemed to be true chimeras at the ordinal level such that our estimate of the proportion of chimeric environmental fungal ITS sequences in INSD is 1.5% (182 of the 12 307 entries examined). For 15% (155 of 1038) of the potentially chimeric entries, we were unable to assess the chimeric status of the sequence with reasonable certainty, chiefly because of the lack of similar enough sequences for purposes of comparison. Finally, we found 67% (701 of 1038) of the entries to be nonchimeric; taxonomic misidentification, inconsistent taxonomic hierarchies in INSD and gaps in the taxonomic sampling among the reference sequences were the principal reasons why these entries were flagged as potentially chimeric (see Data S1 for details). Thus, resulting in an 82% reduction in the number of potentially chimeric

sequences, the manual interpretation step seems every bit as important for the present software package as for, e.g., Bellerophon. To quantify the proportion of false negatives (i.e., queries not marked as potentially chimeric but for which a chimeric origin seems very likely), the alignments of 1000 queries not flagged as potentially chimeric were inspected by eye. A full 98.4% (984 of 1000) of them were deemed as clearly nonchimeric, 0% was deemed clearly chimeric, and in 1.6% (16 of 1000) of the cases the absence of sufficiently similar reference sequences precluded any conclusive decision as to the chimeric nature of the query. Thus, with 0% clear false negatives, the software does a precise job locating potentially chimeric sequences at the cost of a 5.7% incidence (701 of 12 307) of false positives.

With respect to taxonomic affiliation, the default setting of the package is to compare sequences at the ordinal level, a level at which fungal systematics is reasonably standardized (cf. Hibbett *et al*. 2007). The classification of fungi remains in a state of flux, however, and the partial lack of stability below the ordinal level (as well as compounding factors such as synonyms and anamorph – teleomorph relationships) may be taken as arguments against *in silico* comparison at the species, genus or family level (Binder *et al*. 2005; Blackwell *et al*. 2006; Hibbett *et al*. 2007). Thus, in its default state, the software package does not find chimeric unions that occurred between fungi of the same order; the fungal family concept is however fairly well developed for at least some groups of fungi such that the user may find that a family-centric comparison may be for the purpose of certain datasets. The package can furthermore be expected to perform suboptimally when only partial sequence data are available for either of ITS1 or ITS2 (or both) or when no reasonably closely related, fully identified reference sequences are available in INSD. The performance of the package on sequences that are doubly chimeric (although probably unlikely for the ITS region) or where the chimeric breakpoint occurs inside the ITS1 or ITS2 remains largely untested. An additional cause for concern is the presence of 'biological chimeras': sequences whose ITS region is the product of partial horizontal transfer with subsequent downpassing of the sequence through the generations (Xie *et al*. 2008).

The present software targets the fungal ITS region but does not require that the query sequences be reasonably closely related, of approximately the same, sizable length, or of a relatively homogeneous nature, all of which are factors that previously have made chimera control for fungal ITS sequences using existing software resources problematic. Although it does require that both ITS1 and ITS2 be present in the query sequence, it needs as little as 100 bp or less of each to perform well. Unlike, e.g., Bellerophon, it does not employ a sliding-window approach over the entire length of the query sequence but instead contrasts the taxonomic signal inherent to the constituent spacers ITS1 and ITS2 with one another. The software has a strong focus on batch mode operation as to allow processing of arbitrarily large datasets and to admit incorporation into software pipelines for sequence analysis and processing (e.g., Nilsson *et al*. 2009). Jointly, these features distinguish the present software from existing chimera control resources. ITS1 and ITS2 are extracted using HMMs tailored from large alignments of the neighbouring ribosomal genes, which makes the present release specific to fungi. These genes are variable enough that HMMs should be tailored for each separate group of organisms in the interest of precision. To adapt the software for operation on other groups of organisms is straightforward, however, and includes computing such HMMs in HMMER (Eddy 1998) from corresponding alignments for the target taxa and compiling a database of the relevant INSD entries to serve as reference sequences. HMMs for animals and oomycetes are bundled with the present release, and additional HMMs are in preparation.

Acknowledgements

RHN and KA gratefully acknowledge infrastructural support from the Fungi in Boreal Forest Soils network and from the Frontiers in Biodiversity Research Centre of Excellence (University of Tartu, Estonia). Figure 1 and 2 were compiled in collaboration with Bomb Mediaproduktion.

References

Altschul SF, Madden TL, Schaffer AA *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial errors. *Applied and Environmental Microbiology*, **71**, 7724–7736.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. *Nucleic Acids Research*, **36**, D25–D30.

Bidartondo MI, Bruns TD, Blackwell M *et al*. (2008) Preserving accuracy in GenBank. *Science*, **319**, 1616.

Binder M, Hibbett DS, Larsson K-H *et al*. (2005) The phylogenetic distribution of resupinate forms across the major clades of mushroom-forming fungi (*Homobasidiomycetes*). *Systematics and Biodiversity*, **3**, 113–157.

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

Bueé M, Reich M, Murat C *et al*. (2009) 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist*, **184**, 449–456.

Christen R (2008) Global sequencing: a review of current molecular data and new methods available to assess microbial diversity. *Microbes and Environments*, **23**, 253–268.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Feibelman TP, Bayman P, Cibula WG (1994) Length variation in the internal transcribed spacer of ribosomal DNA in chanterelles. *Mycological Research*, **98**, 614–618.

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**, 221–224.

Hibbett DS (2007) After the gold rush, or before the flood? Evolutionary morphology of mushroom-forming fungi (Agaricomycetes) in the early 21st century *Mycological Research*, **111**, 1001–1018.

Hibbett DS, Binder M, Bischoff JF *et al.* (2007) A higher-level phylogenetic classification of the Fungi. *Mycological Research*, **111**, 509–547.

Hibbett DS, Ohman A, Kirk PM (2009) Fungal ecology catches fire. *New Phytologist*, **184**, 279–282.

Horton TR, Bruns TD (2001) The molecular revolution in ectomycorrhizal ecology: peeking into the black-box. *Molecular Ecology*, **10**, 1855–1871.

Huber T, Faulkner G, Hugenholtz P (2004) Bellerophon; a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*, **20**, 2317–2319.

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, **9**, 286–298.

Kõljalg U, Larsson K-H, Abarenkov K *et al.* (2005) UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist*, **166**, 1063–1068.

Mullineux T, Hausner G (2009) Evolution of rDNA ITS1 and ITS2 sequences and RNA secondary structures within members of the fungal genera Grosmannia and Leptographium. *Fungal Genetics and Biology*, **46**, 855–867.

Nilsson RH, Kristiansson E, Ryberg M, Larsson K-H (2005) Approaching the taxonomic affiliation of unidentified sequences in public databases – an example from the mycorrhizal fungi. *BMC Bioinformatics*, **6**, 178.

Nilsson RH, Ryberg M, Kristiansson E *et al.* (2006) Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS ONE*, **1**, e59.

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.

Nilsson RH, Bok G, Ryberg M, Kristiansson E, Hallenberg N (2009) A software pipeline for processing and identification of fungal ITS sequences. *Source Code in Biology and Medicine*, **4**, 1.

O'Brien HE, Parrent JL, Jackson JA, Moncalvo JM, Vilgalys R (2005) Fungal community analysis by large-scale sequencing of environmental samples. *Applied and Environmental Microbiology*, **71**, 5544–5550.

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences, USA*, **85**, 2444–2448.

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**, 3037–3050.

Ryberg M, Nilsson RH, Kristiansson E *et al.* (2008) Mining metadata from unidentified ITS sequences in GenBank: a case study in *Inocybe* (*Basidiomycota*). *BMC Evolutionary Biology*, **8**, 50.

Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH (2009) An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity. *New Phytologist*, **181**, 471–477.

Seifert KA (2009) Progress towards DNA barcoding of fungi. *Molecular Ecology Resources*, **9**, 83–89.

Subramanian AR, Kaufmann M, Morgenstern B (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for Molecular Biology*, **3**, 6.

Taylor JW, Jacobson DJ, Kroken S *et al.* (2000) Phylogenetics species recognition and species concepts in Fungi. *Fungal Genetics and Biology*, **31**, 21–32.

Taylor DL, Booth MG, McFarland JW *et al.* (2008) Increasing ecological inference from high throughput sequencing of fungi in the environment through a tagging approach. *Molecular Ecology Resources*, **8**, 742–752.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, **25**, 4876–4882.

Vialle A, Feau N, Allaire M *et al.* (2009) Evaluation of mitochondrial genes as DNA barcode for *Basidiomycota*. *Molecular Ecology Resources*, **9**, 99–113.

Wang G, Wang Y (1997) Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Applied and Environmental Microbiology*, **63**, 4645–4650.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

Xie J, Fu Y, Jiang D *et al.* (2008) Intergeneric transfer of ribosomal genes between two fungi. *BMC Evolutionary Biology*, **8**, 87.

Supporting Information

Additional Supporting information may be found in the online version of this article.

Data S1 The software package together with its documentation, reference sequences from INSD and a test data set (including, for illustratory purposes, 10 sequences that are clearly chimeric, 10 sequences that are clearly nonchimeric and 10 sequences that cannot be processed for the lack of enough sequence data). In addition, the user will have to install NCBI-BLAST, HMMER, and one of Clustal W, MAFFT, and DIALIGN-TX; detailed installation instructions are provided in the documentation. The underlying directory was archived with tar and compressed with zip.

Data S2 A screenshot of the software package in operation on a dual core MacBook Pro running OS X 10.4.11.

Data S3 Hidden Markov models for the nuclear large and small subunits and the 5.8S for animals and oomycetes (*Oomycota*), provided to facilitate the implementation of the software for other organism groups where ITS sequences are used for scientific purposes. The models are based on inclusive alignments designed to capture the full width of the lineages. As with fungi, however, some taxa can be expected to have sequences deviant enough as to preclude automated attempts at locating the ITS1 and ITS2 with kingdom-level HMMs (cf. Feibelman *et al*. 1994).

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.