UNITE fungal ITS database

UNITE version 7, release date October 5, 2014

This is the non-redundant version of the UNITE+INSDC fungal ITS database as presented in Kõljalg et al. 2013. UNITE mirrors the International Nucleotide Sequence Database Collaboration so as to comprise all public Sanger-derived fungal ITS sequences (and a fair number of NGS sequences too). The present release features only ITS sequences that are more or less full length. It can be used as reference corpus for studies employing only ITS1, only ITS2, full-length ITS sequences, or anything in between.

Clustering redundancy information (see Kõljalg et al. 2013 for details)

UNITE clusters all fungal ITS sequences to approximately the species level. All OTUs composed of two or more sequences are referred to as species hypotheses (SHs). The QIIME release of UNITE comprises the representative/reference sequences of all SHs in UNITE. By default, a sequence from the most common sequence type of an SH is chosen as representative sequence for that SH. The choice is thus done by the clustering software. Registered users can manually override the choice of what sequence is chosen to represent an SH in UNITE to account for expert knowledge, sequences from type material etc. No single similarity threshold will accurately reflect the species level throughout the fungal kingdom. Thus, while the present release does include both the representative/reference sequences of all SHs after 97% and 99% similarity clustering, we feel that the core release is the "dynamic" representative/reference sequence file. Here, taxonomic experts on various fungal lineages have gone through their lineage of expertise and indicated at what level each species should be delimited. Thus, some SHs were delimited at 97%, some at 97.5%, some at 98% and so on up to 100%. The default threshold value of the dynamic file is 98.5%.

In addition to this readme, there are six (plus six) files in this archive

sh_taxonomy_qiime_ver7_97_01.08.2015.txt
sh_refs_qiime_ver7_97_01.08.2015.fasta (17,421 sequences)
sh_taxonomy_qiime_ver7_99_01.08.2015.txt
sh_refs_qiime_ver7_99_01.08.2015.fasta (23,201sequences)
sh_taxonomy_qiime_ver7_dynamic_01.08.2015.txt
sh_refs_qiime_ver7_dynamic_01.08.2015.fasta (22,774 sequences)
developer (folder)

The FASTA files are the representative sequences for 97% clustering, 99% clustering, and the dynamic file with varying threshold values.

The TAX files are the corresponding taxonomy files with the following syntax:

Identifier (tab)
k__Fungi;p__Phylum;c__Class;o__Order;f__Family;g__Genus;s__Species

Whereas the "Species" column represents the lowest assignment available for that SH, it is not always a full species name. Partial species names (e.g., "Candida sp.") and other expressions of uncertainty (e.g., "Unidentified basidiomycete") are not uncommon – this reflects the current state of molecular identification of fungi.

The SH identifiers are resolved through URIs (e.g., http://unite.ut.ee/sh/SH000011.07FU, which can be followed for information on constituent sequences, ecology, and geography. The SHs remain traceable over time, such that they represent a means for uniform communication on all species of fungi – with or without a full Latin name – known from at least two ITS sequences.

All sequences in the above files were run through ITSx (Bengtsson-Palme et al. 2013) to remove any larger parts of SSU (18S) and LSU (28S). For most applications relating to ITS-based molecular identification of fungi, the SSU and LSU are unwanted in that they skew clustering attempts and sequence similarity searchers. The "developer" folder contains the same six files in their untrimmed

state, that is, exactly as the sequence data are in INSD. They are provided in the interest of completeness – our feeling is that they should not be used for everyday QIIME use.

UNITE offers extended capacities for web-based, third-party sequence annotation to its registered users (and registration is free). Thus, anyone in the position to improve the usefulness and metadata of the sequences in UNITE – through, e.g., replacing incorrect names, adding names to unidentified sequences, highlighting particularly reliable, high-quality sequences (or excluding compromised sequences), and/or adding data on ecology and geography to sequences – are welcome to join. All such changes will be incorporated in the next QIIME release of UNITE. Several annotation efforts of various groups of fungi are ongoing. Three such efforts have been published so far (Tedersoo et al. 2011; Nilsson et al. 2012; Kõljalg et al. 2013). That said, some lineages of the fungal tree of life are better vetted than others in UNITE.

This is the first full release of UNITE for QIIME. It is released in the hope that QIIME users will let us know of any problems, bugs, or shortcomings.

URL: http://unite.ut.ee Contact: kessy.abarenkov@ut.ee - henrik.nilsson@bioenv.gu.se

REFERENCES

Bengtsson-Palme et al. 2013. ITSx: Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for use in environmental sequencing. Methods in Ecology and Evolution 4: 914-919.

Kõljalg et al. 2013. Towards a unified paradigm for sequence-based identification of fungi. Molecular Ecology 22: 5271-5277.

Nilsson et al. 2012. Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. MycoKeys 4: 37-63.

Tedersoo et al. 2011. Tidying up International Nucleotide Sequence Databases: ecological, geographical, and sequence quality annotation of ITS sequences of mycorrhizal fungi. PLoS ONE 6: e24940.