NB! This is a separate paragraph of the PlutoF manual. Please use current manual at http://unite.ut.ee/temp/plutof2/files/PlutoF_2.5_Manual_small.pdf if needed or cited in this document.

6. Global key annotations
6.1. Finding the clusters of sequences
   6.1.1. Using menu Search and edit => All sequences
   6.1.2. Using menu Global Key annotations => GK annotations
6.2. Working with clusters
6.3. Working with species
6.4 Guidelines for the choosing reference sequence
6. Global key annotations

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 hypothesis in all clusters based on identical criteria. RepS allows to name and communicate species until RefS becomes available for given species.

**Name of the reference sequence.** Reference sequence maybe identified on species, genus, family or higher level. The name of the reference sequence is a combination of taxon name and unique INSDC or UNITE accession code.

Example 1: The INSDC sequence EU668254 originate from the plant mycorrhizal root and identified as a Pseudotomentella sp. in UNITE database. If it is selected as a reference sequence then its name is “Pseudotomentella sp. EU668254”.

Example 2: The INSDC sequence EU668254 is originate from the sporocarp and identified as a Pseudotomentella mucidula in UNITE database. If it is selected as a reference sequence then its name is “Pseudotomentella mucidula EU668254”.

**Name of the representative sequence** is formed the same way as for reference sequences (see examples above).

Currently PlutoF cloud supports only fungal ITS based annotations and key which is based on UNITE database (unite.ut.ee). The UNITE database includes core data set of ITS sequences which originate from the fruitbodies identified by experts as well as all INSDC fungal ITS sequences with sufficient quality. Technical description of the selection of fungal ITS sequences and subsequent clustering is available in the end of this manual.
6.1. Finding the clusters

There are two basic ways how to find clusters of the specific taxon.

6.1.1. Using menu Search and edit => all sequences

Please consult paragraph 3.7 for the searching sequences of particular taxa. In Figure 6-1 is shown search results for the species *Tomentella sublilacina*. The direct link to the Global Key cluster is shown in the end of each sequence. We recommend to use Qview option which is much faster because the alignment of the cluster is displayed in black and grey instead of colours. UCL4 and UCL5 are acronyms for the Fungal Global Key versions four and five respectively. Clicking on cluster name (e.g. UCL4_000977) will display this cluster with alignment in full colours.

Figure 6-1. Search results for the *Tomentella sublilacina* sequences.
6.1.2. Using menu Global Key annotations => GK annotations

In the Figure 6-2 is shown Global Key annotations => GK annotations window. Here you can Select global key version to browse; List of global key singletons (sequences which didn’t fall into any cluster); List of global key clusters; Search clusters and singletons by UNITE or INSD taxon names, INSD accession number, cluster code or by ectomycorrhizal (EcM) lineage.
In Figure 6-3 is shown result of the Search clusters and singletons by UNITE name “Tomentella sublilacina”. To display cluster you can preferably click on link “Qview” or on cluster code “UCL5_005194”. Figure 6-4 shows the cluster UCL5_005194 window which is opened when clicking on “Qview”.

Figure 6-4. View of the cluster UCL5_005194.
6.2 Working with clusters

On Fig 6.4 is shown cluster UCL5_005194. The header of the window displays information on version and cluster ID. Next lines list genera (UNITE names only), which appear in this cluster and number of sequences in cluster. The sequence is likely chimeric if its ID and other text in this line is shown in red and it will be removed from the next version. If the text is brown then the sequence is low quality and will be removed in next version as well. UNITE core sequence ID-s are shown in yellow. Ex in the front of Sequence ID shows that expert decided that it will be removed from the next version of the key. Each cluster has following columns: 1) Sequence ID displays UNITE or/and INSD accession code which is hyperlink to the original as well as annotated data which can be edited by expert (see paragraph 3.8 in PlutoF manual). Clicking on “more” will open a small window below the line which displays all alternative identifications if present and allows to add data on specimen as well as mark sequence if it should be removed from next version; 2) UNITE taxon name is name given to the specimen or to the sequence from any other biological sample; 3) INSD taxon name displays name of the sequence in the INSD original data; 4) Country shows the name of the country from where the sequence originates; 5) DNA source shows the type of biological sample from where the sequence originates; 6) Next column allows to choose threshold value for the species discrimination (see also 6.3); 7) Clustering based on allows to switch between full ITS and ITS2 based alignments and species; 8) Order sequences allows to reorder sequences based on mafft alignment or blastclust outputs (default is combined approach); 9) Download alignment as a FASTA file.
6.3 Working with species

Column “DSH” on Figure 6-4 is divided into five strips based on ITS sequence similarity threshold values 99, 98.5, 98, 97.5 and 97% (from left to right). Strip cells of the sequences, which cluster together based on specific threshold value have the same colour. If sequence is not clustering with any sequence then the strip cell is colourless. For example 3rd and 4th sequences on Figure 6-4 are not clustering with 99, 98.5 and 98% threshold values, but do with 97.5 and 97%. Figure 6-5 shows the middle part of the same cluster (UCL5_005194). The cursor is on the left edge of the left strip (99% threshold value based clustering). Clicking on it will display new window shown on Figure 6-6.

Figure 6-5. Middle part of the cluster UCL5_005194.

Figure 6-6. Cluster of nine sequences based on 99% similarity threshold value.
In this window (Figure 6-6) expert can set reference sequence of the species by clicking on button “set ref”. See paragraph “Reference sequence” for the guidelines how to choose it. Clicking on “set ref” of the *Tomentella tenuis* sequence AM412299 will reload this page as shown on Figure 6-7. Reference sequence can be unset by clicking on button “unset ref”. In this widow automatically chosen representative sequence is shown in green colour.

![Image](image1.png)

Figure 6-7. *Tomentella tenuis* sequence AM412299 is set as a reference sequence.

Reloaded window shown on Figure 6-5 will display reference sequence for all strips (Figure 6-8)

![Image](image2.png)

Figure 6-8. Middle part of the cluster UCL5_005194 with reference sequence of *Tomentella tenuis* marked for all five strips.
6.4 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 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 which have reference sequences X and Y are lumped together then one of them will become reference sequence of the new SH. Currently PlutoF will automatically select 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 needs that expert will make the replacement. The alarming system that potentially better reference sequence is available will be implemented in future version.

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 low quality then we recommend 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 a reference sequences except cases when well grounded SH includes cloned sequences only.