# RDP Classifier Fungal ITS Warcup Training Set V2

June 2016

Deshpande, V., Q. Wang, P. Greenfield, M. Charleston, A. Porras-Alfaro, C. R. Kuske, J. R. Cole, D. J. Midgley, and N. Tran-Dinh. 2015. 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

Version 2 of the Warcup training set provides more accurate and consistent classification of fungal ITS sequences. These improvements result from an effort to find and correct a number of small errors and inconsistencies that were present in the reference data used to build the V1 training set.

The principles we followed when creating the V2 training set: 1) that closely related organisms should have similar ITS sequences, and 2) that similar sequences should have similar classifications. The sequences in the training set were repeatedly clustered at 97% or greater sequence identity, and these clusters were checked for consistency. We used the *-cluster_fast* function of USEARCH (Edgar, 2010). These clusters were examined for taxonomic consistency, and, where inconsistencies were identified, the errant classifications were corrected. In broad terms, the types of inconsistencies fell into three categories: Category 1) erroneously labelled sequences; Category 2) synonymies; and Category 3) anamorph/teleomorph inconsistencies in the genus name.

Correcting inconsistencies from Category 1 was generally straightforward, using BLAST against GenBank to identify the aberrant sequence (or sequences) within a cluster. Sequences identified by this method which added no useful information to the cluster and were misleadingly named were removed from the training set. Category 2 errors were resolved using a combination of strategies, generally using both MycoBank and Index Fungorum, but also primary literature, as reference sources for resolving synonymies with a preference for the current valid name. Inconsistencies resulting from Category 3 errors were harder to resolve. Most commonly we preferred the teleomorph name to the anamorph name, though for some species where no teleomorphs are known we have retained the anamorph name. This latter strategy is essentially in keeping with the One Fungus = One Name best practice.

Finally, the newly developed training set (V2) was classified against itself, and this showed significant improvements in consistency and accuracy over Warcup V1. The vast majority of sequences (99.9%) now classify with over 80% confidence at the genus level and (94.5%) now classify with over 80% confidence at the species level.

The next step in the development of the Warcup training set will be to further improve the species-level classifications and to add sequences and classifications in those groups of fungi that are poorly represented in the available reference sets.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19): 2460-2461. doi: 10.1093/bioinformatics/btq461