1 Project report

1.1 General Information

This is the project named "`demo_project`". It contains 5 samples. It is described as "Demo project for the pacmill pipeline".

1.2 Processing

- This report and all the analysis was generated using the `pacmill` python pipeline.
- Documentation and source code is available at:

  https://github.com/xapple/pacmill

- Version `0.3.3` of the pipeline was used.
- This document was generated at `2020-08-26 04:06:58 CEST+0200` on host `bigmille`.

1.3 Samples

Some summary information concerning the samples is given in table 1 below.

Table 1. Summary information for all samples.

#	Name	Description	Reads lost	Reads left
1	**mock**	Mock community with 38 species	75.5%	6'038
2	**p__19**	Sediment sample obtained from a hot spring	72.4%	11'559
3	**pm__3**	Sediment sample taken from below the sea floor	74.8%	8'024
4	**sala**	Black biofilm that was taken in an old silver mine	79.2%	5'359
5	**tns__08**	Sediment sample taken from a submarine hydrothermal vent field	72.6%	7'459

1.4 Input data

Summing the reads from all the samples, we have 34'571 sequences to work on. Before starting the analysis we can look at the length distribution pattern that these reads form in figure 1.

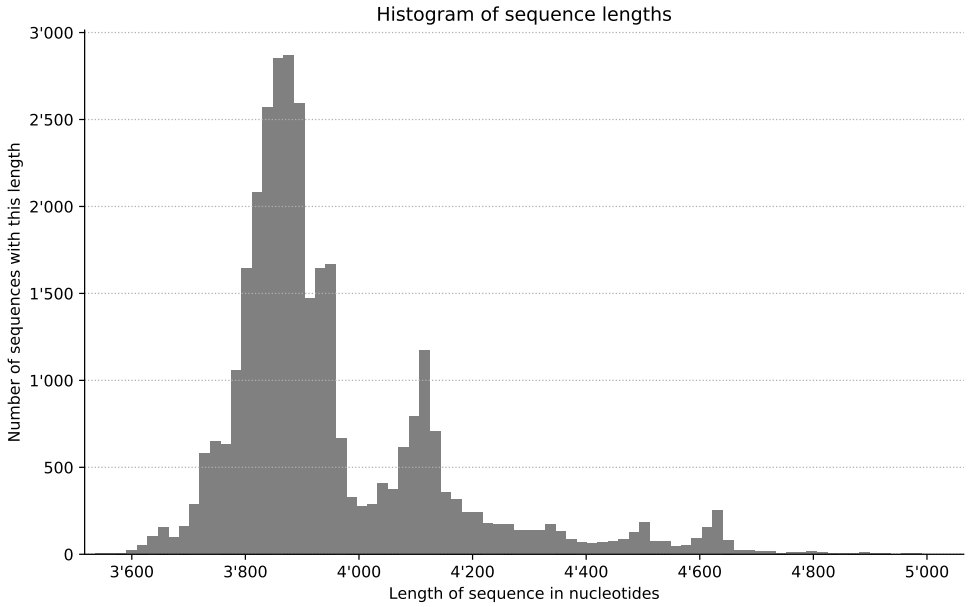

Figure 1. Distribution of sequence lengths at input

1.5 Clustering into OTUs

Two sequences that diverge by no more than a few nucleotides are probably not produced by ecological diversity. They are most likely produced by errors along the laboratory method and the sequencing. Therefore, we place them together in one unit, called an "Operational Taxonomic Unit".

For this clustering, we use the `vsearch --cluster_size` algorithm.

The similarity threshold chosen is 97.0%. Exactly 6'164 OTUs are produced.

1.6 OTU table

A table with OTUs as rows and samples as columns is produced. Each cell within this table tells us how many sequences are participating in the given OTU from the given sample. This table is often too big to be viewed directly here. However, we can plot some of its properties to better gauge the sparsity as seen in figures 2, 3 and 4:

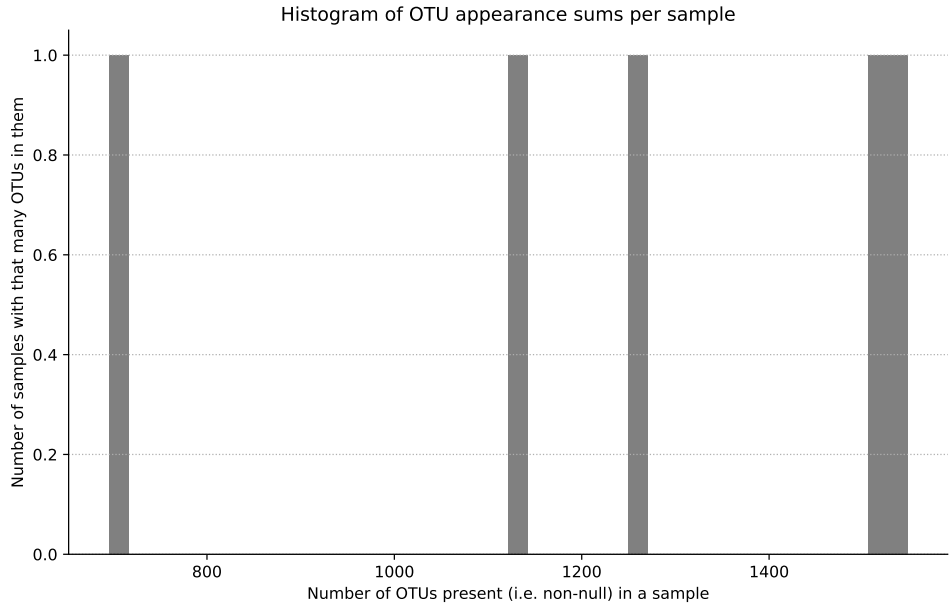

Figure 2. Distribution of OTU presence per OTU

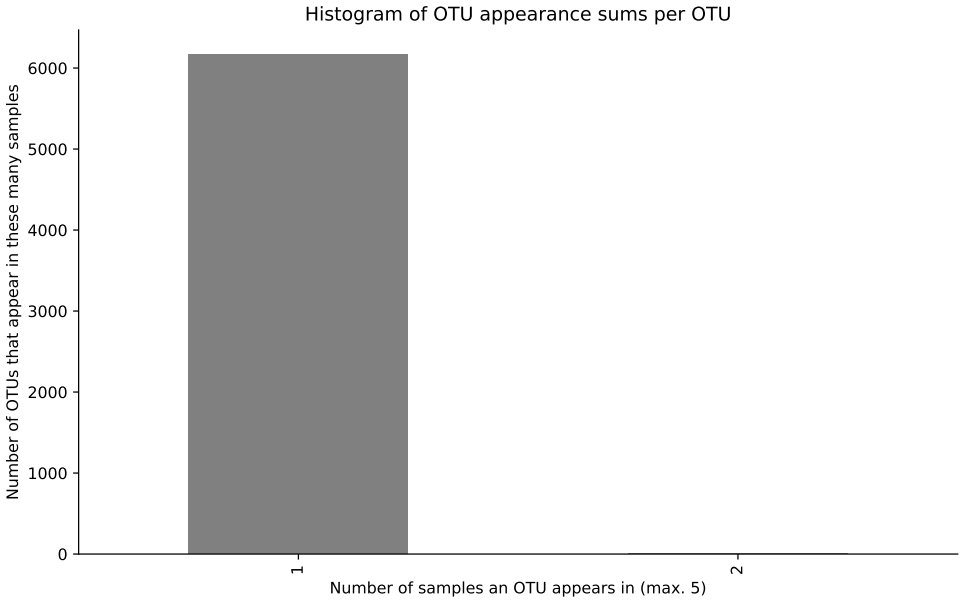

Figure 3. Distribution of OTU presence per sample

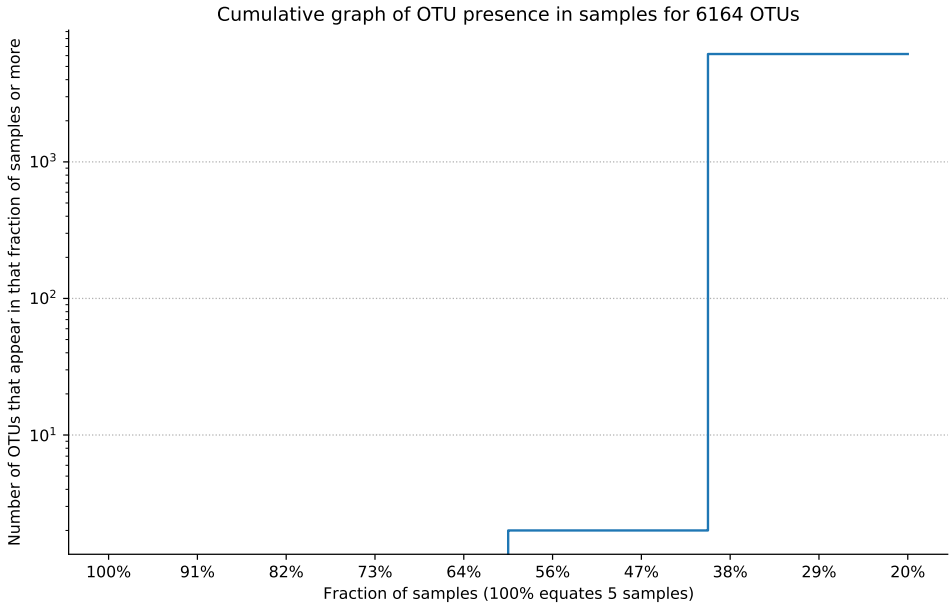

Figure 4. Cumulative number of reads by OTU presence

1.7 Comparison

We now would like to start comparing samples amongst each other to determine which ones are similar or if any clear groups can be observed. A first means of doing that is by using the information in the OTU table and a distance metric such as the "Horn 1966 (adapted from Morisita 1959)" one to place them on an ordination plot. This can be seen in figure 5.

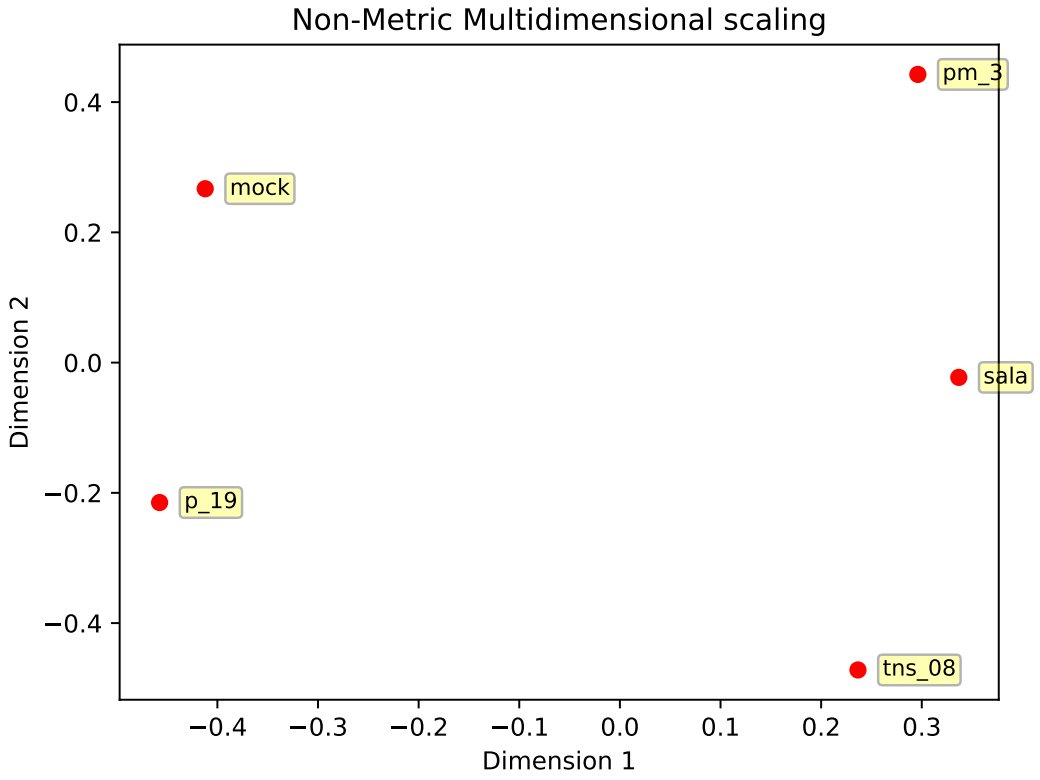

Figure 5. NMDS using the OTU table for 5 samples

These kind of graphs have a random component to them and can be easily influenced by one or two differently looking samples.

1.8 Distances

To compute beta diversity, other distance measures are possible of course. Bray-Curtis and Jaccard distance matrices can be created. We can also explore phylogenetic distance measures such as the UniFrac one. This is also possible and a UniFrac distance matrix can easily be computed. One can also build a hierarchical clustering of the samples from it (not included).