GLOSSary USER GUIDE

GLOSSary (GLobal Ocean 16S Subunit web accessible resource) is a metagenomics platform that allows user friendly explorations of 16S rRNA related sequences from environmental metagenomics efforts.

GLOSSary was conceived and is maintained by the BIOIN*for*MA service at the Stazione Zoologica Anton Dohrn.

Description

GLOSSary can be accessed by two main entry points: a "Keyword search" and a "BLAST search". Both pages can be reached from the GLOSSary main page (Figure 1).

Figure 1. GLOSSary main page. The dark blue buttons redirect to the "Keyword search" or to the "BLAST search" web pages.

GLOSSary currently contains the 16S miTAGs collections from the Tara project [1] and processed sequences from the BIOIN*for*MA service. The processing pipeline is described in Tangherlini *et al.*.

"Keyword search" Web Page

The GLOSSary data can be accessed at the "Keyword search" page by an interactive map and by taxonomy affiliation, by sequence ID, by Tara station ID (Figure 2).

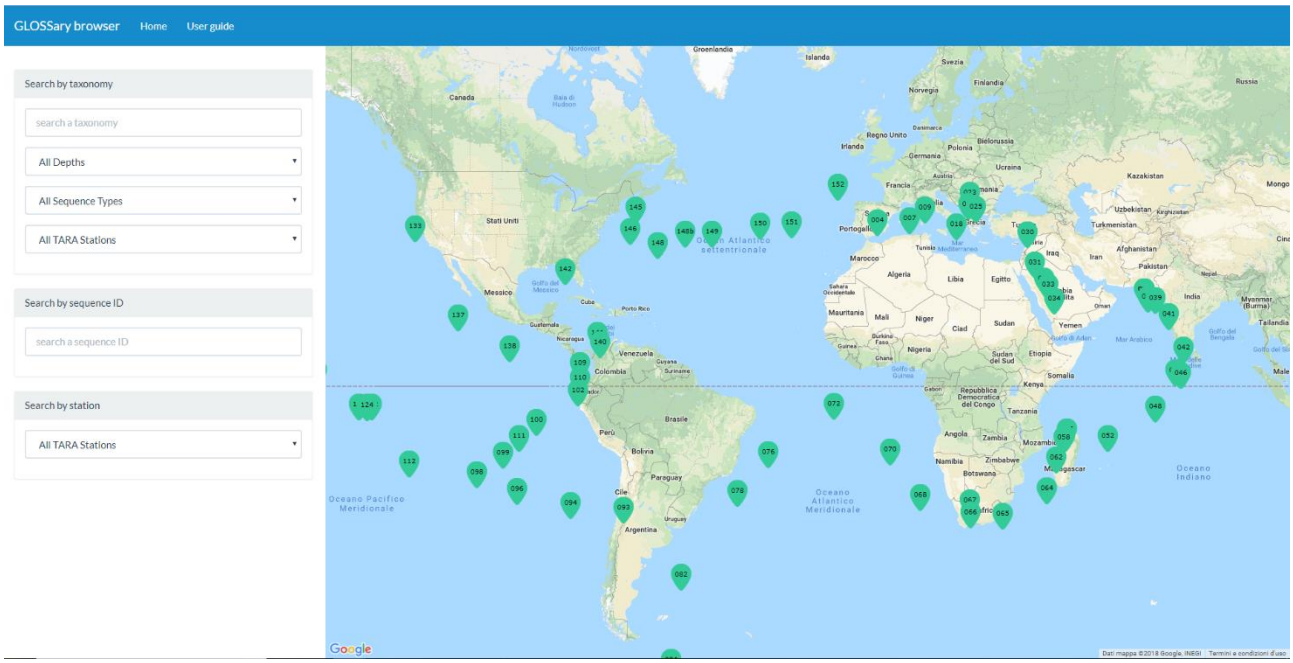

Figure 2. GLOSSary "Keyword search" web page. On the left, there are the "Search by taxonomy", "Search by sequence ID" and the "Search by station" options; on the right, a map with green pointers, indicating the geographical position of the Tara stations, is also displayed.

1) "*Search by taxonomy*"

The "*Search by taxonomy*" option allows to search by the available taxonomic affiliations. While typing in the *taxonomy* search field, a dropdown list will appear, allowing the user to narrow the search to the related terms (Figure 3). The list includes all the taxa that are associated to at least one of the sequences included in the database. Please be aware: if the selected taxonomic affiliation is highly inclusive (e.g. a domain) the query could take a while to be performed. Please be patient!

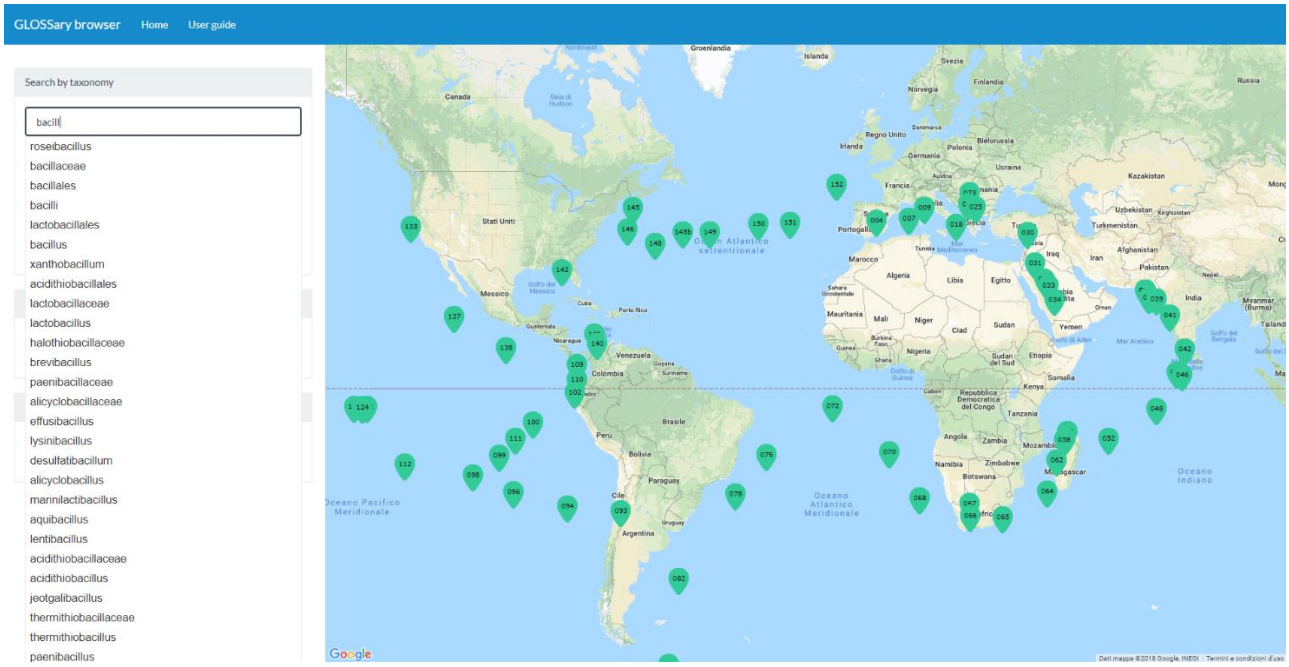

Figure 3. Typing a name into the taxonomy field of the "Search by taxonomy" area. All the matching taxonomy affiliation (i.e. containing the inserted chars, for example "*bacill*") will be displayed.

The page will show the corresponding results: the map will be updated with pointers indicating the Tara stations including hits matching the query. The central section of the page will list all the sequences having the selected taxonomic affiliation. The total number of hits found is shown on the top, and the results, grouped by Tag, are listed below (Figure 4).

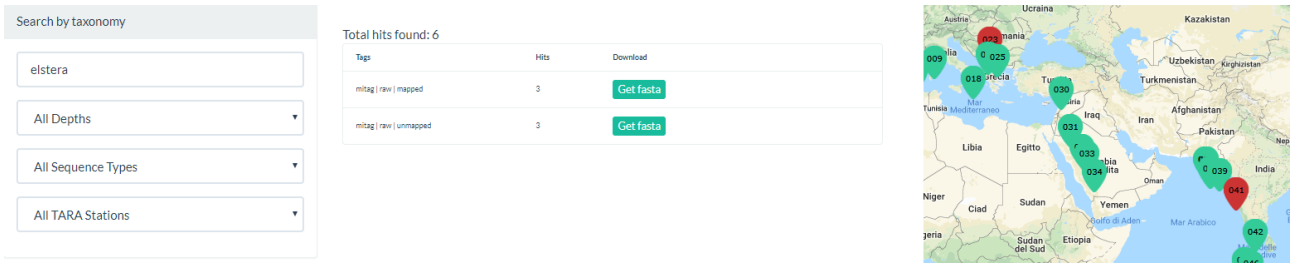

Figure 4. Results of a search by taxonomy. The hits found are grouped by tag. Pointers on the map are updated accordingly (red pointers).

The Tags are labels used to identify features of the extracted sequences according to the processing pipeline. They include, in the case of the contigs, the classification, the length and the contamination status. The miTAGs, instead, are classified in "mapped"

or “unmapped”, indicating if they mapped or not on the contigs, using the VSEARCH aligner [2] (minimum coverage accepted 97%).
Detail of the possible values are described in Table 1.

sequence type	classification	Length	contamination status
Contig	long	> 900bp	chimeraFree chimera borderline
	medium	> 800bp and < 900bp	
	short	< 800bp	
miTAG	mapped		
	Unmapped		

Table 1. Possible values of the Tags fields.

It is possible to further refine the query by selecting specific values for the following fields (Figure 4):

- “**All Depths**”. The dropdown lists the available sampling depths: “*Surface Oceanic*”, “*Deep-Chlorophyll Maximum*”, “*Epipelagic*”, “*Mesopelagic*” (metadata described by the Tara consortium).
- “**All Sequence Types**”. Either the Tara “*miTAGs*” or the “*16S Contigs*” from the GLOSSARY collection.
    - If “*16S Contigs*” is selected, the “*All Contigs Length*” dropdown list will appear. From here it is possible to select the contigs by classification (i.e. length): “*All Contigs Length*”, “*Long (> 900bp)*”, “*Medium (> 800bp and < 900bp)*”, “*Short (< 800 bp)*”;
    - If “*miTAGs*” is selected, the new “*All miTAGs*” dropdown list will appear. From here it is possible to select among the “*Mapped*” or “*Unmapped*” miTAGs;

- "**All TARA Stations**". Tara station IDs are accessible in the dropdown list. The links to primary resources per libraries are provided (both miTAGs and raw reads).

2) **"*Search by Sequence ID*"**

In the "*Search by sequence ID*" box it is possible to search for a specific sequence among miTAG or contigs by its identifier (e.g. TARA-004-DCM_2) (Figure 5).

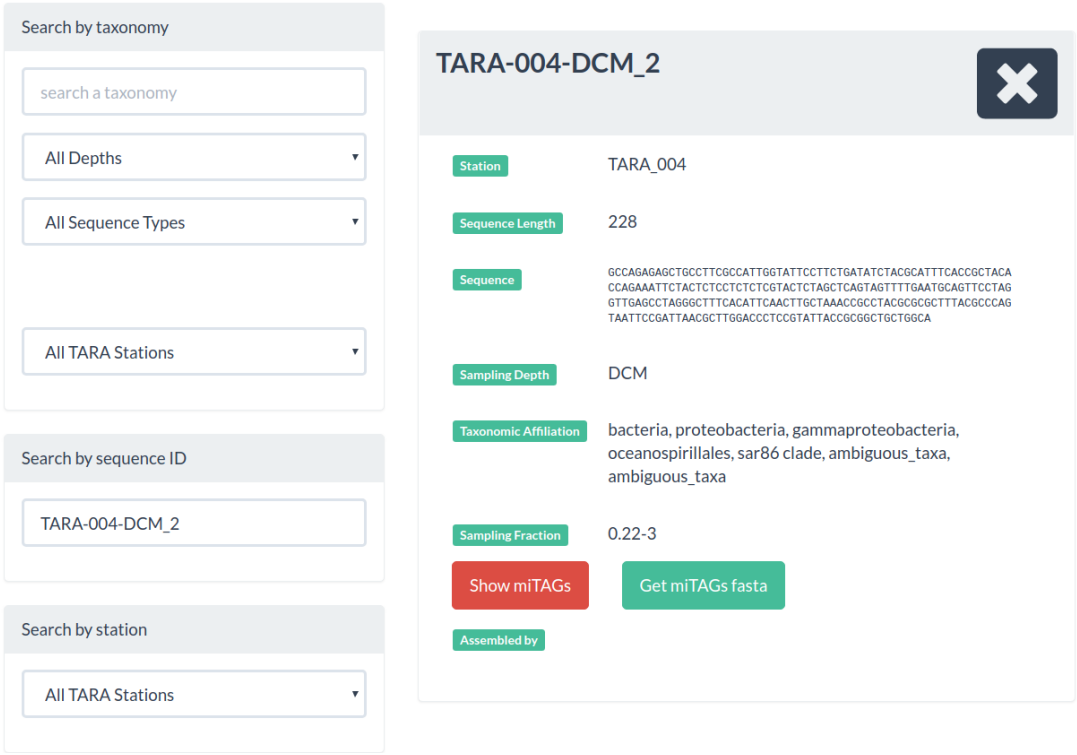

Figure 5. Searching by contig ID. Details on the contig and its features are provided, including the possibility of viewing and downloading the associated miTAGs.

Details on the sequence including taxonomic affiliation, the sequence and its length are shown (see Figure 5 and Figure 6). If the query is a contig, it is possible to view the list of miTAGs sequences that map onto the contig. The sequences can also be downloaded as FASTA formatted files (Figure 5).

If an assembled miTAG sequence is searched, the crosslink to the corresponding contig ID is provided (Figure 6).

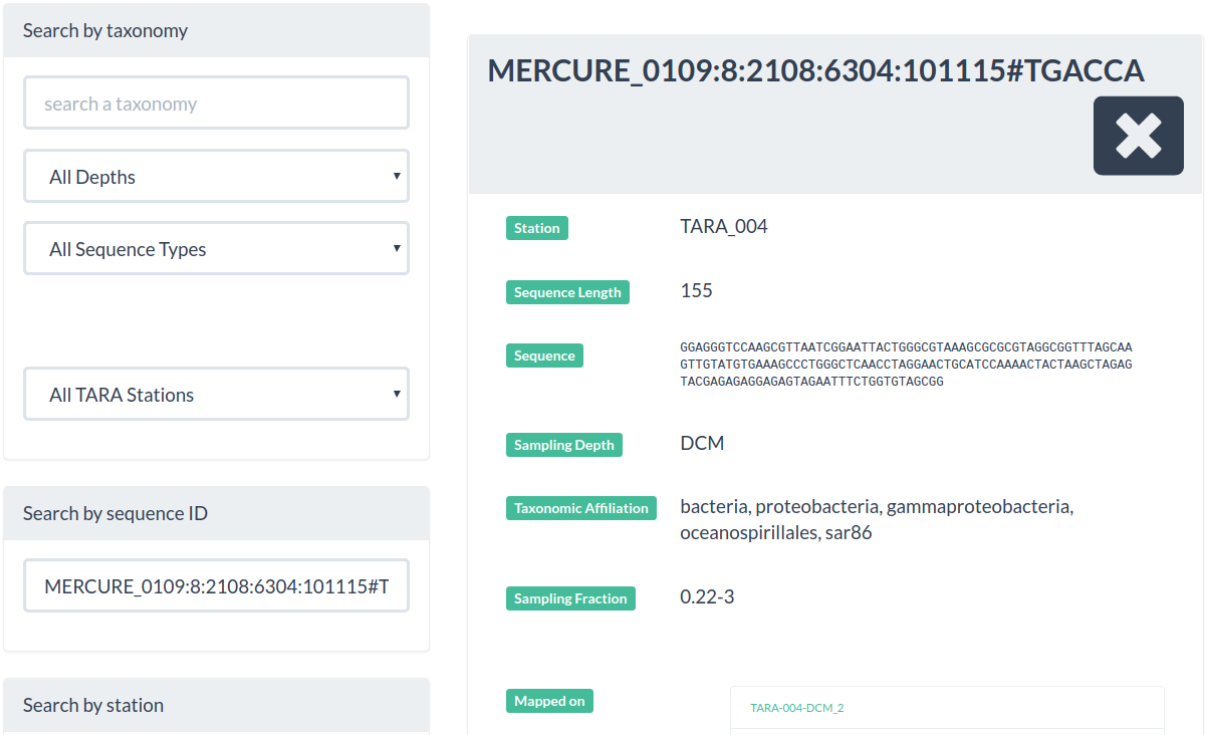

Figure 6. Searching by miTAG ID. Details on the miTAG and its features are given, including the ID of the contig to which it belongs to.

3) "***Search by station***"

In the "*Search by station*" box it is possible to search for a specific Tara station ID selecting from the dropdown list (Figure 7). The results shown include:

- the list of the runs, linked to the libraries at the SRA archives of the NCBI for that station;
- the list of miTAG libraries for the station. Clicking on the link starts the download of the corresponding miTAGs libraries in FASTA format.

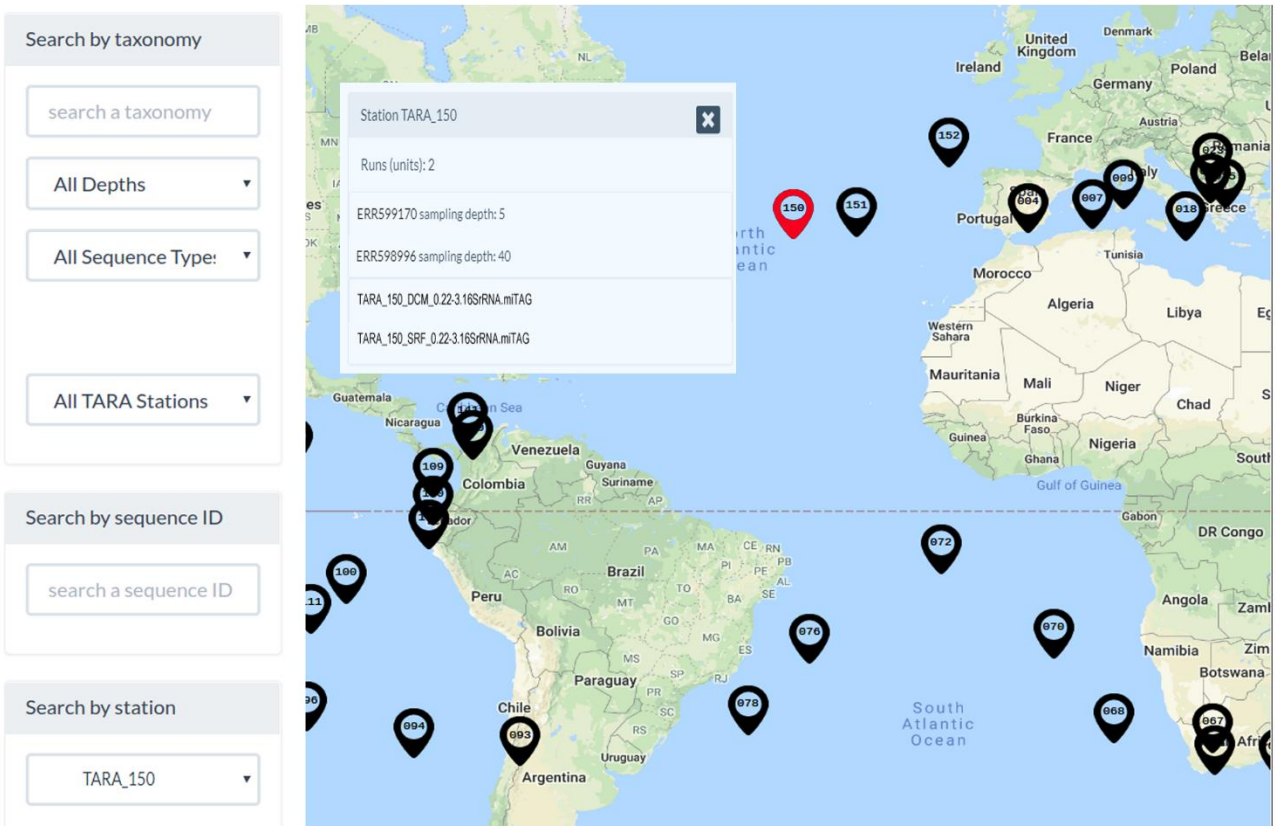

Figure 7. Searching by station ID. The list of the runs and the list of the miTAG libraries for the selected station are shown. Items in the first list link to the SRA archives of the NCBI, the second ones directly download the libraries in FASTA format.

“BLAST search” Web Page

The GLOSSary data can be searched by similarity using the “BLAST search” service (Figure 8).

Figure 8. GLOSSary “BLAST search” web page.

The user can use one or more sequences (in FASTA format) as a query, choosing one or more databases among the available ones:

- 16S Contigs:
  - “long” (>900 bp) chimera free;
  - “long” (>900 bp) chimera;
  - “long” (>900 bp) borderline;
  - “medium” (800 bp<>900 bp) chimera free;
  - “medium” (800 bp<>900 bp) chimera;
  - “medium” (800 bp<>900 bp) borderline;
  - “short” (<800 bp).
- miTAGs:
  - mapped on contigs (i.e. assembled);
  - unmapped on contigs (i.e. not assembled).
- SILVA DB:
  - SSU Ref NR 99 (version 128).

References

1. Sunagawa, S., et al., *Structure and function of the global ocean microbiome.* Science, 2015. **348**(6237): p.1261359.
2. Rognes, T., et al., *VSEARCH: a versatile open source tool for metagenomics.* PeerJ, 2016. **4**: p. e2584.

Please cite

Tangherlini M., Miralto M., Colantuono C., Sangiovanni M., Dell'Anno A., Corinaldesi C., Danovaro R., Chiusano M.L. "*GLOSSARY: the GLobal Ocean 16S Subunit web Accessible Resource*". 2018. *Submitted to BMC Bioinformatics*

For further information please contact us or visit our web site http://bioinfo.szn.it/