Species Hypotheses

PlutoF user manual

This section is designed for the UNITE community
(Last updated: 25.02.2016)

PlutoF

Table of Contents

1. Introduction

PlutoF provides cloud database and computing services for the biology and related disciplines. The purpose of the platform is to provide synergy through common modules for the taxon occurrences, classifications, geography, projects, agents, analytical tools, etc.

This section of the PlutoF manual describes work environment for the Species Hypotheses (SH). More information about "What is Species Hypothesis?" and "What are Reference and Representative sequences?" can be found in terms section in this document, on UNITE homepage (https://unite.ut.ee), and from Kõljalg et al. 2013 (http://onlinelibrary.wiley.com/doi/10.1111/mec.12481/abstract).

NB! Current user manual should to be used together with the PlutoF main user manual (https://plutof.ut.ee/#/manual) that we kindly ask you to read beforehand. Main manual explains the general working principles of the workbench that are not fully covered in this manual.

PlutoF hosts all datasets for the UNITE database (https://unite.ut.ee). Users belonging to *"Sequence annotations"* and *"UNITE Species Hypotheses"* workgroups have access to third-party annotation system of INSD sequence datasets, and working with UNITE SH-s, respectively.

Workgroups can be joined by sending join request to the corresponding workgroup –

Settings => Workgroups => *Search workgroup "UNITE Species Hypotheses"* => *Send Join request*

2. Terms

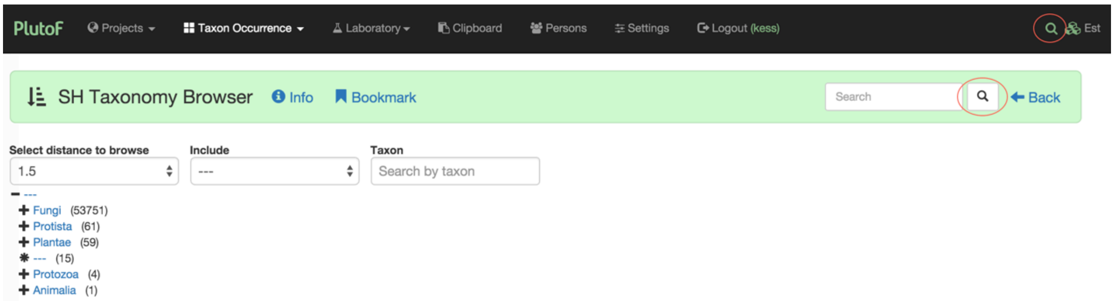
Biological sample – Any physical sample which includes DNA of organism(s). For example, living or collection specimen, soil, water, air, blood, tissue, etc.

Reference sequence (RefS) serves as a name anchor for the species hypothesis and is chosen by the expert. It may originate from any biological sample, viz. herbarium specimen, living culture, soil, water, air, tissue of other organism, etc. RefS is utilised in the scientific communication where identification of organism is based on DNA sequences.

Representative sequence (RepS) serves as a name anchor for the species. It is chosen automatically for all species hypotheses n all clusters based on identical criteria. RepS allows to name and communicate species until RefS becomes available for given species.

3. Finding SH-s using search module

SH search module can be found by following *Laboratory => Molecular Lab => SH Taxonomy Browser* by pasting SH code in the search box and clicking search icon (Figure 1). Alternatively, search module can be entered using the shortcut (look for magnifier glass) on page header.

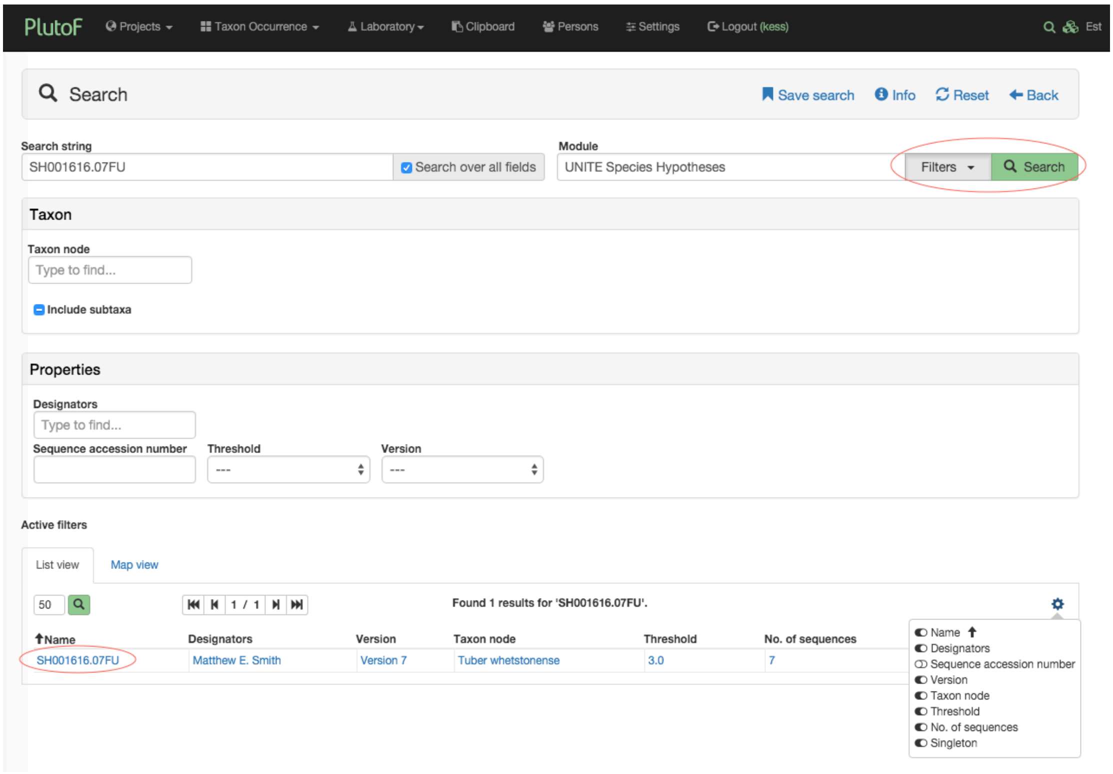


Select distance to browse: 1.5
Include: ---
Taxon: Search by taxon

- ---
  - Fungi (53751)
  - Protista (61)
  - Plantae (59)
  - --- (15)
  - Protozoa (4)
  - Animalia (1)

Figure 1. Entering search module.

User will be directed to advanced search view where additional search filters can be specified and search results ordered (Figure 2). By clicking on search result record, SH detail-view will be opened. “Back” links on navigation bar are a preferred way to easily navigate between all pages in the system.



Search string: SH001616.07FU
☑ Search over all fields
Module: UNITE Species Hypotheses

Taxon

Taxon node: Type to find...
Include subtaxa

Properties

Designators: Type to find...
Sequence accession number:
Threshold: ---
Version: ---

Active filters

50 | 1 / 1

Found 1 results for 'SH001616.07FU'.

Name	Designators	Version	Taxon node	Threshold	No. of sequences
SH001616.07FU	Matthew E. Smith	Version 7	Tuber whetstonense	3.0	7

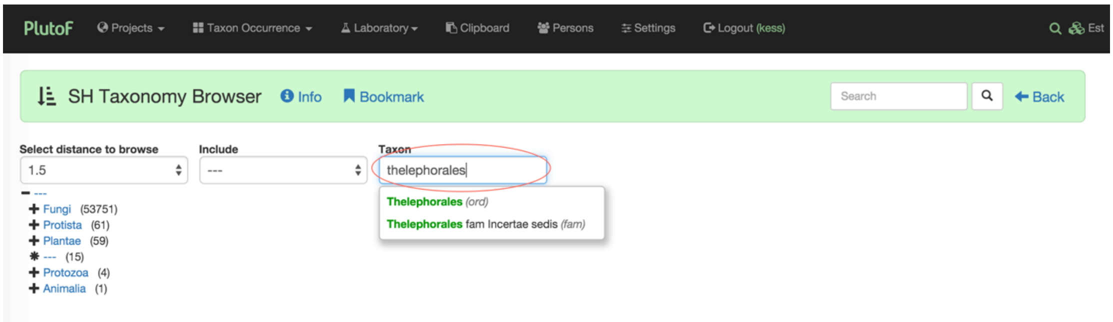
- Name
- Designators
- Sequence accession number
- Version
- Taxon node
- Threshold
- No. of sequences
- Singleton

Figure 2. Advanced SH search view.

4. SH Taxonomy Browser

SH-s can be browsed on hierarchical tree view by following *Laboratory => Molecular Lab => SH Taxonomy Browser* (Figure 1). Taxonomy Browser view is pre-computed at regular time intervals, and it is based on sequence identifications and their additions through third-party annotation.

SH-s can be browsed by clicking on "+" and "-" signs in front of taxon names to open tree branches. User can specify the distance to the closest SH (0.0 – 3.0%) and whether to include singleton SH-s or not when browsing the tree. There is also a taxon name autocomplete available to find and open specific node on the tree (Figure 3).



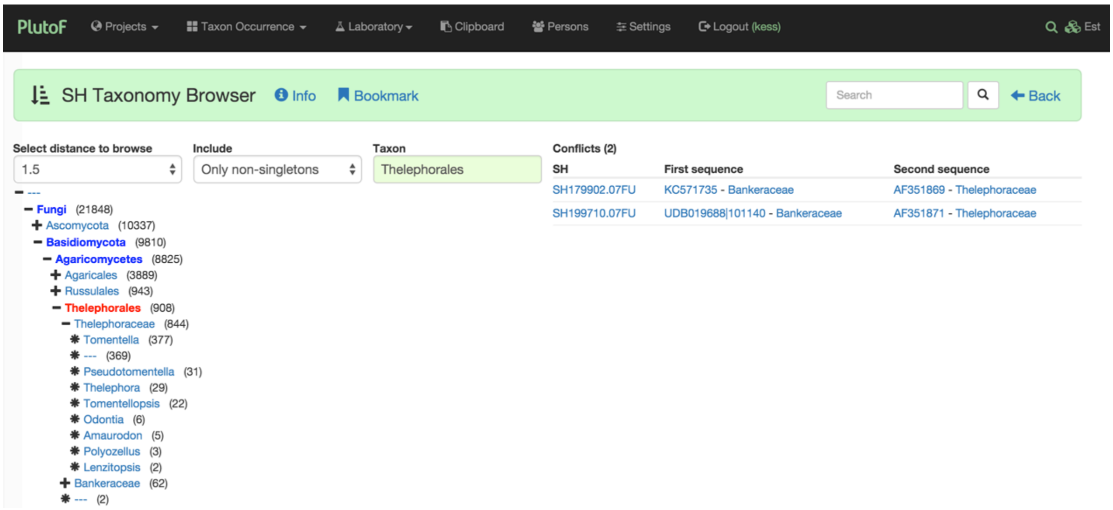
Select distance to browse: 1.5
Include: ---
Taxon: thelephorales

Thelephorales *(ord)*
Thelephorales fam Incertae sedis *(fam)*

```
− ---
  + Fungi   (53751)
  + Protista   (61)
  + Plantae   (59)
  ✱ ---   (15)
  + Protozoa   (4)
  + Animalia   (1)
```

Figure 3. Opening specific node on SH Taxonomy Browser.

When clicking on higher level (kingdom, phylum, class, order, and family) taxon name, a list of conflict SH-s is displayed on the right side of the page (Figure 4) – these are SH-s composed of sequences with conflict identifications (e.g. SH-s containing sequences identified both to *Bankeraceae* and *Thelephoraceae* are shown as conflict SH-s in *Thelephorales*, see SH179902.07FU as an example).



Select distance to browse: 1.5
Include: Only non-singletons
Taxon: Thelephorales

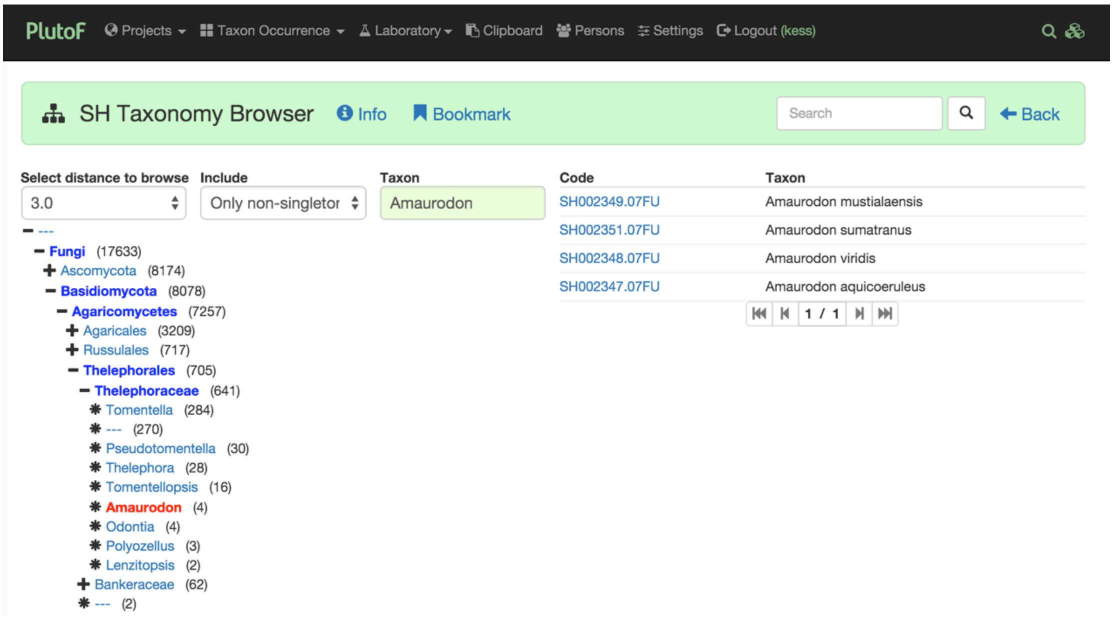
Conflicts (2)

SH	First sequence	Second sequence
SH179902.07FU	KC571735 - Bankeraceae	AF351869 - Thelephoraceae
SH199710.07FU	UDB019688|101140 - Bankeraceae	AF351871 - Thelephoraceae

```
− ---
  − Fungi   (21848)
    + Ascomycota   (10337)
    − Basidiomycota   (9810)
      − Agaricomycetes   (8825)
        + Agaricales   (3889)
        + Russulales   (943)
        − Thelephorales   (908)
          − Thelephoraceae   (844)
            ✱ Tomentella   (377)
            ✱ ---   (369)
            ✱ Pseudotomentella   (31)
            ✱ Thelephora   (29)
            ✱ Tomentellopsis   (22)
            ✱ Odontia   (6)
            ✱ Amaurodon   (5)
            ✱ Polyozellus   (3)
            ✱ Lenzitopsis   (2)
          + Bankeraceae   (62)
          ✱ ---   (2)
```

Figure 4. Browsing SH-s using hierarchical tree view – conflict SHs are shown.

When clicking on genus level SH-s, a list of SH-s belonging to selected genus is displayed on the right side of the page. Numbers next to the taxon name indicates the number of SH-s belonging to this taxon (Figure 5).



Select distance to browse: 3.0

Include: Only non-singletor

Taxon: Amaurodon

- ---
  - Fungi (17633)
    - Ascomycota (8174)
    - Basidiomycota (8078)
      - Agaricomycetes (7257)
        - Agaricales (3209)
        - Russulales (717)
        - Thelephorales (705)
          - Thelephoraceae (641)
            - Tomentella (284)
            - --- (270)
            - Pseudotomentella (30)
            - Thelephora (28)
            - Tomentellopsis (16)
            - Amaurodon (4)
            - Odontia (4)
            - Polyozellus (3)
            - Lenzitopsis (2)
          - Bankeraceae (62)
          - --- (2)

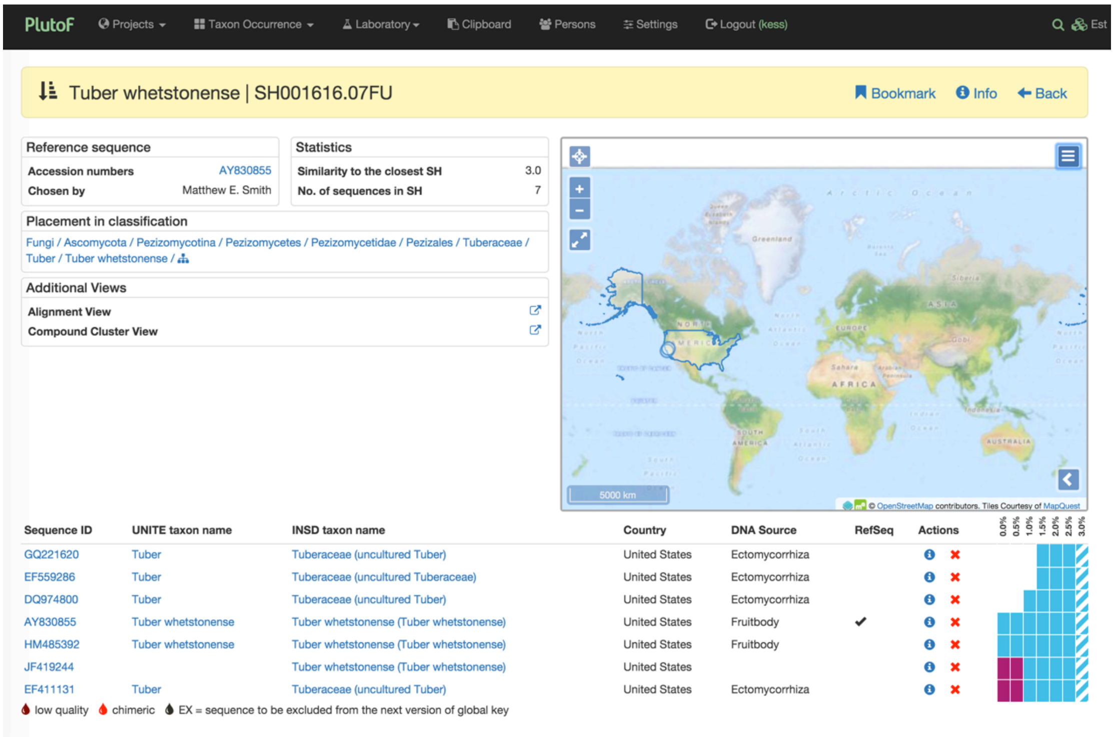
Code	Taxon
SH002349.07FU	Amaurodon mustialaensis
SH002351.07FU	Amaurodon sumatranus
SH002348.07FU	Amaurodon viridis
SH002347.07FU	Amaurodon aquicoeruleus

1 / 1

Figure 5. Browsing SH-s using hierarchical tree view – SHs belonging to genus are shown.

5. SH detail-view

SH detail-view includes list of sequences with metadata, links to alignment and compound cluster views, distribution map for sequences, placement in classification, info on reference or representative sequence, and further statistics on SH (Figure 6).



Tuber whetstonense | SH001616.07FU

Reference sequence

Accession numbers	AY830855
Chosen by	Matthew E. Smith

Statistics

Similarity to the closest SH	3.0
No. of sequences in SH	7

Placement in classification

Fungi / Ascomycota / Pezizomycotina / Pezizomycetes / Pezizomycetidae / Pezizales / Tuberaceae / Tuber / Tuber whetstonense /

Additional Views

Alignment View

Compound Cluster View

Sequence ID	UNITE taxon name	INSD taxon name	Country	DNA Source	RefSeq	Actions	0.0%	0.5%	1.0%	1.5%	2.0%	2.5%	3.0%
GQ221620	Tuber	Tuberaceae (uncultured Tuber)	United States	Ectomycorrhiza		ℹ ✖				🟦	🟦	🟦	▨
EF559286	Tuber	Tuberaceae (uncultured Tuberaceae)	United States	Ectomycorrhiza		ℹ ✖				🟦	🟦	🟦	▨
DQ974800	Tuber	Tuberaceae (uncultured Tuber)	United States	Ectomycorrhiza		ℹ ✖			🟦	🟦	🟦	🟦	▨
AY830855	Tuber whetstonense	Tuber whetstonense (Tuber whetstonense)	United States	Fruitbody	✓	ℹ ✖	🟦	🟦	🟦	🟦	🟦	🟦	▨
HM485392	Tuber whetstonense	Tuber whetstonense (Tuber whetstonense)	United States	Fruitbody		ℹ ✖	🟦	🟦	🟦	🟦	🟦	🟦	▨
JF419244		Tuber whetstonense (Tuber whetstonense)	United States			ℹ ✖	🟪	🟪	🟦	🟦	🟦	🟦	▨
EF411131	Tuber	Tuberaceae (uncultured Tuber)	United States	Ectomycorrhiza		ℹ ✖	🟪	🟪	🟦	🟦	🟦	🟦	▨

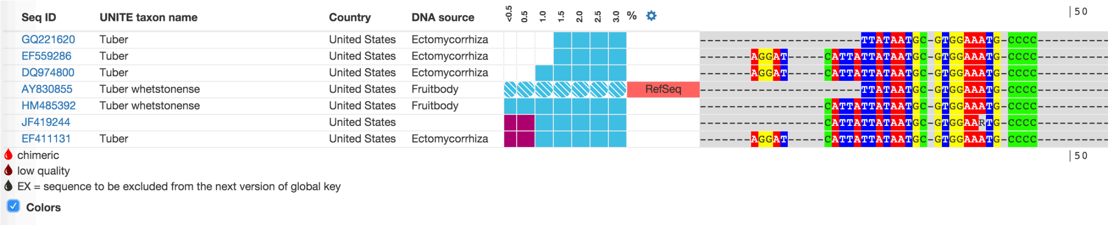
low quality chimeric EX = sequence to be excluded from the next version of global key

Figure 6. Detail-view for SH001616.07FU.

SH sequence alignment view includes indication for RefS (is one has been chosen), available actions with sequences (icons for info pop-up and for excluding sequence from next version). Colored bars indicate the belonging of sequences in distinct SH-s on different distance thresholds, and are clickable for switching between thresholds. White bars refer to formation of singleton SH-s at the corresponding threshold.

For example, on Figure 6, all sequences in SH001616.07FU cluster together on thresholds between 1.5 – 3.0%, but on 1.0% threshold GQ221620 and EF559286 fall out of the cluster to form two singleton SH-s.

When clicking on external link icon ( ) for alignment view, sequence alignment with customizable labels appears (Figure 7) allowing to examine multiple sequence alignment (generated by MAFFT using custom parameters based on matrix size) for sequences forming this SH.



Seq ID	UNITE taxon name	Country	DNA source	<0.5	0.5	1.0	1.5	2.0	2.5	3.0	%	|50
GQ221620	Tuber	United States	Ectomycorrhiza				🟦	🟦	🟦	🟦		---------------------TTATAATGC-GTGGAAATG-CCCC----------
EF559286	Tuber	United States	Ectomycorrhiza				🟦	🟦	🟦	🟦		-------AGGAT-----CATTATTATAATGC-GTGGAAATG-CCCC----------
DQ974800	Tuber	United States	Ectomycorrhiza			🟦	🟦	🟦	🟦	🟦		-------AGGAT-----CATTATTATAATGC-GTGGAAATG-CCCC----------
AY830855	Tuber whetstonense	United States	Fruitbody	▨	▨	▨	▨	▨	▨	▨	RefSeq	---------------------TTATAATGC-GTGGAAATG-CCCC----------
HM485392	Tuber whetstonense	United States	Fruitbody	🟦	🟦	🟦	🟦	🟦	🟦	🟦		----------------CATTATTATAATGC-GTGGAAATG-CCCC----------
JF419244		United States		🟪	🟪	🟦	🟦	🟦	🟦	🟦		----------------CATTATTATAATGC-GTGGAARTG-CCCC----------
EF411131	Tuber	United States	Ectomycorrhiza	🟪	🟪	🟦	🟦	🟦	🟦	🟦		-------AGGAT-----CATTATTATAATGC-GTGGAAATG-CCCC----------

|50

chimeric

low quality

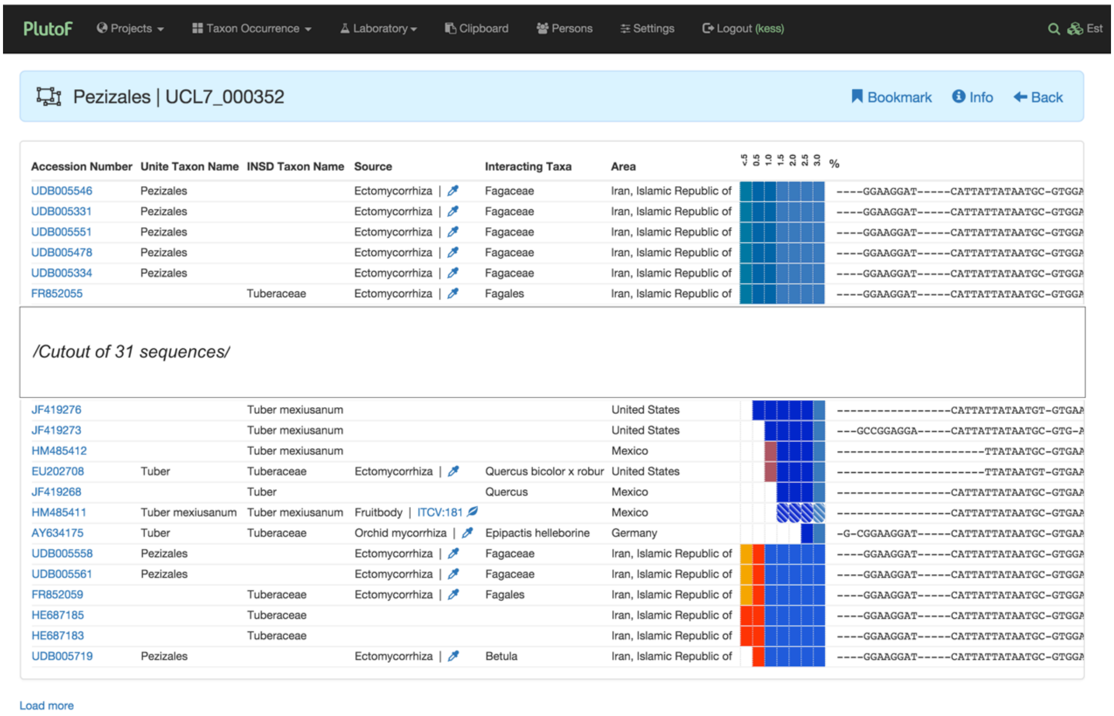
EX = sequence to be excluded from the next version of global key

☑ Colors

Figure 7. Multiple sequence alignment view for SH.

6. Compound cluster view

Clicking on the external link icon for compound cluster in SH detail-view will open compound cluster for this SH (Figure 8). Compound cluster includes sequences clustered approximately on genus/subgenus level that are subjected to dynamic clustering on different thresholds to form SH-s.



Pezizales | UCL7_000352

Accession Number	Unite Taxon Name	INSD Taxon Name	Source	Interacting Taxa	Area	<.5	0.5	1.0	1.5	2.0	2.5	3.0	%
UDB005546	Pezizales		Ectomycorrhiza |	Fagaceae	Iran, Islamic Republic of	🟦	🟦	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA
UDB005331	Pezizales		Ectomycorrhiza |	Fagaceae	Iran, Islamic Republic of	🟦	🟦	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA
UDB005551	Pezizales		Ectomycorrhiza |	Fagaceae	Iran, Islamic Republic of	🟦	🟦	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA
UDB005478	Pezizales		Ectomycorrhiza |	Fagaceae	Iran, Islamic Republic of	🟦	🟦	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA
UDB005334	Pezizales		Ectomycorrhiza |	Fagaceae	Iran, Islamic Republic of	🟦	🟦	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA
FR852055		Tuberaceae	Ectomycorrhiza |	Fagales	Iran, Islamic Republic of	🟦	🟦	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA

/Cutout of 31 sequences/

Accession Number	Unite Taxon Name	INSD Taxon Name	Source	Interacting Taxa	Area	<.5	0.5	1.0	1.5	2.0	2.5	3.0	%
JF419276		Tuber mexiusanum			United States		🟦	🟦	🟦	🟦	🟦	🟦	-----------------CATTATTATAATGT-GTGAA
JF419273		Tuber mexiusanum			United States			🟦	🟦	🟦	🟦	🟦	---GCCGGAGGA-----CATTATTATAATGC-GTG-A
HM485412		Tuber mexiusanum			Mexico			🟥	🟦	🟦	🟦	🟦	----------------------TTATAATGC-GTGAA
EU202708	Tuber	Tuberaceae	Ectomycorrhiza |	Quercus bicolor x robur	United States			🟥	🟦	🟦	🟦	🟦	----------------------TTATAATGT-GTGAA
JF419268		Tuber		Quercus	Mexico				🟦	🟦	🟦	🟦	-----------------CATTATTATAATGC-GTGAA
HM485411	Tuber mexiusanum	Tuber mexiusanum	Fruitbody | ITCV:181		Mexico				▨	▨	▨	▨	-----------------CATTATTATAATGC-GTGAA
AY634175	Tuber	Tuberaceae	Orchid mycorrhiza |	Epipactis helleborine	Germany						🟦	🟦	-G-CGGAAGGAT-----CATTATTATAATGC-GTGAA
UDB005558	Pezizales		Ectomycorrhiza |	Fagaceae	Iran, Islamic Republic of	🟧	🟥	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA
UDB005561	Pezizales		Ectomycorrhiza |	Fagaceae	Iran, Islamic Republic of	🟧	🟥	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA
FR852059		Tuberaceae	Ectomycorrhiza |	Fagales	Iran, Islamic Republic of	🟧	🟥	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA
HE687185		Tuberaceae			Iran, Islamic Republic of	🟥	🟥	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA
HE687183		Tuberaceae			Iran, Islamic Republic of	🟥	🟥	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA
UDB005719	Pezizales		Ectomycorrhiza |	Betula	Iran, Islamic Republic of		🟥	🟦	🟦	🟦	🟦	🟦	----GGAAGGAT-----CATTATTATAATGC-GTGGA

Load more

Figure 8. Compound cluster view for SH001616.07FU (UCL7_000352).

If compound cluster includes more than 50 sequences, “Load more” link will be displayed on the bottom of the page to extend the view.

In compound cluster view, reference sequences on different thresholds are indicated by dashed color bars. Clicking on 3.0% color bar will open SH detail-view where reference sequences can be set.

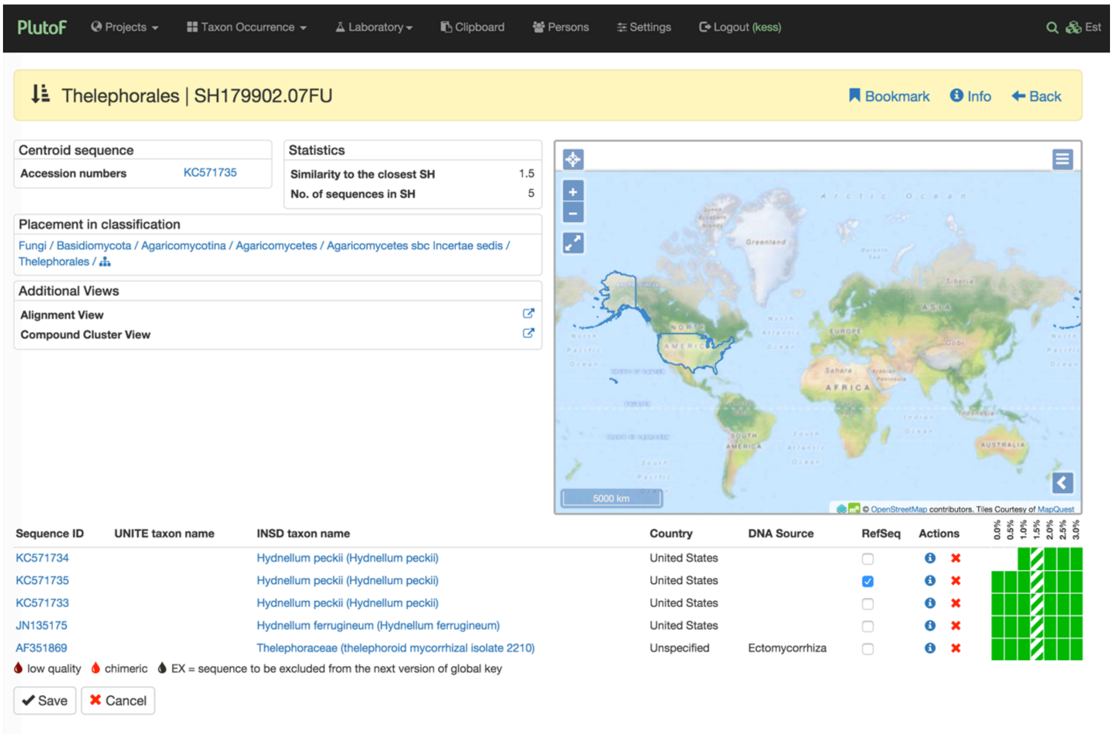
Similar to the functionality of SH-s, compound clusters can be browsed as hierarchical tree, and searched by their code, taxonomy, and sequence accession numbers by following *Laboratory* => *Molecular Lab* => *Compound Clusters*.

7. Choosing reference sequences

Reference sequences (RefS) can be selected by users belonging to the workgroup “UNITE Species Hypotheses” (see introduction section in this document on how to join the workgroup).

If reference sequence is already set for SH, it is indicated with checked icon (✔) in RefSeq column for the selected sequence (see Figure 6).

If reference sequence has not been set, user with access rights can select one by checking one of the boxes indicating which sequence should be set as RefS (see Figure 9). To confirm and complete the action of setting RefS, “Save” button has to be clicked on the bottom of the page.



Thelephorales | SH179902.07FU

Centroid sequence

Accession numbers KC571735

Statistics

Similarity to the closest SH	1.5
No. of sequences in SH	5

Placement in classification

Fungi / Basidiomycota / Agaricomycotina / Agaricomycetes / Agaricomycetes sbc Incertae sedis / Thelephorales /

Additional Views

Alignment View

Compound Cluster View

Sequence ID	UNITE taxon name	INSD taxon name	Country	DNA Source	RefSeq	Actions	0.0%	0.5%	1.0%	1.5%	2.0%	2.5%	3.0%
KC571734		Hydnellum peckii (Hydnellum peckii)	United States		□	ℹ ✖			🟩	▨	🟩	🟩	🟩
KC571735		Hydnellum peckii (Hydnellum peckii)	United States		☑	ℹ ✖	🟩	🟩	🟩	▨	🟩	🟩	🟩
KC571733		Hydnellum peckii (Hydnellum peckii)	United States		□	ℹ ✖	🟩	🟩	🟩	▨	🟩	🟩	🟩
JN135175		Hydnellum ferrugineum (Hydnellum ferrugineum)	United States		□	ℹ ✖	🟩	🟩	🟩	▨	🟩	🟩	🟩
AF351869		Thelephoraceae (thelephoroid mycorrhizal isolate 2210)	Unspecified	Ectomycorrhiza	□	ℹ ✖	🟩	🟩	🟩	▨	🟩	🟩	🟩

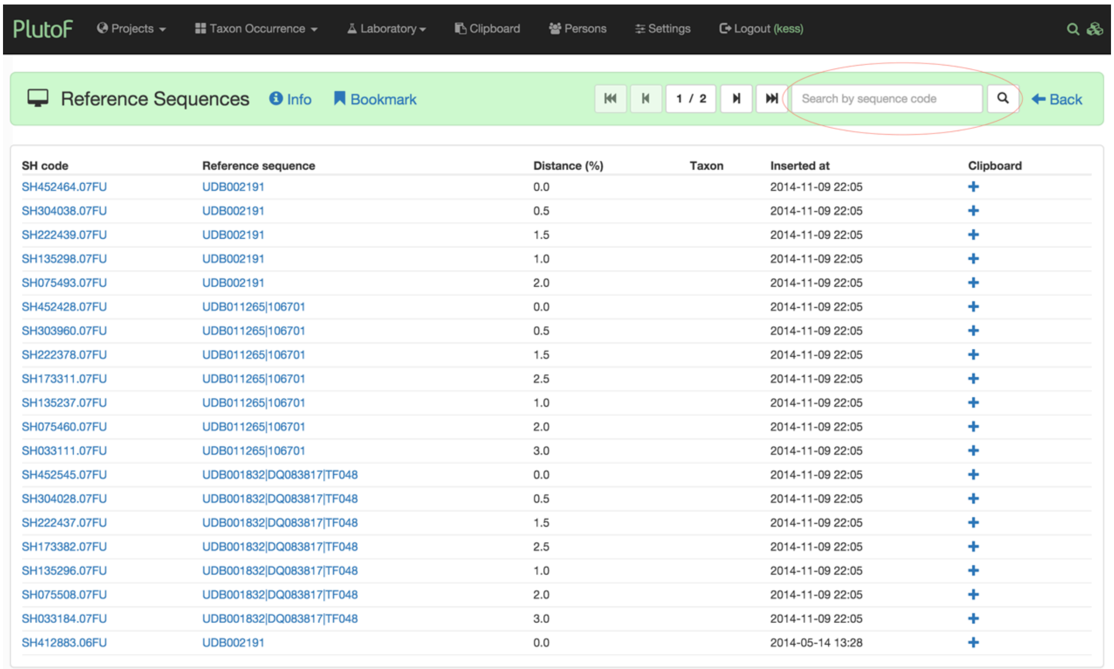
low quality chimeric EX = sequence to be excluded from the next version of global key

✔ Save ✖ Cancel

Figure 9. Choosing reference sequence for SH.

Once chosen, RefS will span over all thresholds where reference sequence has not yet been set. Only the user who originally set reference sequence (in addition to system admin) can change or remove it by unchecking the same checkbox and clicking “Save”. The latter action, unsetting RefS, will also span over all thresholds.

Reference sequences chosen by the user can be found by following *Laboratory => Molecular Lab => Reference Sequences*, where list of all RefS set by user is displayed. Sequences can be further searched by entering the Reference Sequences search module (Figure 10).



SH code	Reference sequence	Distance (%)	Taxon	Inserted at	Clipboard
SH452464.07FU	UDB002191	0.0		2014-11-09 22:05	+
SH304038.07FU	UDB002191	0.5		2014-11-09 22:05	+
SH222439.07FU	UDB002191	1.5		2014-11-09 22:05	+
SH135298.07FU	UDB002191	1.0		2014-11-09 22:05	+
SH075493.07FU	UDB002191	2.0		2014-11-09 22:05	+
SH452428.07FU	UDB011265|106701	0.0		2014-11-09 22:05	+
SH303960.07FU	UDB011265|106701	0.5		2014-11-09 22:05	+
SH222378.07FU	UDB011265|106701	1.5		2014-11-09 22:05	+
SH173311.07FU	UDB011265|106701	2.5		2014-11-09 22:05	+
SH135237.07FU	UDB011265|106701	1.0		2014-11-09 22:05	+
SH075460.07FU	UDB011265|106701	2.0		2014-11-09 22:05	+
SH033111.07FU	UDB011265|106701	3.0		2014-11-09 22:05	+
SH452545.07FU	UDB001832|DQ083817|TF048	0.0		2014-11-09 22:05	+
SH304028.07FU	UDB001832|DQ083817|TF048	0.5		2014-11-09 22:05	+
SH222437.07FU	UDB001832|DQ083817|TF048	1.5		2014-11-09 22:05	+
SH173382.07FU	UDB001832|DQ083817|TF048	2.5		2014-11-09 22:05	+
SH135296.07FU	UDB001832|DQ083817|TF048	1.0		2014-11-09 22:05	+
SH075508.07FU	UDB001832|DQ083817|TF048	2.0		2014-11-09 22:05	+
SH033184.07FU	UDB001832|DQ083817|TF048	3.0		2014-11-09 22:05	+
SH412883.06FU	UDB002191	0.0		2014-05-14 13:28	+

Figure 10. Reference sequences list-view. Red circle on navigation bar indicates where to enter the search module.

8. Guidelines for the choosing reference sequence

Basic guidelines

I. Sequence from type material has priority
Sequence of the type material has no priority if it is too short or of low quality.

II. One reference sequence per species hypothesis
Example 1: Species hypothesis (SH) based on 97% similarity threshold value includes one reference sequence X. If this SH is divided into two species by 98% similarity threshold value then one SH will include reference sequence X, but second SH should receive new reference sequence Y.

Example 2: If two SH-s that have reference sequences X and Y lumped together, then one of them will become reference sequence of the new SH. Currently, PlutoF automatically selects reference sequence which was chosen first. This decision can be amended by expert.

III. Reference sequence can be replaced
Reference sequence X can be replaced by a new sequence Y if its source stands higher in "Reference sequence selection priority list" (see below).

Example: Reference sequence X is derived from soil sample but later sequence Y from living culture becomes available. It falls inside the same SH as reference sequence X and therefore may replace it.

Remark: Current version of the PlutoF requires expert to make the replacement. The alarming system that there is potentially a better reference sequence available, will be implemented in future versions.

Practical recommendations for the selection of reference sequence

Reference sequence selection priority list
The selection priority in decreasing order is as follows (by assuming that sequences are of high quality): type material, specimen in public collection, living culture in public collection, and sequence from any other biological sample.

1. If type specimen is sequenced then it is also reference sequence of this species – it carries the species name.

If the sequence of type specimen is not in the species cluster or if it is of low quality then we recommend the following selection procedures:

2. The sequence from authentic herbarium specimen or living culture, which is identified by expert, should be chosen. The species name of the specimen is also the name of the reference sequence. The locality of the reference sequence should be as close as possible to the type material locality.

3. If species cluster includes only sequences from biological samples like soil, water, air, tissue of other organism, etc. then sequence available in INSD should be chosen. If there are no sequences from INSD then sequence submitted into other public databases like UNITE should be chosen. The name of the reference sequence is accession code accompanied by genus name, if available.

4. Cloned sequences are not recommended as reference sequences, except cases when well grounded SH includes cloned sequences only.