


Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences

Vinita Deshpande, Qiong Wang, Paul Greenfield, Michael Charleston, Andrea Porras-Alfaro, Cheryl R. Kuske, James R. Cole, David J Midgley & Nai Tran-Dinh


To cite this article: Vinita Deshpande, Qiong Wang, Paul Greenfield, Michael Charleston, Andrea Porras-Alfaro, Cheryl R. Kuske, James R. Cole, David J Midgley & Nai Tran-Dinh (2016) Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences, *Mycologia*, 108:1, 1-5, DOI: [10.3852/14-293](https://doi.org/10.3852/14-293)


To link to this article: <https://doi.org/10.3852/14-293>

 View supplementary material 

 Published online: 20 Jan 2017.

 Submit your article to this journal 

 Article views: 718

 View related articles 

 View Crossmark data 

 Citing articles: 3 View citing articles 

Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences

Vinita Deshpande

*School of Information Technologies, University of Sydney,
Sydney, New South Wales, 2006, Australia*

Qiong Wang

*Center for Microbial Ecology, Michigan State University, East
Lansing, Michigan 48823*

Paul Greenfield

CSIRO, North Ryde, New South Wales, 2113, Australia

Michael Charleston

*School of Physical Sciences, University of Tasmania, Sandy
Bay, Tasmania, 7005; School of Information Technologies,
University of Sydney, Sydney, New South Wales, 2006, Australia*

Andrea Porras-Alfaro

*Department of Biological Sciences, Western Illinois University,
Macomb, Illinois 61455*

Cheryl R. Kuske

*Bioscience Division, Los Alamos National Laboratory, Los
Alamos, New Mexico 87545*

James R. Cole

*Center for Microbial Ecology, Michigan State University, East
Lansing, Michigan 48823*

David J Midgley

CSIRO, North Ryde, New South Wales, 2113, Australia

Nai Tran-Dinh¹

CSIRO, North Ryde, New South Wales, 2113, Australia

Abstract: Fungi are key organisms in many ecological processes and communities. Rapid and low cost surveys of the fungal members of a community can be undertaken by isolating and sequencing a taxonomically informative genomic region, such as the ITS (internal transcribed spacer), from DNA extracted from a metagenomic sample, and then classifying these sequences to determine which organisms are present. This paper announces the availability of the Warcup ITS training set and shows how it can be used with the Ribosomal Database Project (RDP) Bayesian Classifier to rapidly and accurately identify fungi using ITS sequences. The classifications can be down to species level and use conventional literature-based mycological nomenclature and taxonomic assignments.

Key words: Ecogenome, fungal classification, fungi, identification, ITS

INTRODUCTION

Fungi are key participants in many ecological processes and communities. They perform essential roles

in the decomposition of organic matter and in nutrient cycling and exchange throughout the landscape. Fungi also form critical symbioses with algae and most of the higher plants and play important roles in the food web. Current estimates of diversity within the fungal kingdom suggest that there may be more than 5 million species worldwide (Blackwell 2011). Modern molecular methods can be used to select and sequence taxonomically informative regions from environmental samples, generating millions of sequences at low cost. Classifying these sequences using traditional BLAST queries is problematic both because of the time this would take and the uncertain reliability of the entries in the reference databases. The Ribosomal Database Project (RDP) provides a fast and effective classifier for bacterial and archaeal organisms based on the 16S ribosomal RNA gene (Wang et al. 2007). This classifier has been adapted for the identification of fungal sequences using the 18S (Quast et al. 2013) and 28S rRNA subunits (Liu et al. 2012, Quast et al. 2013). The fungal research community now has adopted the ITS (internal transcribed spacer) region as the marker of choice for typing fungal cultures and the exploration of fungal diversity (Koljalg et al. 2013). This region is both ubiquitous and highly variable in both sequence and length, making it a good marker gene for identification, but this same variability has proven to be somewhat challenging for the molecular phylogenetic methods that have been so powerfully applied to other groups with sequences from 16S rRNA and other structured genes. The development of the DOE SFA training set (US Department of Energy Science Focus Area) demonstrated that the RDP Classifier, a tool developed for identification as opposed to phylogenetic inference, could be used with an ITS-derived training set to accurately identify fungal organisms (Koljalg et al. 2013, Porras-Alfaro et al. 2014).

This paper announces the availability of the Warcup ITS training set, named in honor of Australasian mycologist John “Jack” Warcup (1921–1998). This training set can be used with the RDP classifier to rapidly and accurately identify fungi using their ITS sequences, often down to species, and the classifications returned reflect conventional literature-based mycological nomenclature and taxonomy.

The development of the initial Warcup fungal ITS training set started with the 343 809 ITS sequences contained in the UNITE+INSD dataset (Abarenkov

Submitted 8 Nov 2014; accepted for publication 21 Sep 2015.

¹ Corresponding author. E-mail: nai.tran-dinh@csiro.au

TABLE I. Taxonomic coverage of the Warcup training set

Phylum	Classes	Orders	Families	Genera	Species
Ascomycota ^a	16	66	192	1048	5453
Basidiomycota ^a	13	47	134	495	3397
Blastocladiomycota	1	1	1	1	1
Chytridiomycota ^a	2	5	7	10	19
Fungi_Incertae sedis	1	1	1	2	3
Glomeromycota	1	4	9	16	97
Neocallimastigomycota	1	1	1	2	2
Zygomycota ^a	5	6	19	48	143

^a Missing orders: Ascomycota: Arachnomycetales, Boliniales, Coronophorales, Gyalectales, Halosphaeriales, Hysteriales, Laboulbeniales, Lichinales, Medeolariales, Meliolales, Microthyriales, Mytilinidiales, Protomycetales, Spathulosporales, Trichotheliales, Trypetheliales. Basidiomycota: Attractiellales, Classiculales, Cryptomycocolacales, Doassansiales, Geogefischeriales, Helicobasidiales, Mixiales, Naohideales, Pachnocybales, Septobasidiales, Spiculogloeales, Urocystales. Chytridiomycota: Geoglossales. Zygomycota: Dimargaritales, Harpellales, Zoopagales.

et al. 2010), as downloaded from <http://unite.ut.ee> (release 10 May 2013). This dataset contained fungal ITS sequences and classifications from both the INSD database (<http://www.insdc.org/>) and the UNITE set, and each sequence had been assigned an organism name and taxonomy from one or both of the INSD or UNITE databases. Any sequences without taxonomic or lineage information were discarded because the RDP classifier requires all sequences in a training set to have an assigned taxonomy. Sequences from the nonfungal phyla (principally Arthropoda and Microspora spp.) also were discarded. Any sequences that were shorter than 250 bp were also not included, as were those that did not contain a part of the 5.8S region. Some sequences also contained 18S or 28S rDNA regions, and these were removed with Fungal-ITSExtractor (Nilsson et al. 2010). Any sequences containing ambiguous bases (Ns) also were discarded.

The UNITE and INSD taxonomies for all the sequences in the set then were harmonized as follows:

- When a sequence was classified in only one of the taxonomies, then that classification was given to the sequence.
- When a sequence was classified in both the UNITE and INSD taxonomies, the more taxonomically informative classification was chosen.
- Identical sequences with differing taxonomies were resolved by majority vote.
- Singleton sequences were removed as these were deemed to be potentially unreliable. Duplicate sequences deposited from the same study were removed for the same reason.
- Sequences supposedly from the same species were clustered together, and only sequences with an identity greater than 98% were accepted.

The initial release of the Warcup training set covers 9115 species with 24 447 ITS sequences. The coverage of the fungal taxonomic space is shown (TABLE I). The training set covers all fungal phyla and most orders. Thirty-two orders have been identified as missing or poorly represented in the training set. The authors invite community efforts to improve coverage of the Warcup ITS training set by suggesting additional curated ITS sequences and to incorporate the Warcup improvements into the primary ITS repositories such as UNITE.

CLASSIFICATION PERFORMANCE AND ACCURACY

The Warcup ITS training set was evaluated by running leave-one-sequence-out and leave-one-taxon-out tests (Wang et al. 2007, Liu et al. 2012). We also calculated taxa similarity metrics (Sab scores), both within taxa and between taxa. The results from these tests were compared to those obtained through use of a UNITE-based set and the DOE SFA training set (Porras-Alfaro et al. 2014). The UNITE set consists of 145 019 UNITE core sequences (excluding chimeric and low quality) for each dynamic species hypothesis, provided by Kessy Abarenkov of UNITE on 4 Jul 2014. Each sequence in the UNITE set had two species designations: a UNITE species hypothesis accession code number (Koljalg et al. 2013); and a more traditional UNITE taxon name based on Index Fungorum. Both the Warcup and UNITE training sets support classification to species, while the DOE SFA demonstrator training set only supports classification to genus. The full report on the testing process and the methods used is available at http://rdp.cme.msu.edu/download/posters/fungalITSreport_062014.pdf

We generated two RDP training sets from the UNITE training set, the first (UNITE_name) by grouping sequences into terminal taxa using the UNITE

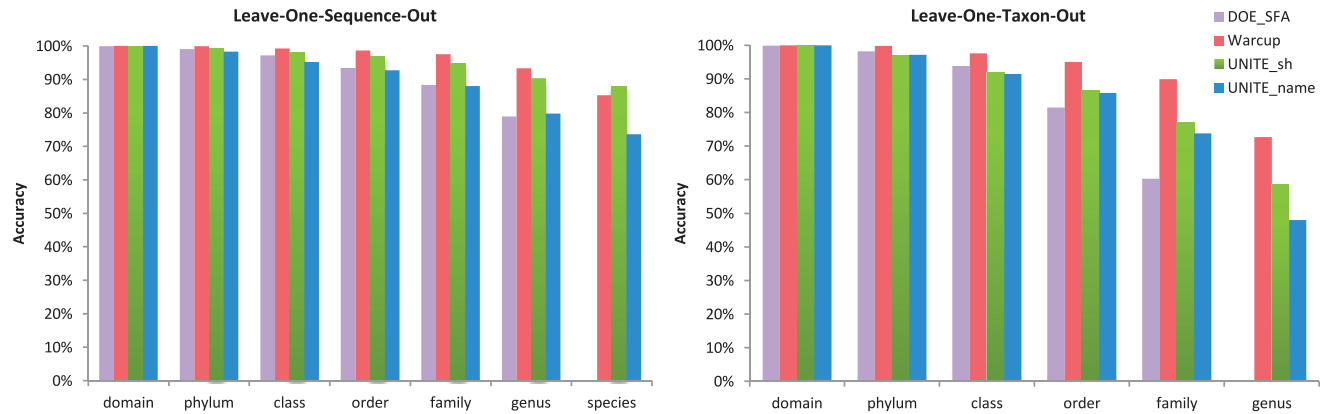


FIG. 1. Classification accuracy at each major taxon rank from leave-one-out testing. The RDP Classifier was trained on the each of the four fungal ITS training sets. No bootstrap cutoff was applied in the accuracy calculation. Note that the DOE_SFA set does not classify to species rank. Left: leave-one-sequence-out. Right: leave-one-taxon-out.

taxon name and a second (UNITE_sh) with a concatenation of the UNITE species hypotheses accession code number and the UNITE terminal taxon name to group sequences into terminal taxa (UNITE_sh). All tests were run with both of these UNITE-based training sets. The UNITE_name training set is the most comparable to the Warcup training set, in that it uses a more traditional taxonomy for its sequences rather than returning less-familiar species hypothesis classifications.

We used Sab scores as a measure of similarity between sequences. The Sab score was calculated as the number of unique k-mers of size 8 shared by the two sequences divided by the lowest number of unique 8-mers in either of the two sequences. This Sab score method is similar to the one used by the RDP SequenceMatch tool except the latter uses k-mers of size 7. Warcup revealed the highest and tightest similarity within species with median at 96%. UNITE_sh revealed a median similarity within species of 90% but with a large range from 72% (2nd percentile) to 99% (98th percentile). UNITE_name had the lowest median similarity of 37% within species. DOE_SFA does not group at species rank, but the median Sab score within genera was 56% and dropped to 31% among families.

In leave-one-sequence-out testing we iteratively chose a sequence from each of the four starting sets, rebuilt the corresponding training set without it and then classified the chosen sequence with this new training set. We then compared the taxonomy assigned to this sequence to its original taxonomic label to measure classification accuracy. All Warcup and DOE_SFA sequences were used in these tests, and one sequence from each species was chosen randomly for testing from the UNITE_sh and

UNITE_name sets due to their size. Warcup and UNITE_sh performed equally well in this test, revealing 85% and 88% accuracy at species level and 93% and 90% at genus, respectively. The more conventional UNITE_name training set did not perform as well, returning a species-level accuracy of only 74% and 80% at genus level. The DOE_SFA training set returned a genus-level accuracy of 79%.

The leave-one-taxon-out testing was similar to the leave-one-sequence-out testing except for each chosen sequence, all other sequences assigned to the same lowest taxon (genus for the DOE_SFA training set and species for the other three training sets) also were removed from the training set. This test was intended to determine whether the classifier could correctly assign a sequence to an appropriate higher level taxon when the corresponding terminal taxon was not present in the training set, simulating a novel query taxon. Leave-one-taxon-out testing revealed that, if the entire species was removed from the training set, use of the Warcup set resulted in the correct genus being returned for 73% of the sequences, with UNITE_sh coming next at 58% and UNITE_name returning 47%. With the target species removed, the Warcup set gave the correct family assignment 90% of the time, followed by UNITE_sh at 77%, UNITE_name at 74% and 60% for the DOE_SFA set. The results from the leave-one-sequence-out and leave-one-taxon-out tests are summarized (FIG. 1).

The Warcup and UNITE_sh training sets were compared for their ability to assign taxonomic rank to authentic NGS sequences. We downloaded the large fungal ITS dataset used in Tedersoo et al. (2014) and clustered these into 29 658 OTUs (SUPPLEMENTARY MATERIAL 1) using the USEARCH pipeline (Edgar

2010). A representative sequence for each OTU then was classified using the RDP classifier with both the Warcup and UNITE_sh training sets (SUPPLEMENTARY MATERIAL 2). In broad terms Warcup identified almost twice as many taxa at species level and ~35% more taxa at genus level (SUPPLEMENTARY MATERIAL 3). Above genus level the two training sets performed similarly. It is noteworthy that data on subphylum and subclass were not available from UNITE_sh.

We also did some simple speed tests, and the RDP classifier took 80 s to classify a test set of 1000 near-full length ITS sequences using the UNITE training sets on a single CPU on 3.2 GHz Intel Core i5 processor. The Warcup training set was approximately twice as fast on this same test, with performance being roughly proportional to the number of species (terminal taxa) in each of the training sets.

AVAILABILITY OF WARCUP ITS TRAINING SET AND FUTURE DIRECTIONS

The RDP Classifier in conjunction with the Warcup fungal ITS training set is a reliable and accurate tool for identification of a broad range of fungi. Researchers can use the Web interface provided at <http://rdp.cme.msu.edu/classifier/classifier.jsp> to classify small sets of sequences. For larger libraries of sequences both the RDP classifier and the Warcup fungal ITS training set can be downloaded from the RDP repositories (<http://sourceforge.net/projects/rdp-classifier/> and <https://github.com/rdpstaff>) and analyses performed locally.

The main goals in the development of the Warcup fungal ITS training set were to be both broad in coverage and accurate in classification. These two goals are in inherent conflict, and we always favored accuracy over coverage, dropping sequences for poorly studied organisms if we could not be confident that they were accurate. The intent of the development of the Warcup fungal ITS training set was to provide a rapid and reliable classification of fungi to species, using validated ITS sequences from sequence repositories such as UNITE and INSD and returning conventional literature-based mycological nomenclature and taxonomic assignments.

We invite the fungal community to participate in further improving both the accuracy and coverage of the Warcup fungal ITS training set by validating the current classifications and providing additional well-curated ITS sequences to extend its taxonomic range. Regular updates of the Warcup fungal ITS training set will incorporate additional validated ITS sequences from current and novel fungal species. These updates will be released in conjunction with RDP updates.

ACKNOWLEDGMENTS

Development of the DOE SFA ITS training set was supported by the US Department of Energy, Office of Science, Biological and Environmental Research Division through a Science Focus Area grant to CRK (2009LANLF260). The RDP is supported by the Office of Science (Biological and Environmental Research), US Department of Energy (DE-FG02-99ER62848, DE-SC0010715 and DE-FC02-07ER64494). The initial development of the Warcup training set was supported by the CSIRO office of the chief executive through an OCE honors scholarship awarded to Vinita Deshpande. We thank Kessy Abarenkov for providing the UNITE dataset.

LITERATURE CITED

- Abarenkov K, Henrik Nilsson R, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjølner 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 Phytol* 186:281–285, doi:10.1111/j.1469-8137.2009.03160.x
- Blackwell M. 2011. The fungi: 1, 2, 3 ... 5.1 million species? *Am J Bot* 98:426–38, doi:10.3732/ajb.1000298
- Edgar RC. 2010. Search and clustering hundreds of times faster than BLAST. *Bioinformatics* 1–3, doi:10.1093/bioinformatics/btq461
- Koljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, Bates ST, Bruns, Thomas D, Bengtsson-palme J, Callaghan TM, Douglas B, Drenkhan T, Eberhardt U, Duenas M, Grebenc T, Griffith GW, Hartmann M, Kirk PM, Kohout P, Larsson E, Lindahl BD, Lucking R, Martin MP, Matheny PB, Nguyen NH, Niskanen T, Oja J, Peay KG, Peintner U, et al. 2013. Toward a unified paradigm for sequence-based identification of fungi. *Mol Ecol* 22:5271–5277, doi:10.1111/mec.12481
- Liu K-L, Porras-Alfaro A, Kuske CR, Eichorst SA, Xie G. 2012. Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes. *Appl Environ Microbiol* 78:1523–33, doi:10.1128/AEM.06826-11
- Nilsson RH, Veldre V, Hartmann M, Unterseher M, Amend A, Bergsten J, Kristiansson E, Ryberg M, Jumpponen A, Abarenkov K. 2010. 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 Ecol* 3:284–287, doi:10.1016/j.funeco.2010.05.002
- Porras-Alfaro A, Liu K-L, Kuske CR, Xie G. 2014. From genus to phylum: large-subunit and internal transcribed spacer rRNA operon regions show similar classification accuracies influenced by database composition. *Appl Environ Microbiol* 80:829–40, doi:10.1128/AEM.02894-13
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41:D590–596, doi:10.1093/nar/gks1219

- Tedersoo L, Bahram M, Polme S, Koljalg U, Yorou NS, Wijesundera R, Ruiz LV, Vasco-Palacios AM, Thu PQ, Suija A, Smith ME, Sharp C, Saluveer E, Saitta A, Rosas M, Riit T, Ratkowsky D, Pritsch K, Poldmaa K, Piepenbring M, Phosri C, Peterson M, Parts K, Partel K, Otsing E, Nouhra E, Njouonkou AL, Nilsson RH, Morgado LN, et al. 2014. Global diversity and geography of soil fungi. *Science* (80–)346:1256688–1256688, doi:10.1126/science.1256688
- Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267, doi:10.1128/AEM.00062-07