Tutorial: SEQenv_v0.8 Pipeline for linking sequences to the environment through text-mining

Umer Zeeshan Ijaz

The continuous drop in the associated costs combined with the increased efficiency of the latest high-throughput sequencing technologies has resulted in an unprecedented growth in sequencing projects. Ongoing endeavours such as the Earth Microbiome Project (Ref: www.earthmicrobiome.org) and the Ocean Sampling Day (Ref: www.microb3.eu/osd) are transcending national boundaries and are attempting to characterise the global microbial taxonomic and functional diversity for the benefit of mankind. The collection of sequencing information generated by such efforts is vital to shed light on the ecological features and the processes characterising different ecosystems, yet, the full knowledge discovery potential can only be unleashed if the associated meta data is also exploited to extract hidden patterns. For example, with the majority of genomes submitted to NCBI, there is an associated PubMed publication and in some cases there is a GenBank field called "isolation sources" that contains rich environmental information. With the advances in community-generated standards and the adherence to recommended annotation guidelines such as those of MIxS (Ref: gensc.org/gc_wiki/index.php/MIxS) of the Genomics Standards Consortium, it is now feasible to support intelligent queries and automated inference on such text resources. The Environmental Ontology (EnvO) (Ref: http://environmentontology.org/) will be a critical part of this approach as it gives the ontology for the concise, controlled description of environments. It thus provides structured and controlled vocabulary for the unified meta data annotation, and also serves as a source for naming environmental information. Thus, we have developed the SEQenv pipeline capable of annotating sequences with environment descriptive terms occurring within their records and/or in relevant literature. Given a set of sequences, SEQenv retrieves highly similar sequences from public repositories (NCBI GenBank). Subsequently, from each of these records, text fields carrying environmental context information (such as the reference title and the isolation source) are extracted. Additionally, the associated PubMed links are followed and the relevant abstracts are collected. Once the relevant pieces of text for each matching sequence have been gathered, they are then processed by a text mining module capable of identifying EnvO terms mentioned in them. The identified EnvO terms along with their frequencies of occurrence are then subjected to clustering analysis and multivariate statistics. As a result, tagclouds and heatmaps of environment descriptive terms characterizing different sequences/samples are generated. The SeqEnv pipeline can be applied to any set of nucleotide and protein sequences. Annotation of metagenomic samples, in particular 16S rRNA sequences is also supported.

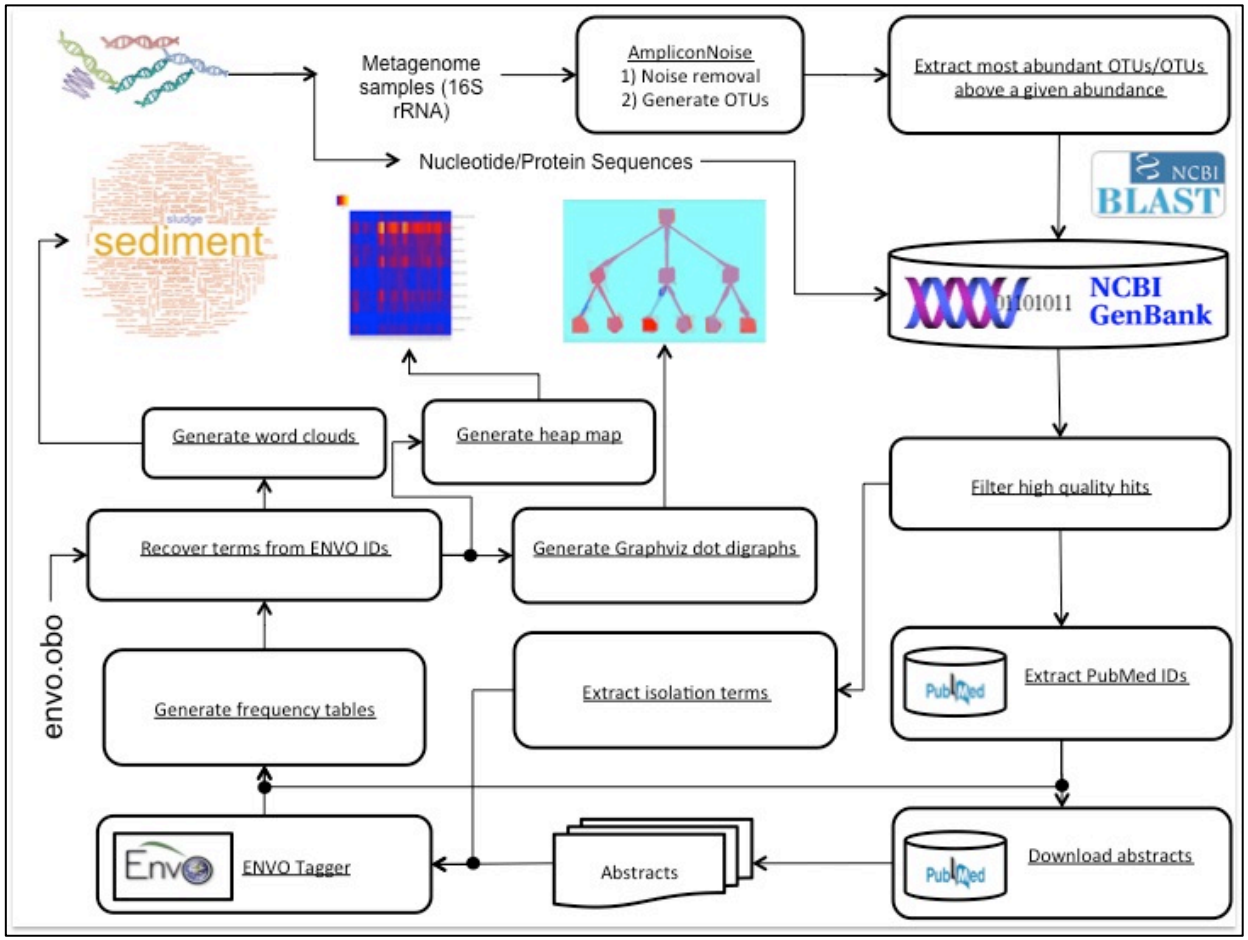


Figure 1: Workflow of SEQenv_v0.8

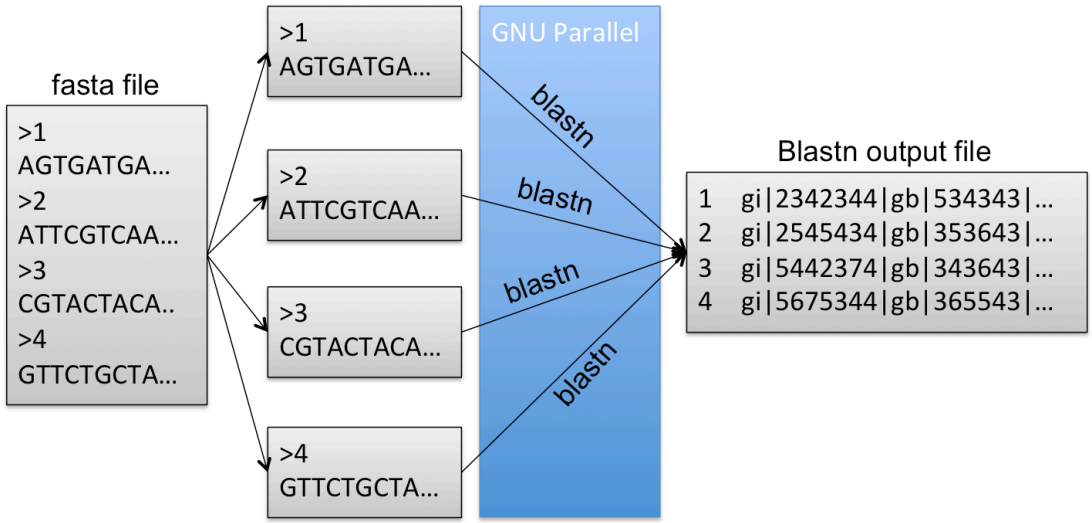


Figure 2: Splitting sequence files to speed up matches using GNU Parallel. Processing sequences through blastn is the most computationally intensive step in the pipeline. To minimize the execution time, we use GNU Parallel, a shell tool for executing jobs in parallel on multicore computers. We split the sequence file into fixed size chunks and then run blastn in parallel on these chunks on separate cores. We have tested the parallel version on an Illumina dataset comprising 6 million reads by matching against a reference dataset comprising 59 genomes. One run for a single reference match using 45 cores gave a running time of 2.5 minutes as compared to running a single instance of blastn that took 86 minutes on a 48 core server running CentOS operating system. For a 16SrRNA dataset comprising 1000 most abundant OTU sequences, matching at most 100 reference sequences against a local NCBI's NT database took 18.9 minutes on 45 cores. A speedup of 30 times or more is beneficial as the whole analysis can be done in few hours. We also use the same principle when searching for PubMed IDs and isolation terms for given sequences by splitting the blastn output files.

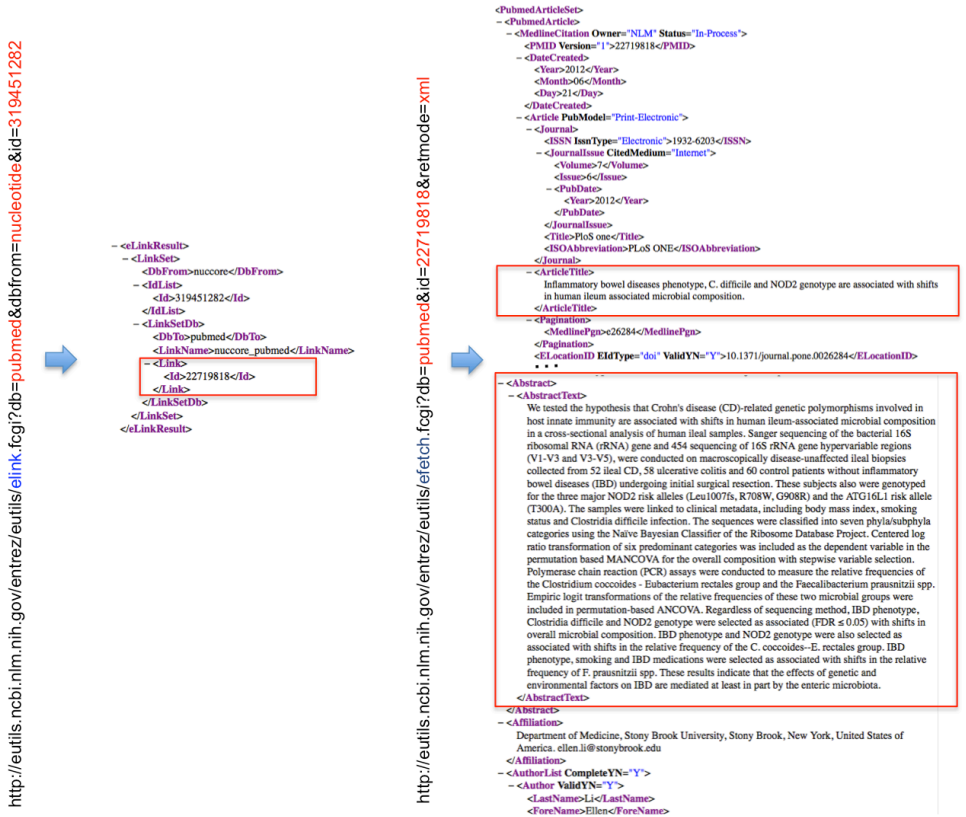


Figure 3: NCBI's E-utilities (Ref: http://www.ncbi.nlm.nih.gov/books/NBK25499/) allow the ability to communicate with the databases maintained at NCBI using HTTP POST and QUERY methods. Notable among these are Esearch, Epost, Esummary, and ELink services that are used in the pipeline to extract data. The data generated by the web requests can be retrieved in XML format as shown by the two example requests in the figure. Most popular languages support parsers to manipulate XML files and allow us to extract specific sections (highlighted with red rectangles). Examples of these parsers include XML::DOM in Perl, xml.dom.minidom in Python, org.w3c.dom and javax.xml.parsers in Java, and Xerces in C++ to name a few.

This work was developed in the following hackathons:

Event title: "From Signals to Environmentally Tagged Sequences II" (Ref: ECOST-MEETING-ES1103-100613-031037)
Location: Hellenic Centre for Marine Research, Crete, Greece
Dates: from 10-06-2013 to 13-06-2013

Event title: "From Signals to Environmentally Tagged Sequences" (Ref: ECOST-MEETING-ES1103-050912-018418)
Location: Hellenic Centre for Marine Research, Crete, Greece
Dates: from 27-09-2012 to 29-09-2012

For a list of contributors to this project, visit the page: http://envo.her.hcmr.gr/seqenv.html

Version 0.8 has the following features:

- We have two scripts, SEQenv_samples.sh (for processing 16SrRNA sequences when species abundance file (OTU table) is available), and SEQenv_sequences.sh (for processing nucleotide/protein sequences) (See Figure 1)
- In SEQenv_samples.sh, we have an additional -l switch to disregard weights in species abundance file. Initially, in the pipeline, we get an S X E OTU frequency table, and multiply that by N X S species abundance file to generate NXE samples frequency table. With -l switch N X S table is converted to have only 0/1 values thus removing weights.
- Both isolation sources (-t 1) and PubMed abstracts (-t 2) are supported in SEQenv_samples.sh and SEQenv_sequences.sh (See Figure 3)
- SEQenv_sequences.sh can now run on both protein and nucleotide sequences using either -s nucleotide or -s protein switch. With -s nucleotide switch, blastn is run against nt database, where as with -s protein switch, blastp is run against nr database.
- Filtering data is different for blastn and blastp. -perc_ident is only supported in blastn and not blastp. This doesn't cause any problem as we can still remove blast hits as both blastp and blastn export percentage identity in the hit file. The execution path is as follows:
  nucleotide sequences → run blastn with minimum percentage identity → filter out hits based on query coverage
  protein sequences → run blastp → filter out hits based on query coverage and percentage identity

- In SEQenv_samples.sh, we can now run the pipeline on most abundant OTUs (-o 1) but also on OTUs where column sum >= threshold (-o 2), i.e. chucking out rare OTUs. In both cases, you can provide the threshold using -n switch
- When you run the pipeline, a "document" folder is generated which contains PubMed abstracts when run with -t 2, and unique isolation sources when run with -t 1 setting
- Both SEQenv_samples.sh and SEQenv_sequences.sh generate a word cloud for overall community profiling. The png file is *_overall_labels.png* stored in the current folder.
- Both SEQenv_samples.sh and SEQenv_sequences.sh can run in parallel mode (See Figure 2)

Run the scripts without any arguments to get the usage information:

```
[seqenv@quince-srv2 ~]$ bash ~/SEQenv_v0.8/SEQenv_samples.sh
SEQenv -samples- Pipeline to link sequences to environmental descriptive terms when species abundance file is given

Usage:
    bash SEQenv_samples.sh -f <fasta_file.fasta> -s <species_abundance_file.csv> [options]

Options:
    -t Text source (1: GenBank record "isolation source" field, 2: PubMed abstracts (Default: 1))

    -l Presence/absence flag for species abundance file

    -p Parallelize using GNU Parallel flag
    -c Number of cores to use (Default: 10)

    -o Filtering method (1: -n most abundant OTUs, 2: minimum OTUs sum >= -n (Default: 1))
    -n Filtering threshold (Default: 1000)

    -m Minimum percentage identity in blastn (Default: 97)
    -q Minimum query coverage in blastn (Default: 97)
    -r Number of reference matches (Default: 10)

    -d Extract terms for the given ENVO ID (Default: all)
        all=Consider all terms
```

```
        Examples:
        ENVO:00010483=Environmental Material
        ENVO:00002297=Environmental Features
        ENVO:00000428=Biome
        ENVO:00002036=Habitat

[seqenv@quince-srv2 ~]$ bash ~/SEQenv_v0.8/SEQenv_sequences.sh
SEQenv -sequences- Pipeline to link nucleotide/protein sequences to environmental descriptive terms

Usage:
    bash SEQenv_sequences.sh -f <fasta_file.fasta> [options]
Options:
    -t Text source (1: GenBank record "isolation source" field, 2: PubMed abstracts (Default: 1))

    -p Parallelize using GNU Parallel
    -c Number of cores to use (Default: 10)

    -m Minimum percentage identity in blastn/blastp (Default: 97)
    -q Minimum query coverage in blastn/blastp (Default: 97)
    -r Number of reference matches (Default: 10)

    -s Sequence type (nucleotide/protein) (Default: nucloetide)
    -d Extract terms for the given ENVO ID (Default: all)
        all=Consider all terms
        Examples:
        ENVO:00010483=Environmental Material
        ENVO:00002297=Environmental Features
        ENVO:00000428=Biome
        ENVO:00002036=Habitat
```

We will first run the pipeline on a16S rRNA dataset using “isolation sources” as a text source. All_GoodT_C03.csv is species abundance file (3% OTUs) processed through AmpliconNoise software and All_GoodT_C03.fa contains the corresponding sequences for the OTUs.

```
[seqenv@quince-srv2 ~/SEQenv_samples_test]$ ls
```

```
All_GoodT_C03.csv
All_GoodT_C03.fa
[seqenv@quince-srv2 ~/SEQenv_samples_test]$ bash ~/SEQenv_v0.8/SEQenv_samples.sh -o 2 -n 1 -f All_GoodT_C03.fa -s
All_GoodT_C03.csv -m 99 -q 99 -r 100
[2013-07-20 04:38:43] SEQenv -samples- v0.8
[2013-07-20 04:38:43] Using  /home/opt/ncbi-blast-2.2.28+/bin/blastn
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/SEQenv_tagger/seqenv
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_parse_environments.sh
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_blast_concat_PMID3.py
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_get_linked_pmids.py
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_extract_ids_level.pl
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_gen_word_cloud.R
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo.obo
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_extract_records.pl
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_gen_heapmap.R
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_get_abstract.py
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/fastagrep.pl
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_blast_concat_isolation_terms.py
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_filter_blast.pl
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_gen_OTUs_freq.pl
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_gen_samples_freq.R
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_process_weight_matrix.R
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_extract_OTUs.R
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_gen_dot_files.pl
[2013-07-20 04:38:43] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_ids_to_terms.pl
[2013-07-20 04:38:43] STEP 1: Get sequences where minimum OTUs sum >= 1.
[2013-07-20 04:38:43] All_GoodT_C03_N1.fa already exists! Skipping this step.
[2013-07-20 04:38:43] STEP 2: Blast against NCBI's nt database with minimum percentage identity of 99%, maximum of
100 reference sequences, and evalue of 0.0001 in blastn.
[2013-07-20 04:38:43] All_GoodT_C03_N1_blast.out already exists! Skipping this step.
[2013-07-20 04:38:43] STEP 3: Filter out low quality hits with query coverage < 99%.
[2013-07-20 04:38:43] All_GoodT_C03_N1_blast_F.out already exists! Skipping this step.
[2013-07-20 04:38:43] STEP 4: Download data from NCBI.
[2013-07-20 04:40:46] Retrieved GenBank -isolation source- field linked to GIs. Unique entries are saved in the
documents folder.
[2013-07-20 04:40:46] STEP 5: Concatenate GenBank -isolation source- field IDs to blast file.
[2013-07-20 04:40:46] All_GoodT_C03_N1_blast_F_PMID.out already exists! Skipping this step.
```

```
[2013-07-20 04:40:46] STEP 6: Run SEQenv_tagger on the documents folder and generate ENVO hits file.
[2013-07-20 04:40:47] All_GoodT_C03_N1_blast_F_ENVO.txt is successfully generated.
[2013-07-20 04:40:47] STEP 7: Generate word cloud for overall community profile.
[2013-07-20 04:40:51] All_GoodT_C03_N1_blast_F_ENVO_overall_labels.png is successfully generated.
[2013-07-20 04:40:51] STEP 8: Generate frequency tables for OTUs.
[2013-07-20 04:40:51] All_GoodT_C03_N1_blast_F_ENVO_OTUs.csv is successfully generated using single document
matching and aggr document counting.
[2013-07-20 04:40:51] divrowsum normalization applied to All_GoodT_C03_N1_blast_F_ENVO_OTUs.csv successfully!
[2013-07-20 04:40:51] STEP 9: Generate frequency tables for samples.
[2013-07-20 04:40:52] All_GoodT_C03_N1_blast_F_ENVO_samples.csv is successfully.
[2013-07-20 04:40:52] STEP 10: Generate dot files for frequency tables of OTUs.
[2013-07-20 04:40:52] Processing C1735.csv
[2013-07-20 04:40:54] Processing C694.csv
[2013-07-20 04:40:55] Processing C696.csv
[2013-07-20 04:40:57] Processing C700.csv
[2013-07-20 04:40:59] Processing C710.csv
[2013-07-20 04:41:01] Processing C732.csv
[2013-07-20 04:41:02] Processing C733.csv
[2013-07-20 04:41:04] Processing C734.csv
[2013-07-20 04:41:06] Processing C753.csv
[2013-07-20 04:41:08] Processing C760.csv
[2013-07-20 04:41:09] Processing C768.csv
[2013-07-20 04:41:11] Processing C782.csv
[2013-07-20 04:41:12] Processing C812.csv
[2013-07-20 04:41:14] Processing C814.csv
[2013-07-20 04:41:16] Processing C821.csv
[2013-07-20 04:41:18] Processing C830.csv
[2013-07-20 04:41:20] Processing C1540.csv
[2013-07-20 04:41:22] Processing C1552.csv
[2013-07-20 04:41:23] Processing C1566.csv
[2013-07-20 04:41:25] Processing C1609.csv
[2013-07-20 04:41:27] Processing C1627.csv
[2013-07-20 04:41:29] Processing C1628.csv
[2013-07-20 04:41:30] Processing C1649.csv
[2013-07-20 04:41:32] Processing C1651.csv
[2013-07-20 04:41:34] Processing C1662.csv
[2013-07-20 04:41:36] Processing C1674.csv
```

```
[2013-07-20 04:41:37] Processing C1688.csv
[2013-07-20 04:41:39] Processing C1708.csv
[2013-07-20 04:41:41] Processing C888.csv
[2013-07-20 04:41:43] Processing C893.csv
[2013-07-20 04:41:44] Processing C899.csv
[2013-07-20 04:41:46] Processing C916.csv
[2013-07-20 04:41:48] Processing C917.csv
[2013-07-20 04:41:49] Processing C919.csv
[2013-07-20 04:41:51] Processing C920.csv
[2013-07-20 04:41:53] Processing C926.csv
[2013-07-20 04:41:55] Processing C927.csv
[2013-07-20 04:41:56] Processing C933.csv
[2013-07-20 04:41:58] Processing C954.csv
[2013-07-20 04:42:00] Processing C985.csv
[2013-07-20 04:42:01] Processing C999.csv
[2013-07-20 04:42:03] Processing C1001.csv
[2013-07-20 04:42:05] Processing C1021.csv
[2013-07-20 04:42:07] Processing C1024.csv
[2013-07-20 04:42:09] Processing C1036.csv
[2013-07-20 04:42:11] Processing C1038.csv
[2013-07-20 04:42:12] Processing C1048.csv
[2013-07-20 04:42:14] Processing C1067.csv
[2013-07-20 04:42:16] Processing C1079.csv
[2013-07-20 04:42:18] Processing C1098.csv
[2013-07-20 04:42:20] Processing C1126.csv
[2013-07-20 04:42:22] Processing C1149.csv
[2013-07-20 04:42:24] Processing C1167.csv
[2013-07-20 04:42:25] Processing C1184.csv
[2013-07-20 04:42:27] Processing C1194.csv
[2013-07-20 04:42:29] Processing C1201.csv
[2013-07-20 04:42:30] Processing C1203.csv
[2013-07-20 04:42:32] Processing C1235.csv
[2013-07-20 04:42:34] Processing C1246.csv
[2013-07-20 04:42:36] Processing C1296.csv
[2013-07-20 04:42:37] Processing C1337.csv
[2013-07-20 04:42:39] Processing C1349.csv
[2013-07-20 04:42:40] Processing C1444.csv
```

```
[2013-07-20 04:42:42] Processing C1479.csv
[2013-07-20 04:42:44] Processing C412.csv
[2013-07-20 04:42:46] Processing C435.csv
[2013-07-20 04:42:47] Processing C446.csv
[2013-07-20 04:42:49] Processing C461.csv
[2013-07-20 04:42:51] Processing C474.csv
[2013-07-20 04:42:52] Processing C477.csv
[2013-07-20 04:42:54] Processing C478.csv
[2013-07-20 04:42:56] Processing C482.csv
[2013-07-20 04:42:58] Processing C495.csv
[2013-07-20 04:42:59] Processing C497.csv
[2013-07-20 04:43:01] Processing C506.csv
[2013-07-20 04:43:03] Processing C508.csv
[2013-07-20 04:43:05] Processing C511.csv
[2013-07-20 04:43:07] Processing C516.csv
[2013-07-20 04:43:08] Processing C524.csv
[2013-07-20 04:43:10] Processing C529.csv
[2013-07-20 04:43:12] Processing C543.csv
[2013-07-20 04:43:13] Processing C2.csv
[2013-07-20 04:43:15] Processing C3.csv
[2013-07-20 04:43:17] Processing C5.csv
[2013-07-20 04:43:18] Processing C14.csv
[2013-07-20 04:43:20] Processing C15.csv
[2013-07-20 04:43:21] Processing C22.csv
[2013-07-20 04:43:23] Processing C25.csv
[2013-07-20 04:43:25] Processing C26.csv
[2013-07-20 04:43:27] Processing C27.csv
[2013-07-20 04:43:29] Processing C30.csv
[2013-07-20 04:43:31] Processing C33.csv
[2013-07-20 04:43:32] Processing C35.csv
[2013-07-20 04:43:34] Processing C37.csv
[2013-07-20 04:43:36] Processing C41.csv
[2013-07-20 04:43:38] Processing C42.csv
[2013-07-20 04:43:39] Processing C45.csv
[2013-07-20 04:43:41] Processing C47.csv
[2013-07-20 04:43:43] Processing C58.csv
[2013-07-20 04:43:45] Processing C60.csv
```

```
[2013-07-20 04:43:47] Processing C64.csv
[2013-07-20 04:43:48] Processing C65.csv
[2013-07-20 04:43:50] Processing C70.csv
[2013-07-20 04:43:52] Processing C75.csv
[2013-07-20 04:43:54] Processing C78.csv
[2013-07-20 04:43:55] Processing C80.csv
[2013-07-20 04:43:57] Processing C83.csv
[2013-07-20 04:43:59] Processing C85.csv
[2013-07-20 04:44:01] Processing C89.csv
[2013-07-20 04:44:03] Processing C94.csv
[2013-07-20 04:44:05] Processing C96.csv
[2013-07-20 04:44:07] Processing C98.csv
[2013-07-20 04:44:08] Processing C109.csv
[2013-07-20 04:44:10] Processing C111.csv
[2013-07-20 04:44:12] Processing C112.csv
[2013-07-20 04:44:13] Processing C116.csv
[2013-07-20 04:44:15] Processing C119.csv
[2013-07-20 04:44:17] Processing C128.csv
[2013-07-20 04:44:18] Processing C129.csv
[2013-07-20 04:44:20] Processing C264.csv
[2013-07-20 04:44:22] Processing C268.csv
[2013-07-20 04:44:24] Processing C272.csv
[2013-07-20 04:44:26] Processing C275.csv
[2013-07-20 04:44:27] Processing C283.csv
[2013-07-20 04:44:29] Processing C286.csv
[2013-07-20 04:44:31] Processing C287.csv
[2013-07-20 04:44:32] Processing C290.csv
[2013-07-20 04:44:34] Processing C298.csv
[2013-07-20 04:44:36] Processing C304.csv
[2013-07-20 04:44:37] Processing C317.csv
[2013-07-20 04:44:39] Processing C321.csv
[2013-07-20 04:44:41] Processing C323.csv
[2013-07-20 04:44:42] Processing C324.csv
[2013-07-20 04:44:44] Processing C328.csv
[2013-07-20 04:44:46] Processing C344.csv
[2013-07-20 04:44:47] Processing C359.csv
[2013-07-20 04:44:49] Processing C360.csv
```

```
[2013-07-20 04:44:51] Processing C364.csv
[2013-07-20 04:44:52] Processing C366.csv
[2013-07-20 04:44:55] Processing C368.csv
[2013-07-20 04:44:56] Processing C370.csv
[2013-07-20 04:44:58] Processing C372.csv
[2013-07-20 04:45:00] Processing C376.csv
[2013-07-20 04:45:02] Processing C377.csv
[2013-07-20 04:45:04] Processing C378.csv
[2013-07-20 04:45:06] Processing C384.csv
[2013-07-20 04:45:07] Processing C392.csv
[2013-07-20 04:45:09] Processing C397.csv
[2013-07-20 04:45:11] Processing C552.csv
[2013-07-20 04:45:13] Processing C555.csv
[2013-07-20 04:45:14] Processing C556.csv
[2013-07-20 04:45:16] Processing C557.csv
[2013-07-20 04:45:18] Processing C580.csv
[2013-07-20 04:45:20] Processing C581.csv
[2013-07-20 04:45:21] Processing C582.csv
[2013-07-20 04:45:23] Processing C584.csv
[2013-07-20 04:45:25] Processing C585.csv
[2013-07-20 04:45:26] Processing C586.csv
[2013-07-20 04:45:28] Processing C588.csv
[2013-07-20 04:45:30] Processing C593.csv
[2013-07-20 04:45:32] Processing C598.csv
[2013-07-20 04:45:33] Processing C602.csv
[2013-07-20 04:45:35] Processing C604.csv
[2013-07-20 04:45:37] Processing C608.csv
[2013-07-20 04:45:39] Processing C609.csv
[2013-07-20 04:45:41] Processing C611.csv
[2013-07-20 04:45:43] Processing C635.csv
[2013-07-20 04:45:44] Processing C641.csv
[2013-07-20 04:45:47] Processing C644.csv
[2013-07-20 04:45:48] Processing C660.csv
[2013-07-20 04:45:50] Processing C666.csv
[2013-07-20 04:45:52] Processing C670.csv
[2013-07-20 04:45:53] Processing C683.csv
[2013-07-20 04:45:56] Processing C154.csv
```

```
[2013-07-20 04:45:57] Processing C163.csv
[2013-07-20 04:45:59] Processing C164.csv
[2013-07-20 04:46:01] Processing C166.csv
[2013-07-20 04:46:02] Processing C171.csv
[2013-07-20 04:46:04] Processing C173.csv
[2013-07-20 04:46:06] Processing C185.csv
[2013-07-20 04:46:08] Processing C186.csv
[2013-07-20 04:46:10] Processing C189.csv
[2013-07-20 04:46:11] Processing C190.csv
[2013-07-20 04:46:13] Processing C192.csv
[2013-07-20 04:46:15] Processing C206.csv
[2013-07-20 04:46:16] Processing C207.csv
[2013-07-20 04:46:18] Processing C213.csv
[2013-07-20 04:46:20] Processing C218.csv
[2013-07-20 04:46:22] Processing C229.csv
[2013-07-20 04:46:24] Processing C233.csv
[2013-07-20 04:46:25] Processing C234.csv
[2013-07-20 04:46:27] Processing C235.csv
[2013-07-20 04:46:29] Processing C238.csv
[2013-07-20 04:46:31] Processing C243.csv
[2013-07-20 04:46:33] Processing C253.csv
[2013-07-20 04:46:35] Processing C255.csv
[2013-07-20 04:46:37] Processing C263.csv
[2013-07-20 04:46:38] Folder OTUs_dot is successfully generated.
[2013-07-20 04:46:38] STEP 11: Generate dot files for frequency tables of samples.
[2013-07-20 04:46:40] Folder samples_dot is successfully generated.
[2013-07-20 04:46:40] STEP 12: Generate labels for frequency tables of OTUs (all).
[2013-07-20 04:46:42] All_GoodT_C03_N1_blast_F_ENVO_OTUs_labels.csv is successfully generated.
[2013-07-20 04:46:42] STEP 13: Generate labels for frequency tables of samples (all).
[2013-07-20 04:46:44] All_GoodT_C03_N1_blast_F_ENVO_samples_labels.csv is successfully generated.
[2013-07-20 04:46:44] STEP 14: Generate word clouds for samples (all).
[2013-07-20 04:47:05] Folder samples_wc is successfully generated.
[2013-07-20 04:47:05] STEP 15: Generate word clouds for OTUs (all).
[2013-07-20 04:47:43] Folder OTUs_wc is successfully generated.
[2013-07-20 04:47:43] Generate heapmap for samples (all).
[2013-07-20 04:47:48] Generated All_GoodT_C03_N1_blast_F_ENVO_samples_labels.png successfully.
[2013-07-20 04:47:48] Folder samples_heapmaps is successfully generated.
```

```
[2013-07-20 04:47:48] Finished processing!
```

Once the pipeline has finished processing, you will have the following contents in the current folder:

```
[seqenv@quince-srv2 ~/SEQenv_samples_test]$ ls
documents
OTUs_dot
OTUs_wc
samples_dot
samples_heapmaps
samples_wc
All_GoodT_C03.csv
All_GoodT_C03.fa
All_GoodT_C03_N1.fa
All_GoodT_C03_N1_blast.out
All_GoodT_C03_N1_blast_F.out
All_GoodT_C03_N1_blast_F_PMID.out
All_GoodT_C03_N1_blast_F_ENVO.txt
All_GoodT_C03_N1_blast_F_ENVO_OTUs.csv
All_GoodT_C03_N1_blast_F_ENVO_OTUs_labels.csv
All_GoodT_C03_N1_blast_F_ENVO_overall.csv
All_GoodT_C03_N1_blast_F_ENVO_overall_labels.csv
All_GoodT_C03_N1_blast_F_ENVO_overall_labels.png
All_GoodT_C03_N1_blast_F_ENVO_samples.csv
All_GoodT_C03_N1_blast_F_ENVO_samples_labels.csv
SEQenv.log
```

The “documents” folder contains all the isolation sources for the reference matches. The IDs have no significance as they were generated to keep a record of unique isolation sources.

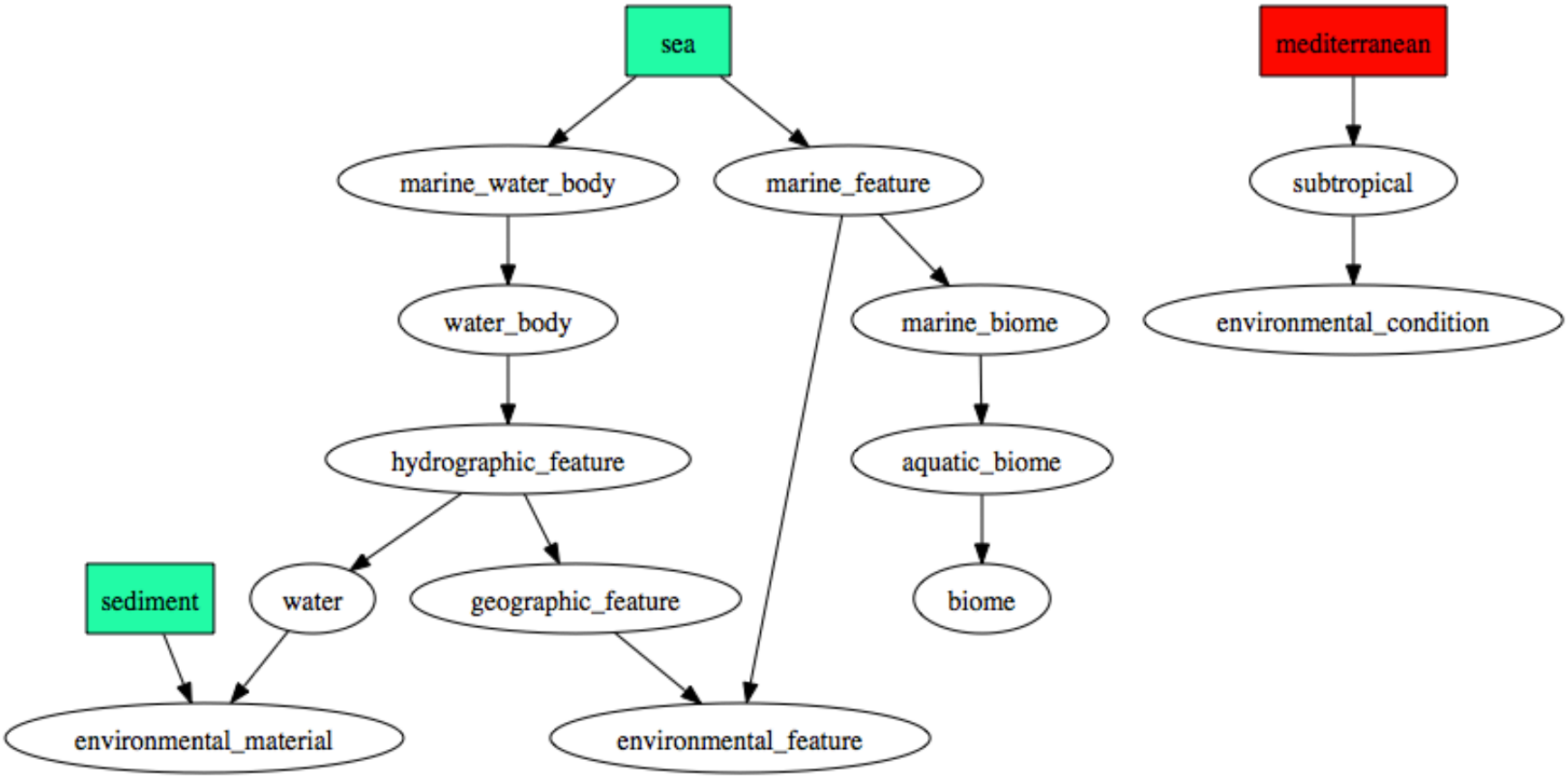
```
[seqenv@quince-srv2 ~/SEQenv_samples_test]$ ls documents/*
documents/IS_1.txt      documents/IS_195.txt    documents/IS_290.txt
documents/IS_10.txt     documents/IS_196.txt    documents/IS_291.txt
documents/IS_100.txt    documents/IS_197.txt    documents/IS_292.txt
documents/IS_101.txt    documents/IS_198.txt    documents/IS_293.txt
documents/IS_102.txt    documents/IS_199.txt    documents/IS_294.txt
```

```
documents/IS_103.txt    documents/IS_2.txt      documents/IS_295.txt
documents/IS_104.txt    documents/IS_20.txt     documents/IS_296.txt
documents/IS_105.txt    documents/IS_200.txt    documents/IS_297.txt
documents/IS_106.txt    documents/IS_201.txt    documents/IS_298.txt
documents/IS_107.txt    documents/IS_202.txt    documents/IS_299.txt
documents/IS_108.txt    documents/IS_203.txt    documents/IS_3.txt
documents/IS_109.txt    documents/IS_204.txt    documents/IS_30.txt
documents/IS_11.txt     documents/IS_205.txt    documents/IS_300.txt
documents/IS_110.txt    documents/IS_206.txt    documents/IS_301.txt
documents/IS_111.txt    documents/IS_207.txt    documents/IS_302.txt
documents/IS_112.txt    documents/IS_208.txt    documents/IS_303.txt
documents/IS_113.txt    documents/IS_209.txt    documents/IS_304.txt
documents/IS_114.txt    documents/IS_21.txt     documents/IS_305.txt
documents/IS_115.txt    documents/IS_210.txt    documents/IS_306.txt
documents/IS_116.txt    documents/IS_211.txt    documents/IS_307.txt
documents/IS_117.txt    documents/IS_212.txt    documents/IS_308.txt
documents/IS_118.txt    documents/IS_213.txt    documents/IS_309.txt
documents/IS_119.txt    documents/IS_214.txt    documents/IS_31.txt
documents/IS_12.txt     documents/IS_215.txt    documents/IS_310.txt
documents/IS_120.txt    documents/IS_216.txt    documents/IS_311.txt
documents/IS_121.txt    documents/IS_217.txt    documents/IS_312.txt
documents/IS_122.txt    documents/IS_218.txt    documents/IS_313.txt
documents/IS_123.txt    documents/IS_219.txt    documents/IS_314.txt
documents/IS_124.txt    documents/IS_22.txt     documents/IS_315.txt
documents/IS_125.txt    documents/IS_220.txt    documents/IS_316.txt
documents/IS_126.txt    documents/IS_221.txt    documents/IS_32.txt
documents/IS_127.txt    documents/IS_222.txt    documents/IS_33.txt
documents/IS_128.txt    documents/IS_223.txt    documents/IS_34.txt
documents/IS_129.txt    documents/IS_224.txt    documents/IS_35.txt
documents/IS_13.txt     documents/IS_225.txt    documents/IS_36.txt
documents/IS_130.txt    documents/IS_226.txt    documents/IS_37.txt
documents/IS_131.txt    documents/IS_227.txt    documents/IS_38.txt
documents/IS_132.txt    documents/IS_228.txt    documents/IS_39.txt
documents/IS_133.txt    documents/IS_229.txt    documents/IS_4.txt
documents/IS_134.txt    documents/IS_23.txt     documents/IS_40.txt
documents/IS_135.txt    documents/IS_230.txt    documents/IS_41.txt
documents/IS_136.txt    documents/IS_231.txt    documents/IS_42.txt
```

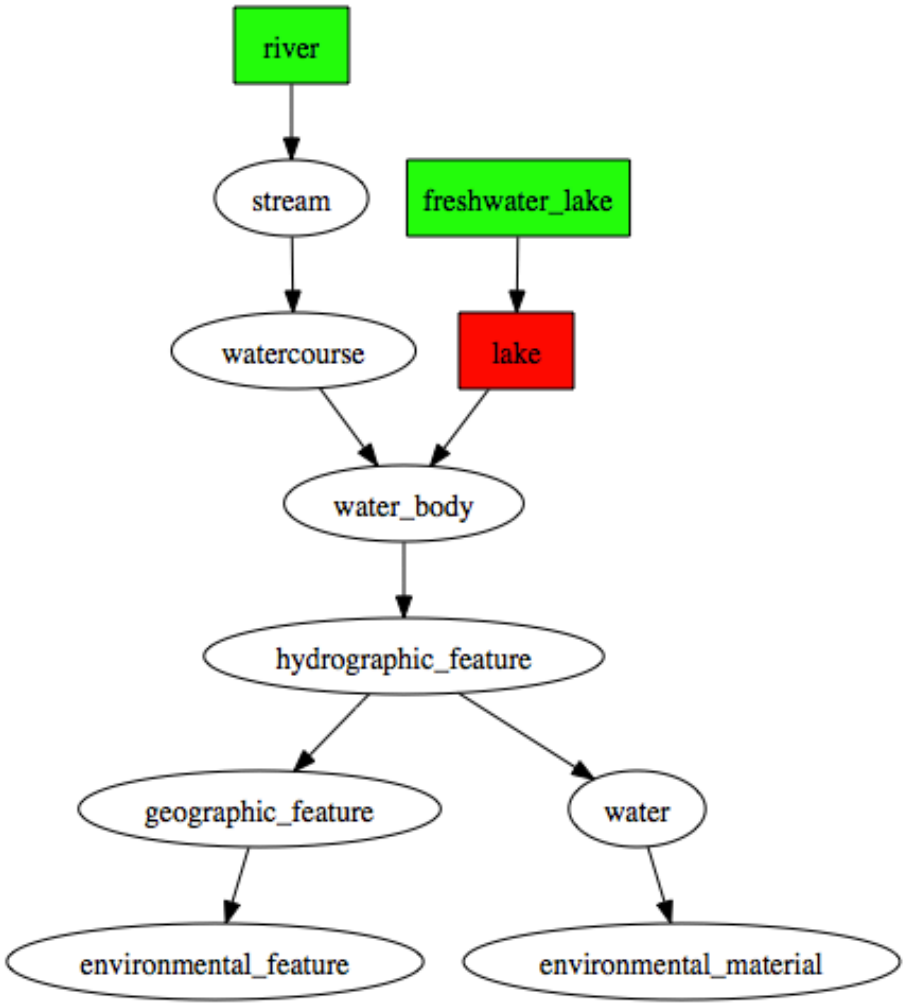
```
documents/IS_137.txt    documents/IS_232.txt    documents/IS_43.txt
documents/IS_138.txt    documents/IS_233.txt    documents/IS_44.txt
documents/IS_139.txt    documents/IS_234.txt    documents/IS_45.txt
documents/IS_14.txt     documents/IS_235.txt    documents/IS_46.txt
documents/IS_140.txt    documents/IS_236.txt    documents/IS_47.txt
documents/IS_141.txt    documents/IS_237.txt    documents/IS_48.txt
documents/IS_142.txt    documents/IS_238.txt    documents/IS_49.txt
documents/IS_143.txt    documents/IS_239.txt    documents/IS_5.txt
documents/IS_144.txt    documents/IS_24.txt     documents/IS_50.txt
documents/IS_145.txt    documents/IS_240.txt    documents/IS_51.txt
documents/IS_146.txt    documents/IS_241.txt    documents/IS_52.txt
documents/IS_147.txt    documents/IS_242.txt    documents/IS_53.txt
documents/IS_148.txt    documents/IS_243.txt    documents/IS_54.txt
documents/IS_149.txt    documents/IS_244.txt    documents/IS_55.txt
documents/IS_15.txt     documents/IS_245.txt    documents/IS_56.txt
documents/IS_150.txt    documents/IS_246.txt    documents/IS_57.txt
documents/IS_151.txt    documents/IS_247.txt    documents/IS_58.txt
documents/IS_152.txt    documents/IS_248.txt    documents/IS_59.txt
documents/IS_153.txt    documents/IS_249.txt    documents/IS_6.txt
documents/IS_154.txt    documents/IS_25.txt     documents/IS_60.txt
documents/IS_155.txt    documents/IS_250.txt    documents/IS_61.txt
documents/IS_156.txt    documents/IS_251.txt    documents/IS_62.txt
documents/IS_157.txt    documents/IS_252.txt    documents/IS_63.txt
documents/IS_158.txt    documents/IS_253.txt    documents/IS_64.txt
documents/IS_159.txt    documents/IS_254.txt    documents/IS_65.txt
documents/IS_16.txt     documents/IS_255.txt    documents/IS_66.txt
documents/IS_160.txt    documents/IS_256.txt    documents/IS_67.txt
documents/IS_161.txt    documents/IS_257.txt    documents/IS_68.txt
documents/IS_162.txt    documents/IS_258.txt    documents/IS_69.txt
documents/IS_163.txt    documents/IS_259.txt    documents/IS_7.txt
documents/IS_164.txt    documents/IS_26.txt     documents/IS_70.txt
documents/IS_165.txt    documents/IS_260.txt    documents/IS_71.txt
documents/IS_166.txt    documents/IS_261.txt    documents/IS_72.txt
documents/IS_167.txt    documents/IS_262.txt    documents/IS_73.txt
documents/IS_168.txt    documents/IS_263.txt    documents/IS_74.txt
documents/IS_169.txt    documents/IS_264.txt    documents/IS_75.txt
documents/IS_17.txt     documents/IS_265.txt    documents/IS_76.txt
```

```
documents/IS_170.txt    documents/IS_266.txt    documents/IS_77.txt
documents/IS_171.txt    documents/IS_267.txt    documents/IS_78.txt
documents/IS_172.txt    documents/IS_268.txt    documents/IS_79.txt
documents/IS_173.txt    documents/IS_269.txt    documents/IS_8.txt
documents/IS_174.txt    documents/IS_27.txt     documents/IS_80.txt
documents/IS_175.txt    documents/IS_270.txt    documents/IS_81.txt
documents/IS_176.txt    documents/IS_271.txt    documents/IS_82.txt
documents/IS_177.txt    documents/IS_272.txt    documents/IS_83.txt
documents/IS_178.txt    documents/IS_273.txt    documents/IS_84.txt
documents/IS_179.txt    documents/IS_274.txt    documents/IS_85.txt
documents/IS_18.txt     documents/IS_275.txt    documents/IS_86.txt
documents/IS_180.txt    documents/IS_276.txt    documents/IS_87.txt
documents/IS_181.txt    documents/IS_277.txt    documents/IS_88.txt
documents/IS_182.txt    documents/IS_278.txt    documents/IS_89.txt
documents/IS_183.txt    documents/IS_279.txt    documents/IS_9.txt
documents/IS_184.txt    documents/IS_28.txt     documents/IS_90.txt
documents/IS_185.txt    documents/IS_280.txt    documents/IS_91.txt
documents/IS_186.txt    documents/IS_281.txt    documents/IS_92.txt
documents/IS_187.txt    documents/IS_282.txt    documents/IS_93.txt
documents/IS_188.txt    documents/IS_283.txt    documents/IS_94.txt
documents/IS_189.txt    documents/IS_284.txt    documents/IS_95.txt
documents/IS_19.txt     documents/IS_285.txt    documents/IS_96.txt
documents/IS_190.txt    documents/IS_286.txt    documents/IS_97.txt
documents/IS_191.txt    documents/IS_287.txt    documents/IS_98.txt
documents/IS_192.txt    documents/IS_288.txt    documents/IS_99.txt
documents/IS_193.txt    documents/IS_289.txt
documents/IS_194.txt    documents/IS_29.txt
```

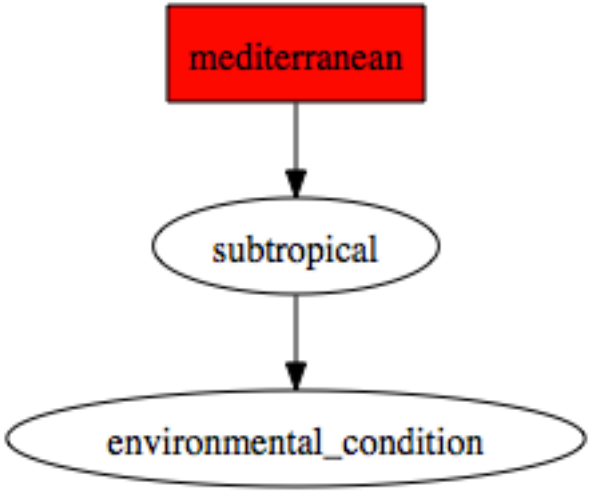
We are particularly interested in 3 OTUs: C15, C26, and C89 that hold importance in this dataset. We will go to “OTUs_dot” folder and open C15.dot, C26.dot, and C89.dot in GraphViz (by right-clicking). In the graphs shown below, the observed terms are drawn as boxes and unobserved terms in the lineage are drawn as ellipses. The boxes are then coloured based on their frequency from blue to red with blue corresponding to low frequency terms and red corresponding to high frequency terms, respectively.


sea
mediterranean
marine_water_body
marine_feature
subtropical
water_body
marine_biome
environmental_condition
hydrographic_feature
aquatic_biome
sediment
water
geographic_feature
biome
environmental_material
environmental_feature

C15

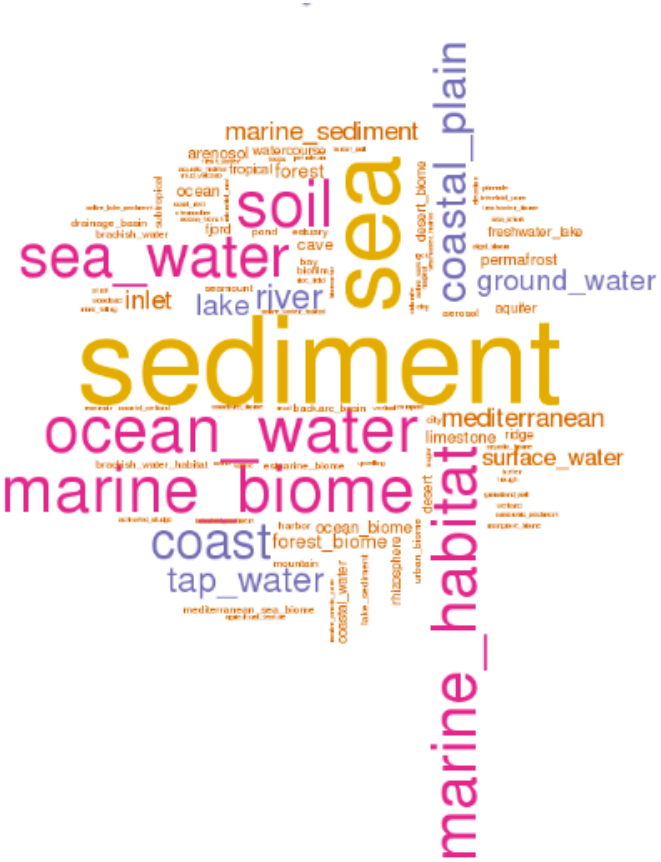

river
stream
freshwater_lake
watercourse
lake
water_body
hydrographic_feature
geographic_feature
water
environmental_feature
environmental_material

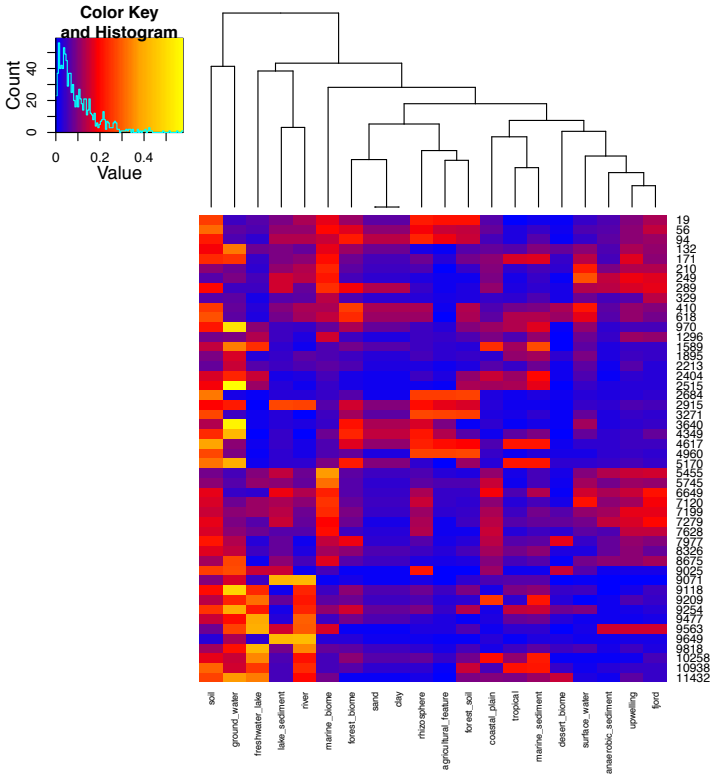
C26


mediterranean
subtropical
environmental_condition

C89

Let us now look at the overall profile word cloud (All_GoodT_C03_N1_blast_F_ENVO_overall_labels.png) and the heapmap generated in the “samples_heapmap” folder. The heapmap is useful for differentiating/grouping samples based on EnvO frequencies.





The folders “OTUs_dot” and “OTUs_wc” contain digraphs and word clouds for each OTU, respectively. Similarly, the folders “samples_dot” and “samples_wc” contain digraphs and word clouds for each sample, respectively. To further process your data, you can use the frequency tables All_GoodT_C03_N1_blast_F_ENVO_OTUs_labels.csv and All_GoodT_C03_N1_blast_F_ENVO_OTUs_labels.csv for multivariate statistical analysis. You can also process them in our TAXAenv pipeline: http://quince-srv2.eng.gla.ac.uk:8080/ Tutorial: http://userweb.eng.gla.ac.uk/umer.ijaz/TAXAenv_tutorial.pdf

If you run the pipeline with –t 2 switch, i.e. using PubMed abstracts as a text source, you will get the similar contents except that the “documents” will contain the PubMed abstracts with their names corresponding to PubMed IDs.

```
[seqenv@quince-srv2 ~/SEQenv_samples_test]$ ls documents/*
documents/10188281.txt  documents/16907756.txt  documents/20865054.txt
documents/10568038.txt  documents/16937025.txt  documents/20874732.txt
documents/10586879.txt  documents/16982820.txt  documents/20875064.txt
documents/10603268.txt  documents/17051209.txt  documents/20880034.txt
documents/10617197.txt  documents/17061822.txt  documents/20941906.txt
documents/10724520.txt  documents/17080290.txt  documents/20943863.txt
documents/10843083.txt  documents/17116968.txt  documents/20980660.txt
documents/10931912.txt  documents/17117985.txt  documents/21057784.txt
documents/10941928.txt  documents/17117990.txt  documents/21148393.txt
documents/11010891.txt  documents/17216330.txt  documents/21189294.txt
documents/11130713.txt  documents/17227412.txt  documents/21229314.txt
documents/11146245.txt  documents/17286994.txt  documents/21239227.txt
documents/11214317.txt  documents/17403162.txt  documents/2124631.txt
documents/11246408.txt  documents/17486969.txt  documents/21303395.txt
documents/11425705.txt  documents/17486973.txt  documents/21304582.txt
documents/11504944.txt  documents/17486974.txt  documents/21326234.txt
documents/11523007.txt  documents/17486977.txt  documents/21332876.txt
documents/11557979.txt  documents/17486979.txt  documents/21380776.txt
documents/11596104.txt  documents/17507782.txt  documents/21390076.txt
documents/11697917.txt  documents/17530165.txt  documents/21462550.txt
documents/11751247.txt  documents/17553708.txt  documents/21523192.txt
documents/11760963.txt  documents/17572336.txt  documents/21554513.txt
documents/11884166.txt  documents/17576099.txt  documents/21558214.txt
documents/12054219.txt  documents/17726297.txt  documents/21566389.txt
```

```
documents/12116929.txt  documents/17899193.txt  documents/21628300.txt
documents/12125757.txt  documents/17903218.txt  documents/21646174.txt
documents/12396494.txt  documents/17944708.txt  documents/21669670.txt
documents/12469312.txt  documents/18043670.txt  documents/21672740.txt
documents/12501365.txt  documents/18180750.txt  documents/21741879.txt
documents/12501399.txt  documents/18218029.txt  documents/21873489.txt
documents/12503682.txt  documents/18266757.txt  documents/22103931.txt
documents/12542710.txt  documents/18368437.txt  documents/22192529.txt
documents/12682873.txt  documents/18469120.txt  documents/22208605.txt
documents/12702293.txt  documents/18490993.txt  documents/22239643.txt
documents/12716990.txt  documents/18544097.txt  documents/22242889.txt
documents/12732534.txt  documents/18545660.txt  documents/22247428.txt
documents/12735797.txt  documents/18584522.txt  documents/22344659.txt
documents/12839741.txt  documents/18604577.txt  documents/22403476.txt
documents/12892148.txt  documents/18621084.txt  documents/22453118.txt
documents/12908085.txt  documents/18693068.txt  documents/22454494.txt
documents/1435237.txt   documents/18759218.txt  documents/22541864.txt
documents/14660370.txt  documents/18768211.txt  documents/22568577.txt
documents/14740910.txt  documents/18771501.txt  documents/22591022.txt
documents/14747975.txt  documents/18801046.txt  documents/22654619.txt
documents/15105500.txt  documents/18820685.txt  documents/22658831.txt
documents/15164237.txt  documents/18984039.txt  documents/22685143.txt
documents/15184153.txt  documents/1901093.txt   documents/22703332.txt
documents/15184155.txt  documents/19128038.txt  documents/22808282.txt
documents/15272195.txt  documents/19150987.txt  documents/22969752.txt
documents/15305792.txt  documents/19172216.txt  documents/23041269.txt
documents/15305796.txt  documents/19189423.txt  documents/23194719.txt
documents/15449591.txt  documents/19465529.txt  documents/23196114.txt
documents/15522504.txt  documents/19467154.txt  documents/23228065.txt
documents/15528550.txt  documents/19484305.txt  documents/23247917.txt
documents/15545489.txt  documents/19515203.txt  documents/23261712.txt
documents/15546421.txt  documents/19539760.txt  documents/23278436.txt
documents/15552061.txt  documents/19583789.txt  documents/23281331.txt
documents/15587708.txt  documents/19641535.txt  documents/23496985.txt
documents/15683926.txt  documents/19749031.txt  documents/23531052.txt
documents/15736863.txt  documents/19767459.txt  documents/23761307.txt
documents/16014017.txt  documents/19799618.txt  documents/2422931.txt
```

```
documents/16044243.txt  documents/19893617.txt  documents/3418693.txt
documents/16104864.txt  documents/20002178.txt  documents/3475703.txt
documents/16156732.txt  documents/20056613.txt  documents/3689320.txt
documents/16171188.txt  documents/2007550.txt   documents/4091818.txt
documents/16204507.txt  documents/20102745.txt  documents/7773393.txt
documents/16292522.txt  documents/20127114.txt  documents/7804250.txt
documents/16309395.txt  documents/20163477.txt  documents/7816825.txt
documents/16329867.txt  documents/20169022.txt  documents/8078407.txt
documents/16338764.txt  documents/20169024.txt  documents/8487639.txt
documents/16353640.txt  documents/20169026.txt  documents/8572692.txt
documents/16389967.txt  documents/20336290.txt  documents/8583907.txt
documents/16405292.txt  documents/20360212.txt  documents/8840501.txt
documents/16427147.txt  documents/20393846.txt  documents/8896371.txt
documents/16597941.txt  documents/20396576.txt  documents/8934908.txt
documents/16598157.txt  documents/20482740.txt  documents/8972871.txt
documents/16672445.txt  documents/20561018.txt  documents/8976608.txt
documents/16672511.txt  documents/20597984.txt  documents/9049276.txt
documents/16677346.txt  documents/20660211.txt  documents/9149422.txt
documents/16691324.txt  documents/20668244.txt  documents/9225445.txt
documents/16691328.txt  documents/20806248.txt  documents/9342352.txt
documents/16778350.txt  documents/20826189.txt  documents/9495032.txt
documents/16820449.txt  documents/20846815.txt  documents/9542099.txt
documents/16885296.txt  documents/20851993.txt  documents/9739550.txt
```

SEQenv_sequences.sh follows the similar workflow as SEQenv_samples.sh. For example, given 80 dummy nucleotide sequences (you can also process protein sequences) in FASTA format as deg_species_filtered.fna, we will run the pipeline as follows:

```
[seqenv@quince-srv2 ~/uzi/test_SEQenv]$ ls
deg_species_filtered.fna
[seqenv@quince-srv2 ~/uzi/test_SEQenv]$
[seqenv@quince-srv2 ~/uzi/test_SEQenv]$ bash ~/SEQenv_v0.8/SEQenv_samples.sh -t 1 -p -c 10 -f
deg_species_filtered.fna –s nucleotide
[2013-07-18 20:36:39] SEQenv -sequences- v0.8
[2013-07-18 20:36:39] Using  /home/opt/ncbi-blast-2.2.28+/bin/blastn
[2013-07-18 20:36:39] Using  /home/opt/ncbi-blast-2.2.28+/bin/blastp
```

```
[2013-07-18 20:36:39] Using  parallel
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/SEQenv_tagger/seqenv
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_parse_environments.sh
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_blast_concat_PMID3.py
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_get_linked_pmids.py
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_extract_ids_level.pl
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_gen_word_cloud.R
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo.obo
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_extract_records.pl
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_gen_heapmap.R
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_get_abstract.py
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/fastagrep.pl
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_blast_concat_isolation_terms.py
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_filter_blast.pl
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_gen_OTUs_freq.pl
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_gen_samples_freq.R
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_process_weight_matrix.R
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_extract_abundant_OTUs.R
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_gen_dot_files.pl
[2013-07-18 20:36:39] Using  /home/seqenv/SEQenv_v0.8/scripts/envo_ids_to_terms.pl
[2013-07-18 20:36:39] STEP 1: Generate mappings for sequence headers in FASTA file.
[2013-07-18 20:36:39] deg_species_filtered_M.map and deg_species_filtered_M.fa are successfully generated.
[2013-07-18 20:36:39] STEP 2: Blast against NCBI's nt database with minimum percentage identity of 97%, maximum of
100 reference sequences, and evalue of 0.0001 in blastn.
[2013-07-18 20:49:30] blastn using GNU parallel took 771 seconds to generate deg_species_filtered_M_blast'.out' from
deg_species_filtered_M.fa.
[2013-07-18 20:49:30] STEP 3: Filter out low quality hits with query coverage < 97%.
[2013-07-18 20:49:30] deg_species_filtered_M_blast_F.out is successfully generated.
[2013-07-18 20:49:30] STEP 4: Download data from NCBI.
[2013-07-18 20:51:39] Retrieved GenBank -isolation source- field linked to GIs. Unique entries are saved in the
documents folder.
[2013-07-18 20:51:39] STEP 5: Concatenate GenBank -isolation source- field IDs to blast file.
[2013-07-18 20:51:39] deg_species_filtered_M_blast_F_PMID.out is successfully generated.
[2013-07-18 20:51:39] STEP 6: Run SEQenv_tagger on the documents folder and generate ENVO hits file.
[2013-07-18 20:51:39] deg_species_filtered_M_blast_F_ENVO.txt is successfully generated.
[2013-07-18 20:51:39] STEP 7: Generate word cloud for overall community profile.
[2013-07-18 20:51:44] deg_species_filtered_M_blast_F_ENVO_overall_labels.png is successfully generated.
```

```
[2013-07-18 20:51:44] STEP 8: Generate frequency tables for sequences.
[2013-07-18 20:51:44] deg_species_filtered_M_blast_F_ENVO_sequences.csv is successfully using single document
matching and aggr document counting.
[2013-07-18 20:51:45] divrowsum normalization applied to deg_species_filtered_M_blast_F_ENVO_sequences.csv
successfully!
[2013-07-18 20:51:45] STEP 9: Generate dot files for frequency tables of sequences.
[2013-07-18 20:51:45] Processing C80.csv
[2013-07-18 20:51:48] Processing C48.csv
[2013-07-18 20:51:50] Processing C49.csv
[2013-07-18 20:51:53] Processing C50.csv
[2013-07-18 20:51:55] Processing C51.csv
[2013-07-18 20:51:58] Processing C60.csv
[2013-07-18 20:52:00] Processing C61.csv
[2013-07-18 20:52:03] Processing C62.csv
[2013-07-18 20:52:06] Processing C63.csv
[2013-07-18 20:52:08] Processing C8.csv
[2013-07-18 20:52:11] Processing C9.csv
[2013-07-18 20:52:13] Processing C10.csv
[2013-07-18 20:52:16] Processing C11.csv
[2013-07-18 20:52:18] Processing C52.csv
[2013-07-18 20:52:21] Processing C53.csv
[2013-07-18 20:52:23] Processing C54.csv
[2013-07-18 20:52:26] Processing C55.csv
[2013-07-18 20:52:29] Processing C16.csv
[2013-07-18 20:52:31] Processing C17.csv
[2013-07-18 20:52:34] Processing C18.csv
[2013-07-18 20:52:36] Processing C19.csv
[2013-07-18 20:52:39] Processing C12.csv
[2013-07-18 20:52:41] Processing C13.csv
[2013-07-18 20:52:44] Processing C14.csv
[2013-07-18 20:52:47] Processing C15.csv
[2013-07-18 20:52:49] Processing C20.csv
[2013-07-18 20:52:52] Processing C21.csv
[2013-07-18 20:52:54] Processing C22.csv
[2013-07-18 20:52:57] Processing C23.csv
[2013-07-18 20:52:59] Processing C36.csv
[2013-07-18 20:53:02] Processing C37.csv
```

```
[2013-07-18 20:53:05] Processing C38.csv
[2013-07-18 20:53:07] Processing C39.csv
[2013-07-18 20:53:10] Processing C28.csv
[2013-07-18 20:53:12] Processing C29.csv
[2013-07-18 20:53:15] Processing C30.csv
[2013-07-18 20:53:18] Processing C31.csv
[2013-07-18 20:53:20] Processing C32.csv
[2013-07-18 20:53:23] Processing C33.csv
[2013-07-18 20:53:25] Processing C34.csv
[2013-07-18 20:53:28] Processing C35.csv
[2013-07-18 20:53:30] Processing C44.csv
[2013-07-18 20:53:33] Processing C45.csv
[2013-07-18 20:53:35] Processing C46.csv
[2013-07-18 20:53:38] Processing C47.csv
[2013-07-18 20:53:41] Processing C4.csv
[2013-07-18 20:53:43] Processing C5.csv
[2013-07-18 20:53:46] Processing C6.csv
[2013-07-18 20:53:48] Processing C7.csv
[2013-07-18 20:53:51] Processing C40.csv
[2013-07-18 20:53:54] Processing C41.csv
[2013-07-18 20:53:56] Processing C42.csv
[2013-07-18 20:53:59] Processing C43.csv
[2013-07-18 20:54:02] Processing C1.csv
[2013-07-18 20:54:04] Processing C2.csv
[2013-07-18 20:54:07] Processing C3.csv
[2013-07-18 20:54:09] Processing C56.csv
[2013-07-18 20:54:12] Processing C57.csv
[2013-07-18 20:54:15] Processing C59.csv
[2013-07-18 20:54:17] Processing C24.csv
[2013-07-18 20:54:20] Processing C25.csv
[2013-07-18 20:54:22] Processing C26.csv
[2013-07-18 20:54:25] Processing C27.csv
[2013-07-18 20:54:28] Processing C64.csv
[2013-07-18 20:54:30] Processing C65.csv
[2013-07-18 20:54:33] Processing C66.csv
[2013-07-18 20:54:35] Processing C67.csv
[2013-07-18 20:54:38] Processing C72.csv
```

```
[2013-07-18 20:54:41] Processing C73.csv
[2013-07-18 20:54:43] Processing C74.csv
[2013-07-18 20:54:46] Processing C75.csv
[2013-07-18 20:54:48] Processing C78.csv
[2013-07-18 20:54:51] Processing C69.csv
[2013-07-18 20:54:54] Processing C70.csv
[2013-07-18 20:54:56] Processing C71.csv
[2013-07-18 20:54:59] Folder sequences_dot is successfully generated.
[2013-07-18 20:54:59] STEP 10: Generate labels for frequency tables of sequences (all).
[2013-07-18 20:55:02] deg_species_filtered_M_blast_F_ENVO_sequences_labels.csv is successfully generated.
[2013-07-18 20:55:02] STEP 11: Generate word clouds for sequences (all).
[2013-07-18 20:55:32] Folder sequences_wc is successfully generated.
[2013-07-18 20:55:32] Generate heapmap for sequences (all).
[2013-07-18 20:55:40] deg_species_filtered_M_blast_F_ENVO_sequences_labels.png is successfully generated.
[2013-07-18 20:55:40] Folder samples_heapmaps is successfully generated.
[2013-07-18 20:55:40] Finished processing!
```

Here are the contents in the current folder:

```
[seqenv@quince-srv2 ~/uzi/test_SEQenv]$ ls
deg_species_filtered.fna                                deg_species_filtered_M_blast_F_ENVO.txt  documents
deg_species_filtered_M_blast_F_ENVO_overall.csv         deg_species_filtered_M_blast_F.out       SEQenv.log
deg_species_filtered_M_blast_F_ENVO_overall_labels.csv  deg_species_filtered_M_blast_F_PMID.out  sequences_dot
deg_species_filtered_M_blast_F_ENVO_overall_labels.png  deg_species_filtered_M_blast.out
sequences_heapmaps
deg_species_filtered_M_blast_F_ENVO_sequences.csv       deg_species_filtered_M.fa                sequences_wc
deg_species_filtered_M_blast_F_ENVO_sequences_labels.csv deg_species_filtered_M.map
[seqenv@quince-srv2 ~/uzi/test_SEQenv]$
```

Since the sequences in FASTA format can have long headers, the pipeline first produces a header map file and a FASTA file with modified headers and then processes the new FASTA file instead. You can check the contents of deg_species_filtered_M.map to see what each ID corresponds to.

```
[seqenv@quince-srv2 ~/uzi/test_SEQenv]$ head deg_species_filtered_M.map
```

C1 gi|219846460|ref|NR_026051.1| Caldicellulosiruptor owensensis OL strain OL 16S ribosomal RNA, complete sequence >gi|2454185|gb|U80596.1|COU80596 Caldicellulosiruptor owensense 16S ribosomal RNA gene, partial sequence
C2 gi|265678524|ref|NR_028828.1| Xylanimonas cellulosilytica DSM 15894 strain XIL07 16S ribosomal RNA, complete sequence >gi|22086567|gb|AF403541.1| Xylanomonas cellulosilytica 16S ribosomal RNA gene, partial sequence
C3 gi|343200548|ref|NR_041235.1| Clostridium clariflavum DSM 19732 strain EBR45 16S ribosomal RNA, complete sequence >gi|51036225|dbj|AB186359.1| Clostridium clariflavum gene for 16S rRNA
C4 gi|343200548|ref|NR_041235.1| Clostridium clariflavum DSM 19732 strain EBR45 16S ribosomal RNA, complete sequence >gi|51036225|dbj|AB186359.1| Clostridium clariflavum gene for 16S rRNA
C5 gi|68989453|gb|DQ089673.1| Enterobacter cloacae strain CP1 16S ribosomal RNA gene, complete sequence
C6 gi|343201115|ref|NR_041822.1| Actinosynnema mirum DSM 43827 strain IMSNU 20048T (IFO 14064T) 16S ribosomal RNA, complete sequence >gi|33340587|gb|AF328679.1| Actinosynnema mirum 16S ribosomal RNA gene, partial sequence
C7 gi|343201115|ref|NR_041822.1| Actinosynnema mirum DSM 43827 strain IMSNU 20048T (IFO 14064T) 16S ribosomal RNA, complete sequence >gi|33340587|gb|AF328679.1| Actinosynnema mirum 16S ribosomal RNA gene, partial sequence
C8 gi|41387515|gb|AY445592.1| Ruminococcus albus strain B199 16S ribosomal RNA gene, complete sequence
C9 gi|41387516|gb|AY445593.1| Ruminococcus flavefaciens strain C94 16S ribosomal RNA gene, complete sequence
C10 gi|41387517|gb|AY445594.1| Ruminococcus albus strain 8 16S ribosomal RNA gene, complete sequence