FUNGAL ECOLOGY 3 (2010) 284–287

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/funeco


Methodological Advances

# 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

*R. Henrik NILSSON*[a,b,*], *Vilmar VELDRE*[b], *Martin HARTMANN*[c], *Martin UNTERSEHER*[d], *Anthony AMEND*[e], *Johannes BERGSTEN*[f], *Erik KRISTIANSSON*[g,h], *Martin RYBERG*[i], *Ari JUMPPONEN*[j], *Kessy ABARENKOV*[b]

[a]*Department of Plant and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden*
[b]*Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai St., 51005 Tartu, Estonia*
[c]*Department of Microbiology and Immunology, University of British Columbia, Life Sciences Centre, 4504-2350 Health Sciences Mall, Vancouver, BC V6T 1Z3, Canada*
[d]*Ernst-Moritz-Arndt University Greifswald, Institute of Botany and Landscape Ecology, Grimmer Str. 88, D-17487 Greifswald, Germany*
[e]*Department of Plant and Microbial Biology, University of California at Berkeley, 321 Koshland Hall, Berkeley, CA 94720-3102, USA*
[f]*Department of Entomology, Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden*
[g]*Department of Zoology, University of Gothenburg, Box 463, 405 30 Göteborg, Sweden*
[h]*The Sahlgrenska Academy at the University of Gothenburg, Department of Neuroscience and Physiology, Box 434, 405 30 Göteborg, Sweden*
[i]*Department of Ecology and Evolutionary Biology, University of Tennessee at Knoxville, TN 37996-1610, USA*
[j]*Division of Biology, Kansas State University, Manhattan, KS 66506, USA*

ARTICLE INFO

*Article history:*
Received 15 January 2010
Revision received 5 May 2010
Accepted 12 May 2010
Available online 26 June 2010
*Corresponding editor:* Anne Pringle

*Keywords:*
Environmental sampling
*ITS*
Molecular ecology
Nuclear ribosomal genes
Sequence identification

ABSTRACT

We introduce an open source software utility to extract the highly variable *ITS1* and *ITS2* subregions from fungal nuclear *ITS* sequences, the region of choice for environmental sampling and molecular identification of fungi. Inclusion of parts of the neighbouring, very conserved, ribosomal genes in the sequence identification process regularly leads to distorted results. The utility is available for UNIX-type operating systems, including MacOS X, and processes about 1 000 sequences per minute.



* *Corresponding author*. Department of Plant and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden. Tel.: +46 31 786 2623; fax: +46 31 786 2560.
E-mail address: henrik.nilsson@dpes.gu.se (R.H. Nilsson).
1754-5048/$ – see front matter © 2010 Elsevier Ltd and The British Mycological Society. All rights reserved.
doi:10.1016/j.funeco.2010.05.002

Fungi form a ubiquitous group of heterotrophic organisms with a largely subterranean or otherwise inconspicuous life cycle (Blackwell *et al.* 2006). The poor correspondence between aboveground fruit bodies and the diversity of the fungal community below ground (or in the substrate) has precluded a detailed understanding of the species composition of fungal communities, although the access to DNA sequence data is starting to change this (Peay *et al.* 2008; Hibbett *et al.* 2009). The most commonly sequenced genetic marker for molecular identification of fungi from environmental samples is the internal transcribed spacer (*ITS* ) region of the nuclear ribosomal repeat unit (Ryberg *et al.* 2009; Abarenkov *et al.* 2010). It is cumbersome to attain the sequence depth needed for reasonably accurate views of the underlying community, but emerging sequencing technologies such as massively parallel ("454") pyrosequencing (Margulies *et al.* 2005) address this and shift the focus from sequence depth to sequence processing and quality control (Huse *et al.* 2007; Shendure & Ji 2008; Galand *et al.* 2009; Kunin *et al.* 2010).

Massively parallel pyrosequencing generates reads shorter than those obtained by traditional Sanger sequencing. As a result, sequencing the *ITS* region in full (450–650+ bp.) is presently impossible. The *ITS* offers three potential target subregions for which numerous PCR primers are readily available: *ITS1*, *5.8S*, and *ITS2* (Fig 1). The *ITS1* is highly variable and about 180-base pairs (bp.) in length. The *ITS2* is nearly as variable although slightly shorter (~170 bp.); the lengths of *ITS1* and *ITS2* do, however, vary substantially among taxa (Nilsson *et al.* 2008). The intercalary *5.8S* gene (~160 bp.) is very conserved and can be aligned across the fungal phyla. The flanking ribosomal genes nuclear small subunit (*nSSU/18S* ) upstream of *ITS1* and nuclear large subunit (*nLSU/28S* ) downstream of *ITS2* make good primer anchors (for *ITS1* and *ITS2*, respectively), with the intercalary *5.8S* serving as the second anchor region (*cf.* Bueé *et al.* 2009; Jumpponen & Jones 2009). Depending on how far into these genes the primer sites are, however, the residual portions of the genes left in the *ITS* sequence may skew sequence similarity searches involving, e.g., BLAST (Altschul *et al.* 1997). These extra sequence segments, many times more conserved than *ITS1* and *ITS2*, will always find matches in the sequence databases – even when *ITS1* and *ITS2* do not – and so will invariably add to the length of the BLAST alignment. This makes automated interpretation of the BLAST results problematic and regularly has the effect that a different sequence or even species is presented as the best match than if *ITS1* or *ITS2* alone had been analyzed (Bruns & Shefferson 2004). This was observed for 11 % of the 86 000 *ITS*-based BLAST searches studied by Nilsson *et al.* (2009). Sequence clustering into hypothetically conspecific taxonomic units may similarly be distorted by these segments.

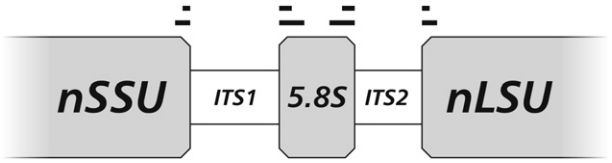


**Fig 1 – Overview of the fungal *ITS* region. The spacer *ITS1* is found between the 3′ end of the *nSSU* (*18S* ) gene and the 5′ end of the *5.8S* gene, and the *ITS2* is found between the 3′ end of *5.8S* and the 5′ end of the *nLSU* (*28S* ) gene. The long bars above the genes indicate the location of the long HMM for that particular gene, and the short bars indicate the location of the short HMMs. All HMMs are positioned in such a way as to cover the very end and the very beginning of the respective genes. The individual lengths of the HMMs were adjusted to reflect the position of the most commonly used primers.**

Though conserved, the *nSSU* and *nLSU* are variable enough that they cannot be located and deleted using regular expressions or pattern matching for a wide selection of fungi. They can be removed manually given a multiple alignment and a primer chart or an annotated reference alignment such as the one provided by Hibbett *et al.* (1995), but this becomes unfeasible as datasets grow. The present study introduces a software utility to extract the *ITS1* and *ITS2* from large fungal *ITS* datasets. The software accounts for partial sequence data – such as when only *nSSU* and half of *ITS1* are available – as well as input sequences in the reverse complementary direction. It is available at http://www.emerencia.org/FungalITSextractor.html (Supplementary material 1) for UNIX-based operating systems, including MacOS X.

The software is written in Perl and processes FASTA format (Pearson & Lipman 1988) input sequences sequentially. *ITS1* and *ITS2* are located using long (30–50 bp.) and short (18–25 bp.) Hidden Markov models (HMMs) computed in the HMMER package (v. 2; Eddy 1998) from inclusive alignments of the *nSSU* (3′ region; Tehler *et al.* 2003), *5.8S* (5′ and 3′ regions; Nilsson *et al.* 2008), and *nLSU* (5′ region; James *et al.* 2006). The query sequence is compared to the long HMMs for each of *nSSU*, *5.8S* (5′ and 3′), and *nLSU* using the HMMER package. If the boundaries of these genes can be detected, *ITS1* and *ITS2* are extracted from the sequence based on those positions. If not, an attempt is made to locate the genes using the shorter HMMs to account for sequences with shorter included conserved regions. Partial extractions are performed if one or more, but not all, genes are detected; if, for instance, only the 5′ end of *5.8S* is found, the *ITS1* is extracted as the region upstream of *5.8S*. A set of FASTA files comprises the core output of the software; these include all *ITS1* and *ITS2* sequences extracted from the input sequences and all sequences for which neither subregion could be found. The program outputs detailed information to the screen, including a summary of the extraction process for each input sequence and the absolute position of each of the subregions in the sequence. The feature to highlight sequences for which none of the flanking regions could be found is of particular relevance to massively parallel pyrosequencing, where artificial sequences consisting entirely of noise are sometimes produced and may pass filters based solely on quality scores (Quince *et al.* 2009). Furthermore, provided that the query sequences feature the 5′ region of *5.8S*, reverse complementary sequences are indicated as such and given in the correct direction.

The software performs better the more of *nSSU*/*5.8S*/*nLSU* are available (up to about 50 bp.). We found that the use of HMMs down to *ca.* 18 bp. in size (matching as little as 18 bp. of the distal part of any of the genes) still provide satisfactory

matches with few false positives. Any HMMs shorter than that are difficult to construct if they still are to be used for a wide selection of fungi. Sequences of poor read quality may pose a problem to the software insofar as the region to which the HMM is compared is obfuscated by incorrect or ambiguous nucleotides. The software may also perform suboptimally on taxa with very deviant rRNA genes — notably the genera *Cantharellus* and *Tulasnella* as well as some basal lineages such as the *Microsporidia* (Feibelman *et al.* 1994; James *et al.* 2006; Moncalvo *et al.* 2006; Taylor & McCormick 2008) — and taxa with large insertions or deletions in the regions targeted by the HMMs (*cf.* Shinohara *et al.* 1996; Bhattacharya *et al.* 2000; Holst-Jensen *et al.* 2004). To seek to modify general fungal HMMs to also include these deviant lineages may detract from the usefulness of the HMMs for the non-deviant lineages; instead, such taxa should be addressed using tailored HMMs.

We evaluated the software on 1 500 *ITS* sequences from all fungal phyla in GenBank (Benson *et al.* 2008) and from the *Quercus* phyllosphere pyrosequencing data of Jumpponen & Jones (2009) (Supplementary material 2). All sequences were compared with Hibbett *et al.* (1995) to identify the subregions, and the results were juxtaposed with those obtained from the software. The respective subregions were identified and extracted successfully for 1 462 (97.5 %) of the sequences. The 38 cases where the extraction of either subregion failed were explained by poor sequence data (17 instances), the failure of the HMMs to identify the regions due to the deviant nature of the taxon under scrutiny (11 instances), and false negatives (6 instances). In addition 4 (0.3 %) false positives (incorrect extractions) were observed. The user may be able to enhance the performance of the software further by tailoring the HMMER E-values (Eddy 1998) to suit any specific property of the target sequences, such as taxonomic affiliation. The very nature of environmental sequencing does, however speak against static assumptions about which taxa are present in samples at hand.

## Acknowledgements

RHN and KA gratefully acknowledge support from the Frontiers in Biodiversity Research Centre of Excellence (University of Tartu) and the Fungi in Boreal Forest Soils Network. Dr. David Taylor of Unilever is thanked for valuable input on the project.

## Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.funeco.2010.05.002.

## 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, 2010. The UNITE database for molecular identification of fungi — recent updates and future perspectives. *New Phytologist* **186**: 281–285.

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**: 3389–3402.

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

Bhattacharya D, Lutzoni F, Reeb V, Simon D, Nason J, Fernandez F, 2000. Widespread occurrence of spliceosomal introns in the rDNA genes of ascomycetes. *Molecular Biology and Evolution* **17**: 1971–1984.

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

Bruns TD, Shefferson RP, 2004. Evolutionary studies of ectomycorrhizal fungi: milestones and future directions. *Canadian Journal of Botany* **82**: 1122–1132.

Bueé M, Reich M, Murat C, Nilsson RH, Uroz S, Martin F, 2009. 454 pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist* **184**: 449–456.

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.

Galand P, Casamayor E, Kirchman D, Lovejoy C, 2009. Ecology of the rare microbial biosphere of the Arctic ocean. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 22427–22432.

Hibbett DS, Tsuneda A, Fukumasa-Nakai Y, Donoghue MJ, 1995. Phylogenetic diversity in shiitake inferred from nuclear ribosomal DNA sequences. *Mycologia* **87**: 618–638.

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

Holst-Jensen A, Vrålstad T, Schumacher T, 2004. On reliability. *New Phytologist* **161**: 11–13.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM, 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**: R143.

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 GH, Johnson D, O'Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schussler A, Longcore JE, O'Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies DR, 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, Lucking R, Budel 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**: 818–822.

Jumpponen A, Jones KL, 2009. Massively parallel 454 sequencing indicates hyperdiverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. *New Phytologist* **184**: 438–448.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P, 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* **12**: 118–123.

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 GP, 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 H, 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**: 376–380.

Moncalvo J-M, Nilsson RH, Koster B, Dunham SM, Bernauer T, Matheny PB, McLenon T, Margaritescu S, Weiss M, Garnica S, Danell E, Langer E, Langer G, Larsson E, Larsson K-H, Vilgalys R, 2006. The cantharelloid clade: dealing with incongruent gene trees and phylogenetic reconstruction methods. *Mycologia* **98**: 937–948.

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

Pearson WR, Lipman DJ, 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences U.S.A.* **85**: 2444–2448.

Peay KG, Kennedy PG, Bruns TD, 2008. Fungal community ecology: a hybrid beast with a molecular master. *BioScience* **58**: 799–810.

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT, 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods* **6**: 639–641.

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

Shendure J, Ji H, 2008. Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135–1145.

Shinohara ML, LoBuglio KF, Rogers SO, 1996. Group-I intron family in the nuclear ribosomal RNA small subunit genes of *Cenococcum geophilum* isolates. *Current Genetics* **29**: 377–387.

Taylor DL, McCormick K, 2008. Internal transcribed spacer primers and sequences for improved characterization of basidiomycetous orchid mycorrhizas. *New Phytologist* **177**: 1020–1033.

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