Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013 doi: 10.1111/2041-210X.12073

APPLICATION

Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data

Johan Bengtsson-Palme[1]*, Martin Ryberg[2], Martin Hartmann[3,4], Sara Branco[5], Zheng Wang[6], Anna Godhe[7], Pierre De Wit[7], Marisol Sánchez-García[8], Ingo Ebersberger[9], Filipe de Sousa[7], Anthony S. Amend[10], Ari Jumpponen[11], Martin Unterseher[12], Erik Kristiansson[13], Kessy Abarenkov[14], Yann J. K. Bertrand[7], Kemal Sanli[7], K. Martin Eriksson[15], Unni Vik[16], Vilmar Veldre† and R. Henrik Nilsson[7]

[1]*Department of Neuroscience and Physiology, The Sahlgrenska Academy, University of Gothenburg, Box 434, 40530 Göteborg, Sweden;* [2]*Department of Organismal Biology, Uppsala University, Norbyvägen 18D, Uppsala, 75236, Sweden;* [3]*Forest Soils and Biogeochemistry, Swiss Federal Research Institute WSL, Birmensdorf 8903, Switzerland;* [4]*Molecular Ecology, Agroscope Reckenholz-Tänikon Research Station ART, Zurich 8046, Switzerland;* [5]*Department of Plant and Microbial Biology, University of California, Berkeley, CA, 94720-3102, USA;* [6]*Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520-8106, USA;* [7]*Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, Göteborg, 40530, Sweden;* [8]*Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, 37996-1610, USA;* [9]*Department for Applied Bioinformatics, Institute for Cell Biology and Neuroscience, Goethe University, Max-von-Laue Str. 13, Frankfurt, D-60438, Germany;* [10]*Botany Department, University of Hawai'i at Manoa, 3190 Maile Way, Honolulu, HI, 96822, USA;* [11]*Division of Biology, Kansas State University, Manhattan, KS, 66506, USA;* [12]*Institute of Botany and Landscape Ecology, Ernst-Moritz-Arndt University Greifswald, Grimmer Str. 88, Greifswald, D-17487, Germany;* [13]*Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, 41296, Sweden;* [14]*Natural History Museum, University of Tartu, 46 Vanemuise Str., Tartu, 51014, Estonia;* [15]*Department of Shipping and Marine Technology, Chalmers University of Technology, Göteborg, 41296, Sweden; and* [16]*Department of Biosciences, University of Oslo, PO Box 1066, Blindern, Oslo, N-0316, Norway*

Summary

1. The nuclear ribosomal internal transcribed spacer (ITS) region is the primary choice for molecular identification of fungi. Its two highly variable spacers (ITS1 and ITS2) are usually species specific, whereas the intercalary 5.8S gene is highly conserved. For sequence clustering and BLAST searches, it is often advantageous to rely on either one of the variable spacers but not the conserved 5.8S gene. To identify and extract ITS1 and ITS2 from large taxonomic and environmental data sets is, however, often difficult, and many ITS sequences are incorrectly delimited in the public sequence databases.

2. We introduce ITSx, a Perl-based software tool to extract ITS1, 5.8S and ITS2 – as well as full-length ITS sequences – from both Sanger and high-throughput sequencing data sets. ITSx uses hidden Markov models computed from large alignments of a total of 20 groups of eukaryotes, including fungi, metazoans and plants, and the sequence extraction is based on the predicted positions of the ribosomal genes in the sequences.

3. ITSx has a very high proportion of true-positive extractions and a low proportion of false-positive extractions. Additionally, process parallelization permits expedient analyses of very large data sets, such as a one million sequence amplicon pyrosequencing data set. ITSx is rich in features and written to be easily incorporated into automated sequence analysis pipelines.

4. ITSx paves the way for more sensitive BLAST searches and sequence clustering operations for the ITS region in eukaryotes. The software also permits elimination of non-ITS sequences from any data set. This is particularly useful for amplicon-based next-generation sequencing data sets, where insidious non-target sequences are often found among the target sequences. Such non-target sequences are difficult to find by other means and would contribute noise to diversity estimates if left in the data set.

*Correspondence author. E-mail: johan@microbiology.se

†In memory of Vilmar Veldre.

Key-words: fungi, molecular ecology, next-generation sequencing, Perl, ribosomal DNA

Introduction

The fungal kingdom is estimated at 1·5 million extant species and comprises an ecologically heterogeneous assemblage of heterotrophic eukaryotes (Hawksworth 2001; Hibbett, Ohman & Kirk 2009). The subterranean or otherwise inconspicuous nature of much of fungal life tends to cede little ground to scientific scrutiny using traditional means, and molecular (DNA) data have emerged as an integral information source in the pursuit of mycological knowledge (De Vries *et al*. 2011; Hyde *et al*. 2013). Sequence analyses are now routine in systematics, taxonomy, and ecology of fungi (Peay, Kennedy & Bruns 2008; Yang 2011; Ebersberger *et al*. 2012), with the nuclear ribosomal operon being the most frequently targeted genetic region for such endeavours (Begerow *et al*. 2010). The small and large subunit genes (SSU/18S and LSU/28S, respectively) of the ribosomal operon are relatively conserved and are primarily used for large-scale phylogenetic inference and systematics. The *c*. 550 base-pair (bp) long internal transcribed spacer (ITS) region between them is more variable and is applied to decipher genus-level phylogenetic inference, species delimitation and species identification (Eberhardt 2010). It plays a similar role in several other groups of eukaryotes, including plants and animals (e.g. Feliner & Rosselló 2007; Li *et al*. 2011).

The use of the ITS region for molecular identification of fungi goes back to the early 1990s (Horton & Bruns 2001; Seifert 2009). The region is composed of the two highly variable spacers ITS1 and ITS2 which, jointly or separately, are often species specific, and the intercalary, very conserved 5.8S gene (Hillis & Dixon 1991). The sequence conservation in the proximate genes, coupled with numerous copies of the ribosomal operon, makes primer design and PCR amplification of the ITS region straightforward even from low-DNA-quantity substrates such as old herbarium specimens and soil. Indeed, the ITS region was recently designated the formal barcode for fungi for these and other reasons (Schoch *et al*. 2012). The ITS region is nevertheless not a barcoding marker without potential shortcomings. Complications include primer bias (Bellemain *et al*. 2010), differing evolutionary rates in different fungal lineages (Nilsson *et al*. 2008) and the presence of several different copies within a single individual (Lindner *et al*. 2013). A perhaps lesser-known complication with the ITS region in the context of molecular identification lies in its composite nature. The neighbouring SSU (immediately upstream of ITS1) and LSU (immediately downstream of ITS2) genes are very conserved, as is the intercalary 5.8S gene. The ITS1 and ITS2 spacers, on the other hand, are very variable. To subject sequences featuring both variable and conserved parts to similarity searches such as BLAST (Altschul *et al*. 1997) in the International Nucleotide Sequence Databases (INSD; Cochrane, Karsch-Mizrachi & Nakamura 2011) does not always produce the intended or correct results from the perspective of species identification. The conserved sequence parts likely find a match in the databases regardless of whether or not the variable part does, and so the outcome of the BLAST search may be more dependent on the length of the conserved component than the information content in the variable one (Hartmann *et al*. 2010; Kang *et al*. 2010). This would not be a concern if the reference databases featured an exhaustive taxon sampling of sequences of comparable length. Unfortunately, ITS sequence data are available only for a modest *c*. 1·5% of the estimated 1·5 million species of fungi (Hibbett *et al*. 2011), and the public fungal ITS sequences come in very different degrees of coverage of the region (Nilsson *et al*. 2008), cautioning against cursory – or fully automated – inspection of BLAST results. Nilsson *et al*. (2009) reported that 11% of the 86 000 BLAST searches undertaken produced a different result (non-synonymous species name) depending on whether the full ITS region, or just the variable regions, was used in the search. If the goal is to identify species (or finding other sequences from the same species, with or without a full Latin name), the BLAST search using either ITS1 or ITS2 may be preferable to using the full-length sequence.

Differentiating the individual components of the ITS region is not trivial, however. While SSU, 5.8S and LSU are conserved, they are regularly too variable for simple pattern matching approaches via regular expressions for their identification (Keller *et al*. 2009). A multiple ITS sequence alignment inspected in the light of the guidelines offered by Hibbett *et al*. (1995) is a good way to demarcate ITS1 and ITS2, but such manual approaches become unwieldy for larger data sets. To undertake it with data sets produced by high-throughput DNA sequencing techniques such as pyrosequencing (Margulies *et al*. 2005) – where the number of sequences may exceed hundreds of thousands – is intractable. Nilsson *et al*. (2010a) released a UNIX software package – Fungal ITS Extractor – to automatically identify, annotate and extract ITS1 and ITS2 from fungal ITS sequences. Drawing from HMMER version 2 (Eddy 1998), the software centred on profile hidden Markov models (HMMs) computed from large, kingdom-wide alignments for the 3′ end of SSU, the 5′ and 3′ ends of 5.8S, and the 5′ end of LSU. Profile HMMs are statistical models to represent the position-specific variations and dependencies typically observed in multiple sequence alignments; without having to store the full alignment, the HMMs are still able to account for the fact that a certain proportion of the sequences may contain, for example, a 'T' instead of an 'A' in some given position, while other positions appear invariable (Durbin *et al*. 1998). All query sequences were filtered through the HMMs, and extractions were made according to which HMMs that produced significant matches. A second use of the software was to filter out non-ITS sequences from large sequence data sets.

As noted by the authors themselves, however, the Fungal ITS Extractor is not impeccable. In larger fungal ITS data sets of heterogeneous taxonomic coverage, the proportion of missed or incorrect extractions – although typically detected as such by the program – can approach 1%. Further, the extractor does

not provide robust support for the genera *Cantharellus*, *Craterellus* or *Tulasnella*, the nuclear ribosomal genes of which are exceedingly divergent from other fungi (Feibelman, Bayman & Cibula 1994; Moncalvo *et al*. 2006; Taylor & McCormick 2008), and a group of Pezizalean ascomycetes characterized by a disruptive intron in the 3′ end of SSU is similarly problematic. Finally, the extractor operated on a single computer processor and although it processes all *c*. 300 000 fungal ITS sequences in INSD in less than 48 hours, the prospect of running a full pyrosequencing plate with an excess of one million sequences does not seem inviting. To improve the accuracy of the extractions and the runtime of the analysis, we introduce a complete software rewrite – ITSx (Item S1; http://microbiology.se/software/itsx/). We also introduce more than ten new features, including support for nineteen additional eukaryotic groups such as plants, animals, oomycetes and algae.

Software design and operation

Drawing from METAXA (Bengtsson *et al*. 2011, 2012), ITSx relies on the new HMMER version 3 (Eddy 2011) for profile hidden Markov model analysis. Fungal HMMs were computed for a 45-base-pair region of the immediate 3′ end of SSU, the 5′ and 3′ ends of 5.8S, and the 5′ end of LSU based on the kingdom-wide alignments of Tehler, Little & Farris (2003), Nilsson *et al*. (2008) and James *et al*. (2006), respectively. Separate alignments (and HMMs) were compiled for *Cantharellus*, *Craterellus* and *Tulasnella* to maintain alignment integrity in the core fungal alignments (cf. Hartmann *et al*. 2010). All fungal HMMs were then concatenated into a single, composite set of fungal HMMs – one for each gene region – to allow processing of all fungi at once. Separate, group-wide alignments and HMMs were similarly compiled for each of *Alveolata* (alveolates), *Amoebozoa* (amoebozoans), *Apusozoa*, *Bacillariophyta* (diatoms), *Bryophyta* (bryophytes), *Chlorophyta* (green algae), *Euglenozoa*, *Eustigmatophyceae* (eustigmatophytes), *Haptophyceae* (haptophytes), *Marchantiophyta* (liverworts), *Metazoa* (metazoans), *Oomycota* (oomycetes), *Parabasalia* (parabasalids), *Phaeophyceae* (brown algae), *Raphidophyceae* (raphidophytes), *Rhizaria*, *Rhodophyta* (red algae), *Synurophyceae* (synurids) and *Tracheophyta* (vascular plants). The INSD taxon definition was used for all of these groups. Upon starting the program, the user chooses which set of HMMs to employ (e.g. fungi); a composite search among the HMMs of all included groups of organisms is also available to facilitate the processing of mixed-taxon data sets. A slightly elevated risk of false-positive extractions may entail the use of these all-taxon searches, such that they should not be used unless the query data set indeed does span more than one of the eukaryotic groups supported.

ITSx expects query sequences in the FASTA format (Pearson & Lipman 1988), with or without gaps. There is no limit on the number of query sequences. The software first examines the sequences in the default orientation; the search is repeated in the reverse complementary orientation to account for incorrectly cast sequences (cf. Nilsson *et al*. 2011). Reverse complementary sequences are logged, reoriented and treated in the correct orientation in all subsequent steps. Each sequence is examined for matches to the HMMs. If the multiple-processor option is activated, ITSx employs the number of processor cores (or physical/logical processors as applicable) specified by the user, such that the speed of the analysis will roughly scale linearly with the number of CPU cores. An index is built of all regions matched by the HMMs.

The extraction is based on the HMM index of each query. By default, the ITS1 and ITS2 will be extracted from the query sequences and saved as separate FASTA files. The user can opt to also produce separate files for the SSU, 5.8S and LSU. In addition, FASTA files containing only those entries with the entire ITS region, or with the entire ITS1 or ITS2 regions, can be generated. This feature supports, for example, predicting the ITS1 and ITS2 secondary structure, which should be performed on full-length sequences (Koetschan *et al*. 2010). The SSU is extracted as everything from the 5′ end of the query sequence to the 3′ end of SSU as indicated by the HMM match; the ITS1 is extracted 1 bp downstream from the end of SSU and 1 bp upstream of the start of 5.8S; and so on. Partial extractions are supported. If, for example, only the 3′ end of 5.8S is detected, the ITS2 is extracted as everything downstream of that location. Various summary files are also written (see software documentation). A tab-separated file gives the start and stop positions for all markers in each query sequence. A log file records which query sequences, if any, were found to be reverse complementary. Additional, separate files record query sequences for which no HMMs were detected and query sequences for which the HMM matches occurred in an unexpected order. The open-source command line-based software is written in Perl, and it is freely available for UNIX-type operating systems (including MacOS X, LINUX and BSD). Although distributed over the Internet (Item S1; http://microbiology.se/software/itsx/), the software does not require Internet access to run. Computer memory (RAM) roughly 1·5 times the size of the input data set and free disc space corresponding to about 4–5 times the size of the query file are needed to run the software.

Evaluation and discussion

To evaluate the software, we compiled thirteen data sets of known, full-length ITS sequences (4674 sequences in all) from a total of nine major eukaryotic groups (Table S1) through INSD searches (Item S2). The data sets were analysed with ITSx under default settings, and the extraction efficiency was examined. We found the software to perform excellently on all data sets, with all genes detected in all sequences (occasionally some few base-pairs off; Table S1). We also ran ITSx on the raw 12 486-sequence ITS1 pyrosequencing data set of Kauserud *et al*. (2012). A total of 12 410 sequences were identified as fungal ITS1 sequences (Item S3); the remaining 76 sequences were examined by hand and were all found to be of low read quality and/or very short length. The run took 26 min to finish using one 2·2 GHz CPU core on a MacBook Pro laptop. In the light of this satisfactory true-positive performance, we evaluated the proportion of false positives by generating one

million random sequences of 550 bp in EMBOSS 6.2.0 (Rice, Longden & Bleasby 2000). Zero false-positive 'ITS' sequences were detected among the random sequences, suggesting a considerable robustness against spurious 'ITS' extractions (Item S4). The user can easily modify the stringency of the detection process by specifying HMMER cut-off E-values to support detection of sequences with less than *c*. 25 bases of the neighbouring ribosomal genes; however, this should normally be done only for data sets that are known to contain only ITS sequences. Conversely, very stringent settings may reduce sensitivity as sequences with deviant genes (or of low read quality) may be missed. For sequences that produce a match only to a single HMM (such as 3′ SSU), ITSx requires that match to be particularly good in order for the sequence to be scored as an ITS region sequence. This feature keeps the number of false-positive identifications low and can be controlled through the `-allow_single_domain` switch. The default settings of the software are calibrated with Sanger- and pyrosequencing-derived environmental data sets in mind.

While the new version of the software outperforms the old one in all respects examined and offers many new features, it is not intended as a panacea for ITS-based biological research. The extractor cannot identify sequences of SSU, 5.8S or LSU shorter than *c*. 20 bp (25 bp for consistent performance), and it is likely to have problems detecting the regions in sequences of poor read quality. ITS sequences that are chimeric (Nilsson *et al*. 2010b) or reverse complementary chimeric (Hartmann *et al*. 2011) may similarly be incompletely extracted. The extractor has several error detection and correction mechanisms – such as reverse complement control and checking that the matches to the HMMs were found in the correct order – and it is likely to catch many such compromised cases. Importantly, though, it should not be used as a chimera checker or as an arbiter of sequence read quality. Although we presently know of no such case, some lineages in the fungal tree of life may still have ribosomal genes deviant enough – or rich in introns – that they are not properly recovered by the HMMs in the present release. We ask the users to examine any cases where the extraction process does not seem to have worked *when it should have* and to notify us of any such observations. (However, support for the *Microsporidia*, which may or may not be fungal (Voigt & Kirk 2011), is pending, owing to the conceptually divergent configuration of their ribosomal operons.) The user also has the option to create custom HMMs of 45 bp long alignment segments and simply append these to the existing HMMs; after indexation as described in the HMMER documentation, the new HMMs will be integrated into ITSx. Our intention is that all fungal lineages should be fully supported without the need for tweaking of software settings, and we will gradually expand the HMMs as the ITS coverage in the public sequence database grows. The present release also introduces 19 additional sets of HMMs for other groups of eukaryotes for which the ITS region plays a role in molecular identification, species delimitation and phylogenetic inference. These HMMs have all been evaluated for basic performance, but it is likely that additional HMMs will be needed to fully capture the astonishing diversity of the Eukarya. A rule of thumb is to always evaluate the performance of ITSx on a subset of the target taxa prior to committing to full-size data sets.

In conclusion, we present an open-source software utility for robust extraction of the components of the ITS region in fungi and nineteen other groups of eukaryotes. This paves the way for sensitive sequence similarity searches and improves sequence clustering by facilitating the use of only the variable parts of the ITS region. A second use of the software is to sort out, from any given data set, sequences that come, or do not come, from the ITS region. This feature should be particularly useful for next-generation sequencing applications, where even amplicon-based runs often contain non-target sequences that are difficult to catch using other means (Quince *et al*. 2011). If assumed as target sequences, these entries likely exaggerate obtained diversity estimates (Dickie 2010; Tedersoo *et al*. 2010). We evaluated the software and found the proportion of correct extractions to be very high. We also showed that the proportion of false positives was very low. ITSx operates on Sanger and NGS-derived data sets alike, regardless of size and coverage of the ITS region. It is conservative regarding memory and disc space, and it is written to be easily incorporated into software pipelines. It is released with the intent that the research community will evaluate its performance also in the parts of the eukaryote tree we ourselves are less used to treading and – if needed – contribute to the alignment of data required to address also those lineages in a thorough way.

Acknowledgements

The authors have no competing interests to declare. Financial support from FORMAS (grant FORMAS, 215-2011-498) and the Carl Stenholm Foundation to RHN are acknowledged. The North European Forest Mycologists and the GOTBIN networks are acknowledged for infrastructural support. Our co-author Vilmar Veldre regrettably passed away during the making of this study, and the remaining co-authors wish to express their sincere gratitude to him for his remarkable energy and passion.

Data Accessibility

ITSx and related files are available for free download at http://microbiology.se/software/itsx/.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Begerow, D., Nilsson, R.H., Unterseher, M. & Maier, W. (2010) Current state and perspectives of fungal DNA barcoding and rapid identification procedures. *Applied Microbiology and Biotechnology*, **87**, 99–108.

Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P. & Kauserud, H. (2010) ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Microbiology*, **10**, 189.

Bengtsson, J., Eriksson, K.M., Hartmann, M., Wang, Z., Shenoy, B.D., Grelet, G.-A. *et al*. (2011) Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie van Leeuwenhoek*, **100**, 471–475.

Bengtsson, J., Hartmann, M., Unterseher, M., Vaishampayan, P., Abarenkov, K., Durso, L. *et al*. (2012) Megraft: a software package to graft ribosomal small subunit (16S/18S) fragments onto full-length sequences for accurate

species richness and sequencing depth analysis in pyrosequencing-length metagenomes and similar environmental datasets. *Research in Microbiology*, **163**, 407–412.

Cochrane, G., Karsch-Mizrachi, I. & Nakamura, Y. (2011) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, **39**, D15–D18.

De Vries, R.P., Benoit, I., Doehlemann, G., Kobayashi, T., Magnuson, J.K. Panisko, E.A. *et al.* (2011) Post-genomic approaches to understanding interactions between fungi and their environment. *IMA Fungus*, **2**, 81–86.

Dickie, I.A. (2010) Insidious effects of sequencing errors on perceived diversity in molecular surveys. *New Phytologist*, **188**, 916–918.

Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Eberhardt, U. (2010) A constructive step towards selecting a DNA barcode for fungi. *New Phytologist*, **187**, 265–268.

Ebersberger, I., de Matos Simoes, R., Kupczok, A., Gube, M., Kothe, E., Voigt, K. *et al.* (2012) A consistent phylogenetic backbone for the fungi. *Molecular Biology and Evolution*, **29**, 1319–1334.

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

Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Computational Biology*, **7**, e1002195.

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

Feliner, G.N. & Rosselló, J.A. (2007) Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution*, **44**, 911–919.

Hartmann, M., Howes, C.G., Abarenkov, K., Mohn, W.W. & Nilsson, R.H. (2010) V-Xtractor: an open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *Journal of Microbiological Methods*, **83**, 250–253.

Hartmann, M., Howes, C.G., Veldre, V., Schneider, S., Vaishampayan, P.A., Yannarell, A.C. *et al.* (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**, 140–145.

Hawksworth, D.L. (2001) The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycological Research*, **105**, 1422–1432.

Hibbett, D.S., Ohman, A. & Kirk, P.M. (2009) Fungal ecology catches fire. *New Phytologist*, **184**, 279–282.

Hibbett, D.S., Tsuneda, A., Fukumasa-Nakai, Y. & Donoghue, M.J. (1995) Phylogenetic diversity in shiitake inferred from nuclear ribosomal DNA sequences. *Mycologia*, **87**, 618–638.

Hibbett, D.S., Ohman, A., Glotzer, D., Nuhn, M., Kirk, P.M. & Nilsson, R.H. (2011) Progress in molecular and morphological taxon discovery in Fungi and options for formal classification of environmental sequences. *Fungal Biology Reviews*, **25**, 38–47.

Hillis, D.M. & Dixon, M.T. (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. *Quarterly Review of Biology*, **66**, 411–53.

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

Hyde, K.D., Udayanga, D., Manamgoda, D.S., Tedersoo, L., Larsson, E., Abarenkov, K. *et al.* (2013) Incorporating molecular data in fungal systematics: a guide for aspiring researchers. *Current Research in Environmental and Applied Mycology*, **3**, 1–32.

James, T.Y., Kauff, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J. *et al.* (2006) Reconstructing the early evolution of *Fungi* using a six-gene phylogeny. *Nature*, **443**, 818–822.

Kang, S., Mansfield, M.A., Park, B., Geiser, D.M., Ivors, K.L., Coffey, M.D. *et al.* (2010) The promise and pitfalls of sequence-based identification of plant-pathogenic fungi and oomycetes. *Phytopathology*, **100**, 732–737.

Kauserud, H., Kumar, S., Brysting, A.K., Nordén, J. & Carlsen, T. (2012) High consistency between replicate 454 pyrosequencing analyses of ectomycorrhizal plant root samples. *Mycorrhiza*, **22**, 309–315.

Keller, A., Schleicher, T., Schultz, J., Müller, T., Dandekar, T. & Wolf, M. (2009) 5.8S-28S rRNA interaction and HMM-based ITS2 annotation. *Gene*, **430**, 50–57.

Koetschan, C., Förster, F., Keller, A., Schleicher, T., Ruderisch, B., Schwarz, R. *et al.* (2010) The ITS2 Database III - sequences and structures for phylogeny. *Nucleic Acids Research*, **38**, D275–D279.

Li, D.-Z., Gao, L.-M., Li, H.-T., Wang, H., Ge, X.J., Liu, J.Q. *et al.* (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences USA*, **108**, 19641–19646.

Lindner, D.L., Carlsen, T., Nilsson, R.H., Davey, M., Schumacher, T. & Kauserud, H. (2013) Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. *Ecology and Evolution*, in press. DOI: 10.1002/ece3.586.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S. Bemben, L.A. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Moncalvo, J.-M., Nilsson, R.H., Koster, B., Dunham, S.M., Bernauer, T. Matheny, P.B. *et al.* (2006) The cantharelloid clade: dealing with incongruent gene trees and phylogenetic reconstruction methods. *Mycologia*, **98**, 937–948.

Nilsson, R.H., 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, R.H., Ryberg, M., Abarenkov, K., Sjökvist, E. & Kristiansson, E. (2009) The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiology Letters*, **296**, 97–101.

Nilsson, R.H., Veldre, V., Hartmann, M., Unterseher, M., Amend, A., Bergsten, J. *et al.* (2010a) An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. *Fungal Ecology*, **3**, 284–287.

Nilsson, R.H., Abarenkov, K., Veldre, V., Nylinder, S., De Wit, P. Brosché, S. *et al.* (2010b) An open source chimera checker for the fungal ITS region. *Molecular Ecology Resources*, **10**, 1076–1081.

Nilsson, R.H., Veldre, V., Wang, Z., Eckart, M., Branco, S. Hartmann, M. *et al.* (2011) 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**, 278–282.

Pearson, W.R. & Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences USA*, **85**, 2444–2448.

Peay, K.G., Kennedy, P.G. & Bruns, T.D. (2008) Fungal community ecology: a hybrid beast with a molecular master. *BioScience*, **58**, 799–810.

Quince, C., Lanzén, A., Davenport, R.J. & Turnbaugh, P.J. (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.

Rice, P., Longden, I. & Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276–277.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A. *et al.* (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**, 6241–6246.

Seifert, K.A. (2009) Progress towards DNA barcoding of fungi. *Molecular Ecology Resources*, **9**(S1), 83–89.

Taylor, D.L. & McCormick, M.K. (2008) Internal transcribed spacer primers and sequences for improved characterization of basidiomycetous orchid mycorrhizas. *New Phytologist*, **177**, 1020–1033.

Tedersoo, L., Nilsson, R.H., Abarenkov, K., Jairus, T., Sadam, A., Saar, I. *et al.* (2010) 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist*, **188**, 291–301.

Tehler, A., Little, D. & Farris, J.S. (2003) The full-length phylogenetic tree from 1551 ribosomal sequences of chitinous fungi, *Fungi*. *Mycological Research*, **107**, 901–916.

Voigt, K. & Kirk, P.M. (2011) Recent developments in the taxonomic affiliation and phylogenetic positioning of fungi: impact in applied microbiology and environmental biotechnology. *Applied Microbiology and Biotechnology*, **90**, 41–57.

Yang, Z.L. (2011) Molecular techniques revolutionize knowledge of basidiomycete evolution. *Fungal Diversity*, **50**, 47–58.

Received 12 March 2013; accepted 23 May 2013
Handling Editor: Michael Bunce

Supporting Information

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

Table S1. Evaluation of the extraction efficiency of ITSx in nine major groups of eukaryotes. All input sequences covered the full length of the ITS region and featured at least 20 bp of the SSU and LSU genes. Each gene region occupies four columns in the table. Column one indicates

the percentage of cases where the gene region was detected at all. Column two indicates the percentage of cases where the gene region was detected and delimited exactly to the correct base-pair. Column three indicates the percentage of cases where the gene region was detected but where the delimitation was one base-pair off. Column four indicates the percentage of cases where the gene region was detected but where the delimitation was two to five base-pairs off.

Item S1. The software package together with its documentation and a test data set.

Item S2. The test data sets used to evaluate ITSx, plus the output files from the analyses.

Item S3. The output from the evaluation of the pyrosequencing data sets.

Item S4. The output from the one million sequence runs of random DNA sequence data for evaluation of the susceptibility of ITSx to false-positive extractions.