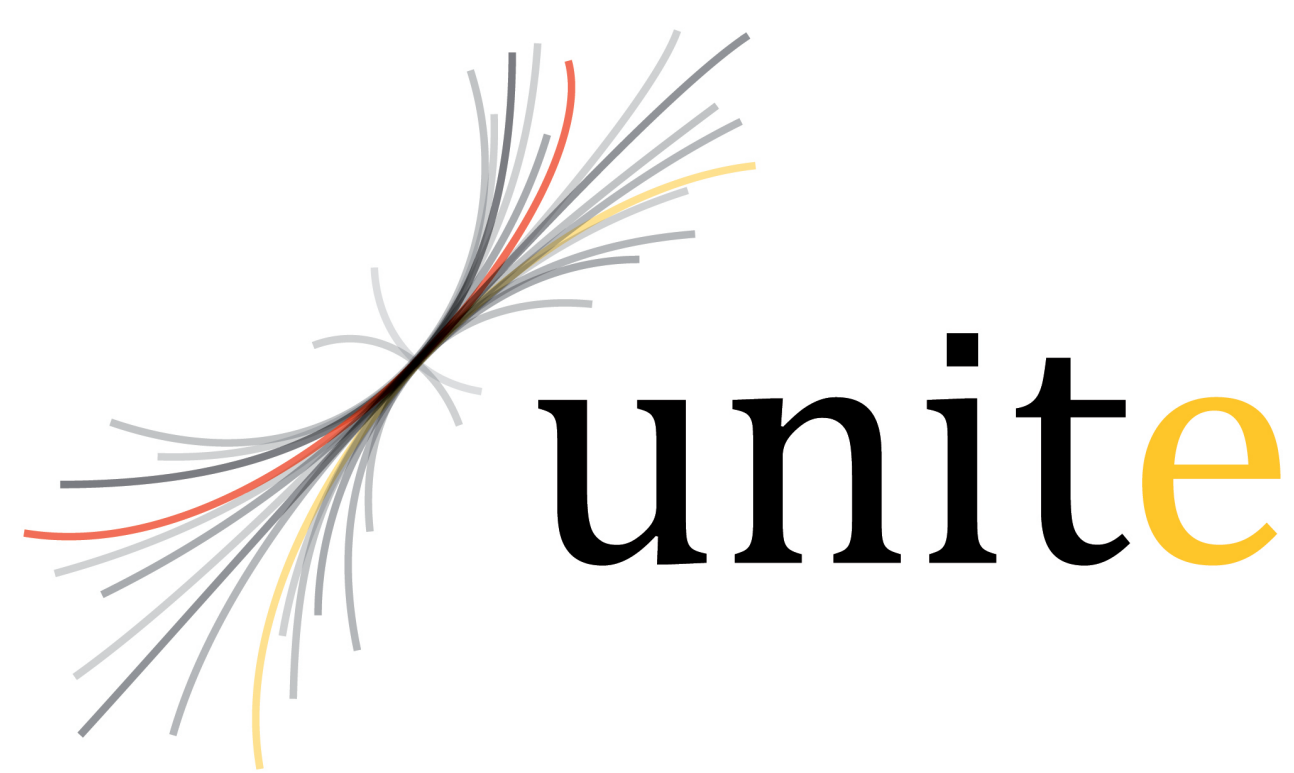




Unified system for the DNA based fungal species linked to classification

Kessy Abarenkov¹, Urmaz Kõljalg^{1,2}, R. Henrik Nilsson³, Andy F. S. Taylor⁴, Karl-Henrik Larsson⁵, UNITE Community⁶

1. Natural History Museum, University of Tartu; 2. Institute of Ecology and Earth Sciences, University of Tartu; 3. Department of Biological and Environmental Sciences, University of Gothenburg; 4. The James Hutton Institute, Craigiebuckler, Aberdeen; 5. Natural History Museum, University of Oslo; 6. <https://unite.ut.ee/acknow.php>

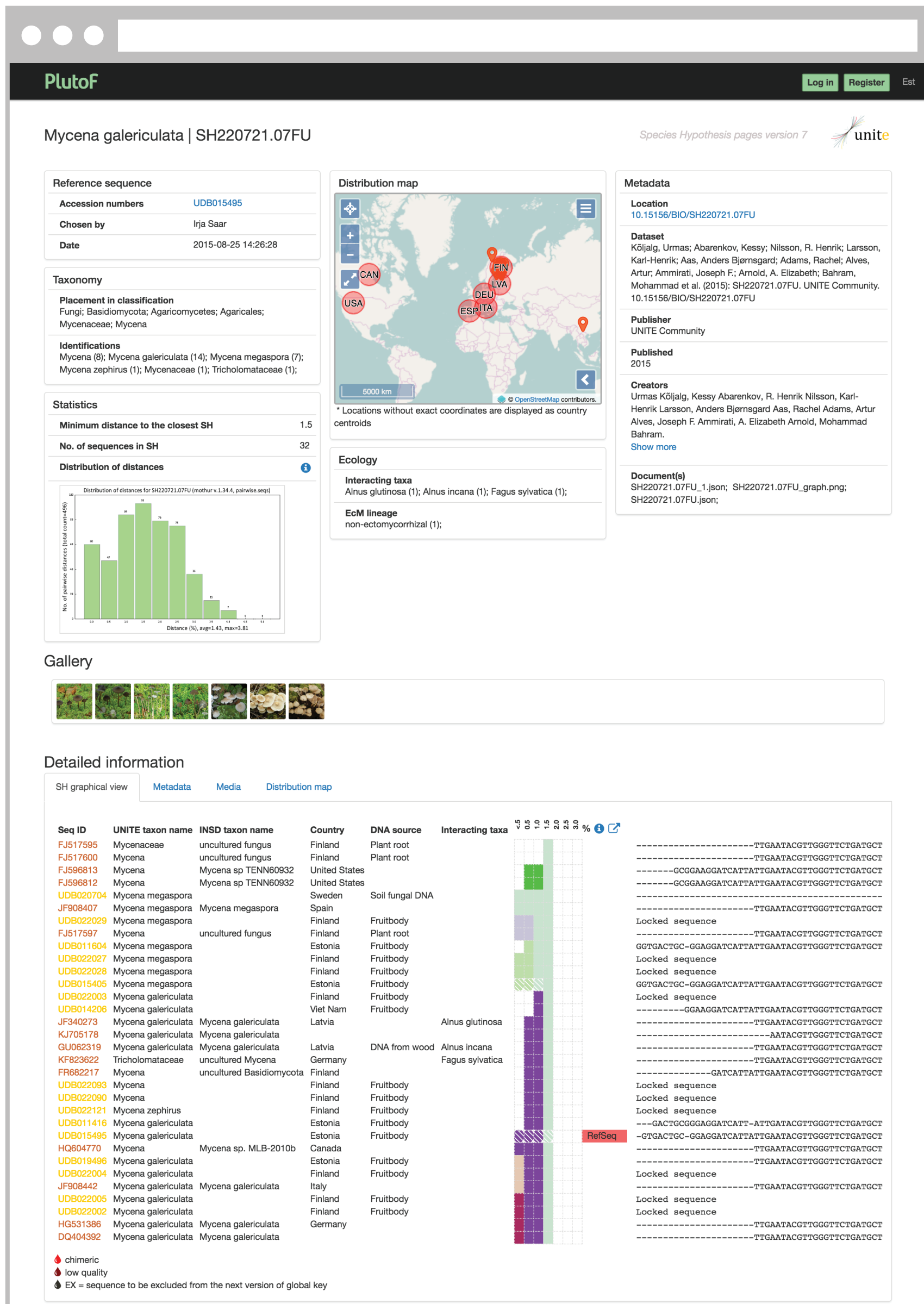
Features of UNITE:

1. UNITE species hypotheses (SH)
2. Resources
3. Search and Statistics
4. Third-party sequence annotation
5. Analysis tools

1 UNITE SH-s

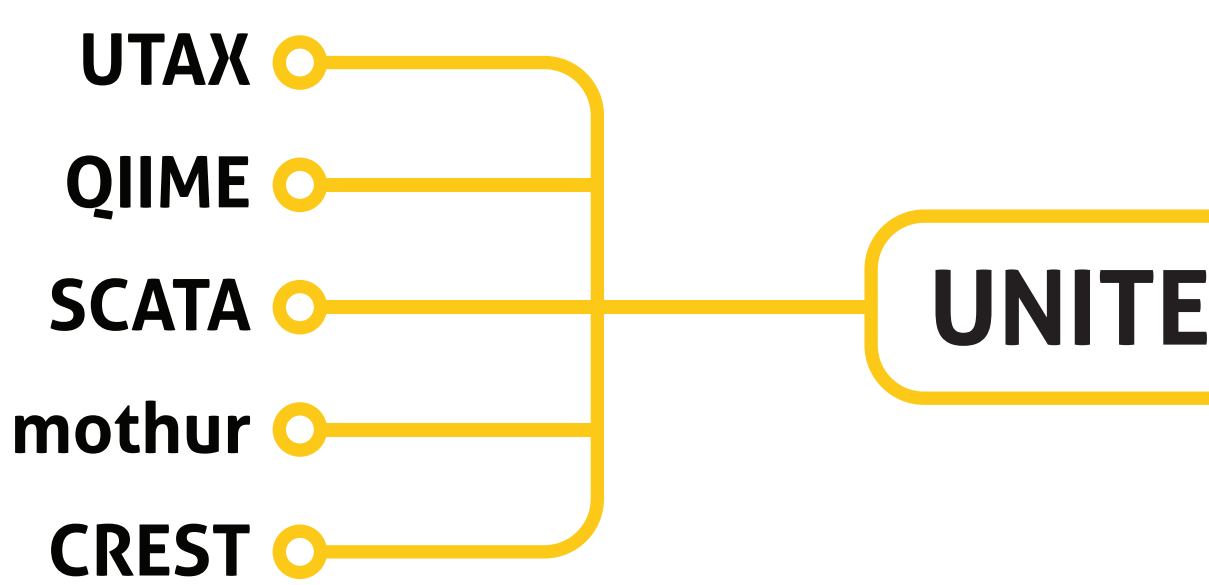
UNITE SH-s are a unique way how to communicate fungal species using stable identifiers. ITS sequences submitted to UNITE and International Nucleotide Sequence Databases (INSD: GenBank, ENA, DDBJ) are grouped into SH-s by following a two-step clustering process, first at the subgeneric/generic level, then at a species level on different distances to the closest SH (e.g. 3%, 2.5%, etc.). All SH-s are linked to the Linnaean classification and they carry a stable accession code and digital object identifier (doi, Figure 1) making it possible to cite and communicate them in published studies.

Figure 1. Example DOI page



2 Resources

UNITE provides reference datasets with annotated sequences and SH-s in various formats for downloading and using locally or in a range of applications (e.g. QIIME, mothur, SCATA).



3 Search and Statistics

Custom search functions and unique views to fungal barcode sequences include extended search filters (e.g. source, locality, habitat, traits) for sequences and SH-s, interactive maps (Figure 2) and graphs (Figure 3), and views to the largest unidentified sequence clusters formed by sequences from multiple independent ecological studies, and for which no metadata currently exists (Figure 4).

Figure 2. Global map of sequences and SH-s

Distribution of UNITE SH-s (1.5% distance) based on ITS sequences carrying location data (71% of sequences included in SH-s). Coloured dots refer to sequences in SH-s unique to one continent. Gray dots refer to sequences in SH-s shared between two or more continents.

Figure 4. Top 50 most wanted search page and view

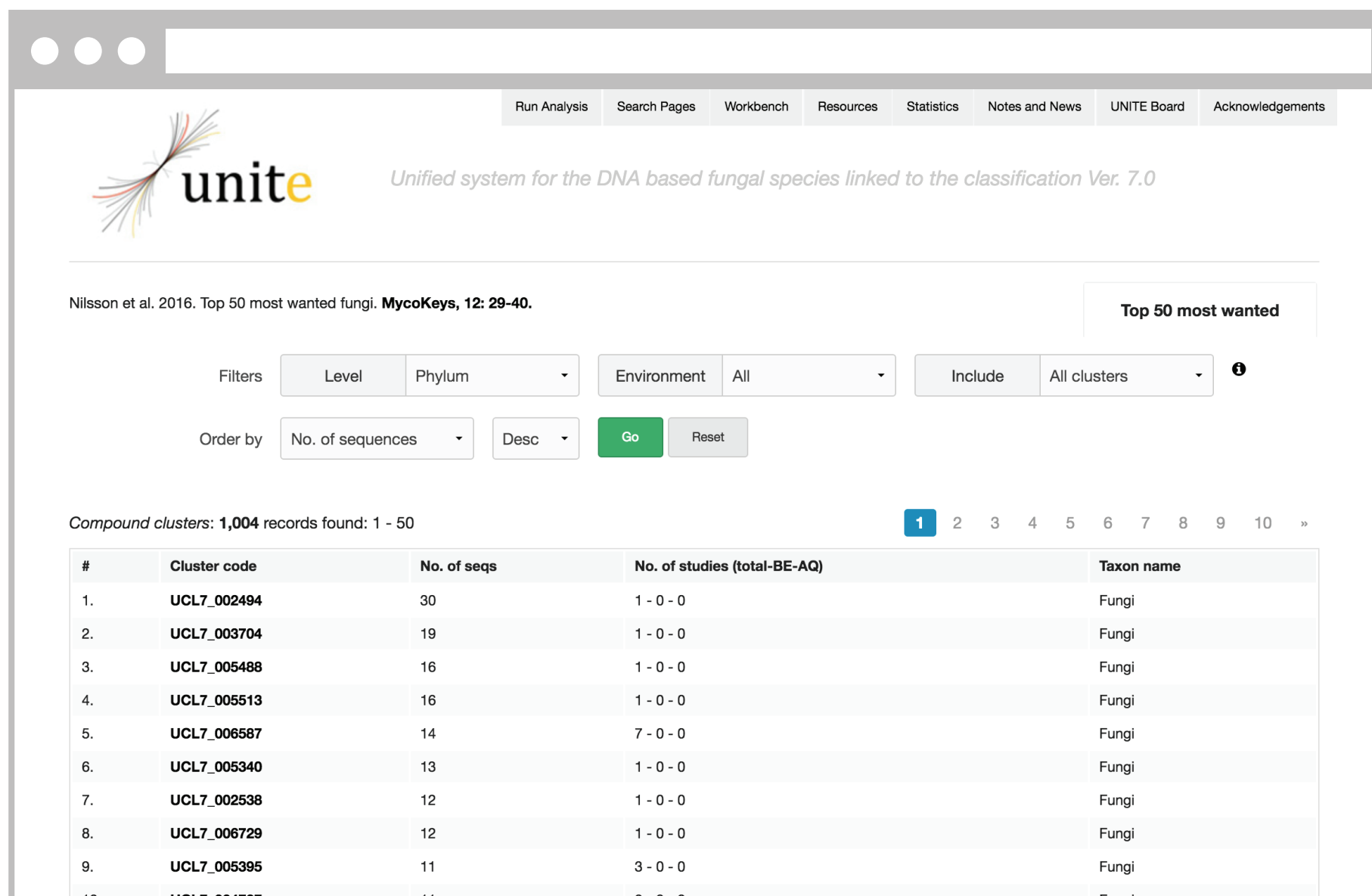
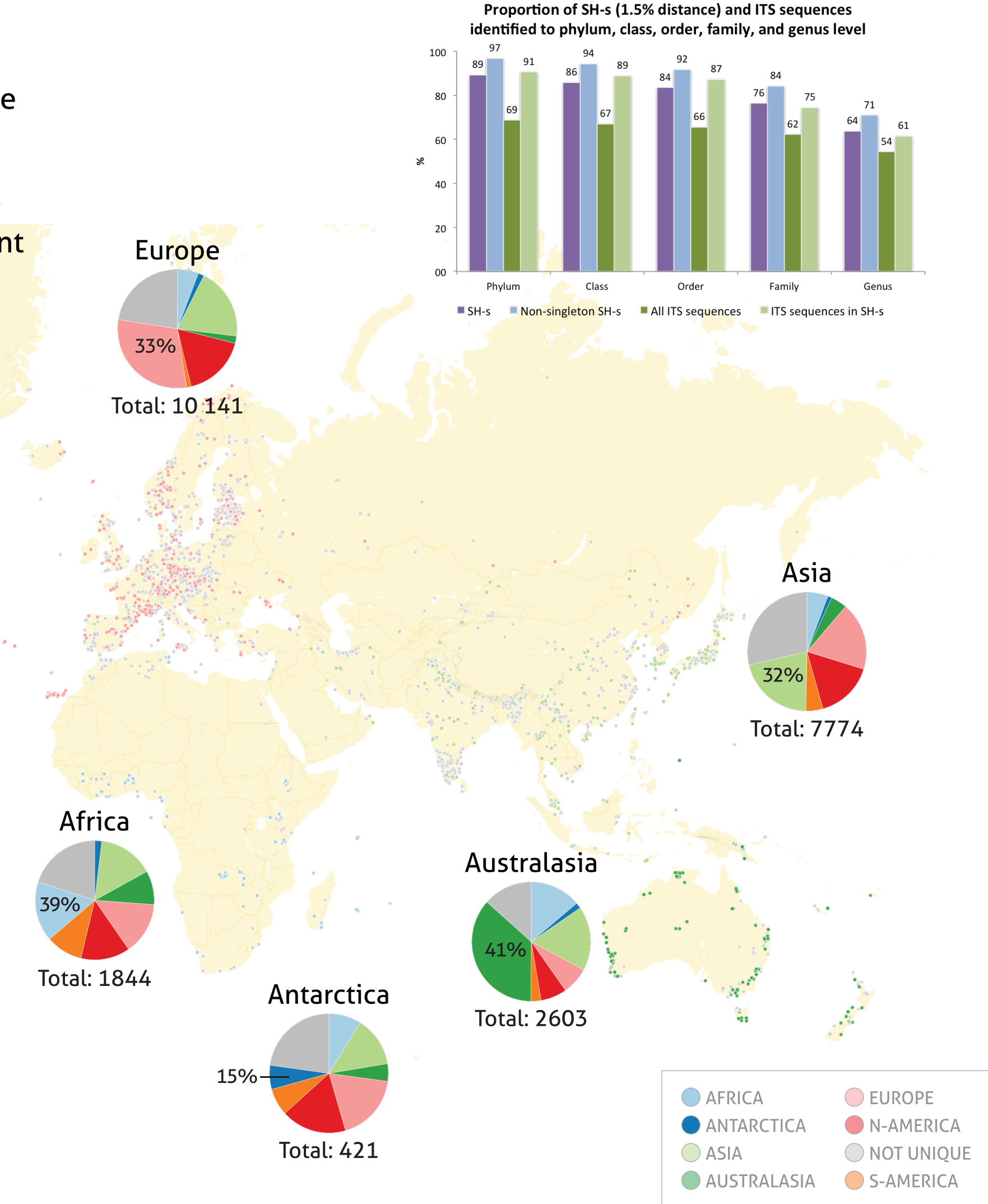


Figure 3. Identified vs unidentified



4 Third-party sequence annotation

Tools for the third-party sequence annotation are used by the UNITE Community to annotate and improve the quality of fungal ITS sequences in INSD. Local copy of INSD dataset is updated on a regular basis. Any third-party annotation added to the INSD dataset (e.g. locality, habitat, source, traits, taxon identifications, and interacting taxa) is made publicly available to the research community through web services and on UNITE homepage.

Table 1. List of annotation efforts

2016	Build mycobiome sequence metadata annotation workshop (Abarenkov et al. 2016)
2016	Top 50 most wanted – annotating the “dark fungal diversity” (Nilsson et al. 2016)
2014	Improving ITS sequence data for identification of plant pathogenic fungi (Nilsson et al. 2014)
2013	UNITE jamboree – fungal ITS sequence annotation workshop (Kõljalg et al. 2013)
2011	Ecological, geographical and sequence quality annotation of ITS sequences of mycorrhizal fungi (Tedersoo et al. 2011)

Table 2. Statistics on metadata annotations

22,217	taxonomic re-annotations
56,720	specifications of ectomycorrhizal lineage
101,035	specifications of country of collection
52,720	specifications of host and interacting taxa
2,977	chimeric and 7,224 low read quality sequences found
6,281	reference sequence specifications
9,560	sequences with specimen/culture metadata annotated
4,518	sequences linked to type specimens/cultures

5 Analysis tools

UNITE provides variety of analysis tools including, for example, mass-BLASTER for blasting hundreds of sequences in one batch, ITSx for detecting and extracting ITS1 and ITS2 regions of ITS sequences from environmental communities, or ATOSH for assigning your unknown sequences to SH-s. Analysis can be run on UNITE homepage (Figure 5) or PlutoF workbench (<https://plutof.com>).

Figure 5. Screenshot of UNITE analysis page

