1 Cluster "ice"

1.1 General Information

This is the cluster named "ice". It contains 40 samples. It corresponds to project code 'ice' ('Jämtland lakes under ice')

1.2 Samples

There are 40 samples in this cluster. Some summary information about them is given in table 1.

#	Name	Reference	Description	Reads lost	Reads left
1	**rl1am**	`run10_sample26`	Ice RL1Am	37.1%	76'540
2	**rl2bm**	`run10_sample27`	Ice RL2Bm	35.9%	208'097
3	**rl3bm**	`run10_sample28`	Ice RL3Bm	37.7%	84'615
4	**rl4am**	`run10_sample29`	Ice RL4Am	36.2%	123'225
5	**rl5bm**	`run10_sample30`	Ice RL5Bm	36.9%	104'348
6	**rl6bm**	`run10_sample31`	Ice RL6Bm	38.1%	39'141
7	**rl7bm**	`run10_sample32`	Ice RL7Bm	37.2%	91'569
8	**rl8bm**	`run10_sample33`	Ice RL8Bm	37.5%	83'646
9	**bt1am**	`run10_sample34`	Ice BT1Am	36.1%	51'998
10	**bt2am**	`run10_sample35`	Ice BT2Am	35.0%	109'352
11	**bt3bm**	`run10_sample36`	Ice BT3Bm	35.9%	42'218
12	**bt4am**	`run10_sample37`	Ice BT4Am	36.0%	76'414
13	**bt5am**	`run10_sample38`	Ice BT5Am	36.7%	56'565
14	**bt6am**	`run10_sample39`	Ice BT6Am	35.4%	146'392
15	**bt7bm**	`run10_sample40`	Ice BT7Bm	38.0%	99'327
16	**bt8am**	`run10_sample41`	Ice BT8Am	37.5%	102'791
17	**lb1bm**	`run10_sample42`	Ice LB1Bm	35.7%	103'512
18	**lb2am**	`run10_sample43`	Ice LB2Am	36.7%	84'476
19	**lb3am**	`run10_sample44`	Ice LB3Am	35.9%	67'326
20	**lb4am**	`run10_sample45`	Ice LB4Am	36.4%	107'560
21	**lb5am**	`run10_sample46`	Ice LB5Am	37.3%	72'760
22	**lb6am**	`run10_sample47`	Ice LB6Am	35.7%	128'608
23	**lb7am**	`run10_sample48`	Ice LB7Am	36.3%	122'885
24	**lb8am**	`run10_sample49`	Ice LB8Am	36.5%	95'190
25	**kt1bm**	`run10_sample50`	Ice KT1Bm	37.2%	113'568
26	**kt2bm**	`run10_sample51`	Ice KT2Bm	35.6%	139'849
27	**kt3am**	`run10_sample52`	Ice KT3Am	38.2%	106'258
28	**kt4am**	`run10_sample53`	Ice KT4Am	36.1%	106'835

#	Name	Reference	Description	Reads lost	Reads left
29	**kt5bm**	`run10_sample54`	Ice KT5Bm	37.1%	83'473
30	**kt6bm**	`run10_sample55`	Ice KT6Bm	39.2%	123'726
31	**kt7bm**	`run10_sample56`	Ice KT7Bm	38.1%	89'134
32	**kt8bm**	`run10_sample57`	Ice KT8Bm	38.1%	80'029
33	**gs1am**	`run10_sample58`	Ice GS1Am	37.5%	104'916
34	**sb1bm**	`run10_sample59`	Ice SB1Bm	36.0%	168'409
35	**sb2am**	`run10_sample60`	Ice SB2Am	36.5%	168'353
36	**sb3bm**	`run10_sample61`	Ice SB3Bm	36.9%	140'008
37	**sb4bm**	`run10_sample62`	Ice SB4Bm	37.8%	101'416
38	**sb5am**	`run10_sample63`	Ice SB5Am	36.1%	114'252
39	**sb6am**	`run10_sample64`	Ice SB6Am	38.6%	141'135
40	**sb7am**	`run10_sample65`	Ice SB7Am	36.8%	131'408

Table 1. Summary information for all samples.

1.3 Processing

- This report (and all the analysis) was generated using the ILLUMITAG project at: http://xapple.github.io/illumitag/
- A more detailed peer reviewed article has been puslished in PLoS ONE describing this method.
- Version `1.0.2` of the pipeline was used.
- This document was generated at `2016-05-11 12:34:26 CEST+0200`.
- The results and all the files generated for this sample can be found at:
  ssh://ww-hmem02.climb.cluster/home/lucas/ILLUMITAG/views/projects/ice/cluster/

1.4 Input data

Summing the reads from all the samples, we have 4'191'324 sequences to work on. Sequence quality information is disregarded from this point on. Before starting the analysis we can look at the length distribution pattern that these reads form in figure 1.

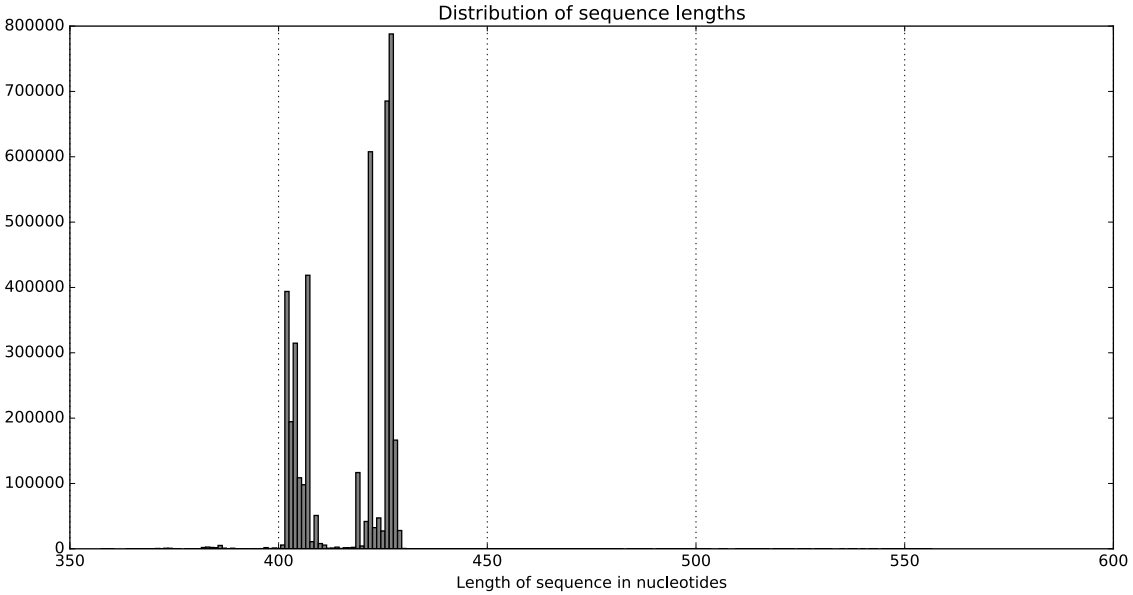

Figure 1. Distribution of sequence lengths at input

1.5 Clustering

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. So we put them together in one unit, called an OTU. On the other hand, a sequence that does not have any such similar-looking brothers is most likely the product of a recombination (chimera) and is discarded. This process is done using the UPARSE denovo picking method (v7.0.1090_i86linux32). The publication is available at:

http://www.nature.com/doifinder/10.1038/nmeth.2604

The similarity threshold chosen is 3.0%. Exactly 9'648 OTUs are produced.

1.6 Classification

Relying on databases of ribosomal genes such as Silva, we can classify each OTU and give it an approximative affiliation. This provides a taxonomic name to each OTU. This is done using the LCAClassifier method (version 2.0 - March 2014). The publication is available at:

http://dx.plos.org/10.1371/journal.pone.0049334

Out of our 9'648 OTUs, exactly 9'620 of them are assigned to a position somewhere in the tree of life (not necessary on a tip though).

At this point we are going to remove some OTUs. All those pertaining to any of the following phyla are discarded: Plastid, Mitochondrion, Thaumarchaeota, Crenarchaeota and Euryarchaeota. This leaves us with 9'136 'good' OTUs. As OTUs contain a varying number of sequences in them, we can plot this distribution in figure 2.

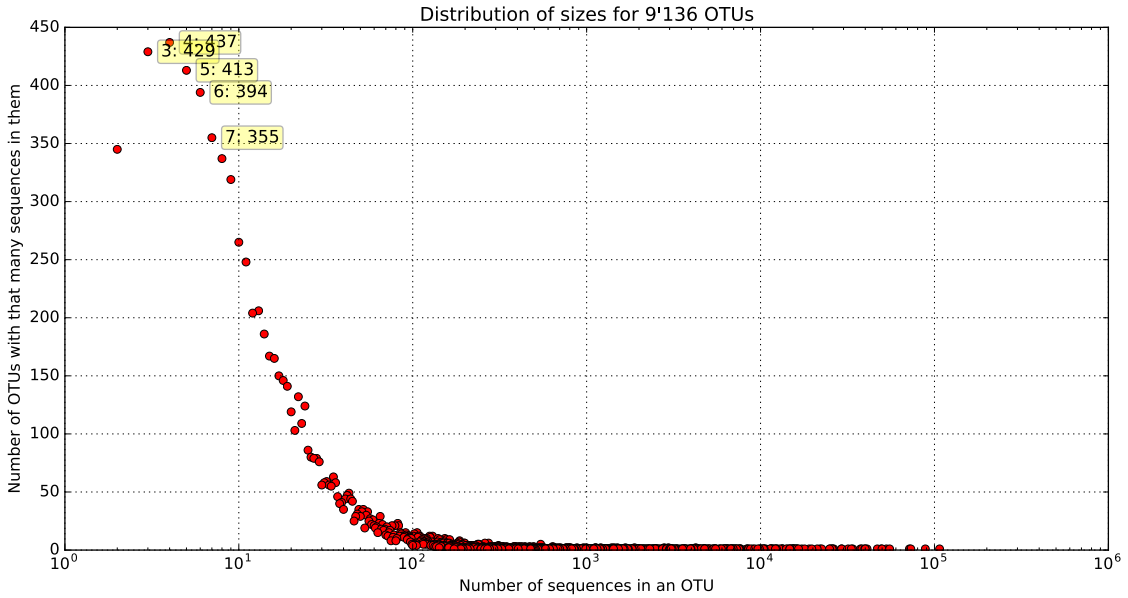

Figure 2. Distribution of OTU sizes

1.7 OTU table

Now we check which sample each sequence of each OTU was coming from and make a count table with OTUs as rows (9'136) and samples as columns (40). Each cell tells us how many sequences are pertaining to this OTU from this sample. This table is too big to be viewed directly here. However we can plot some of its properties to better understand how sparse it is as seen in figures 3, 4 and 5:

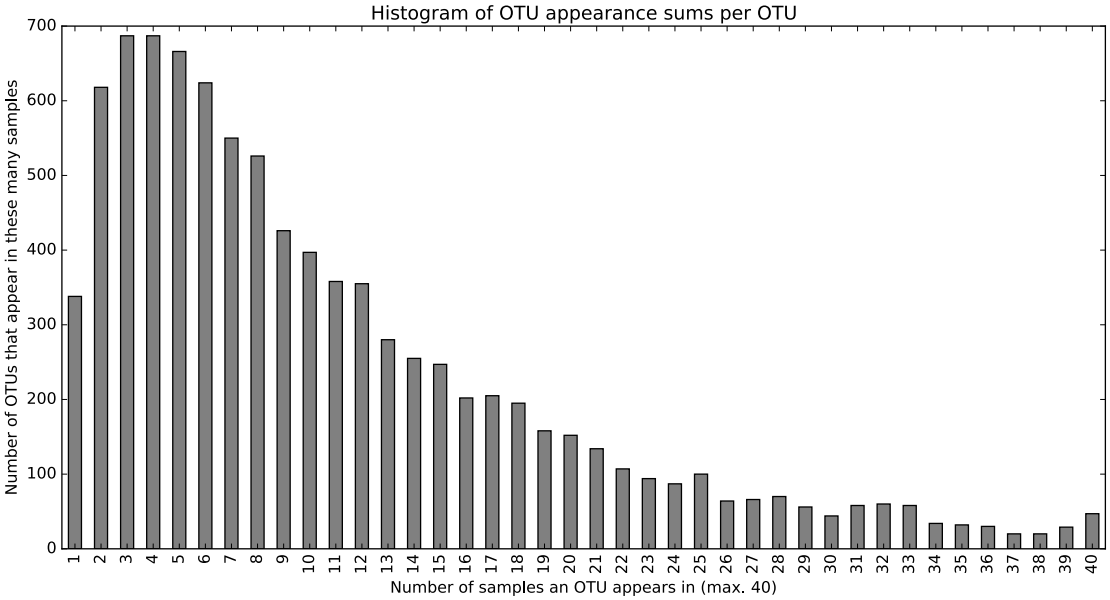

Figure 3. Distribution of OTU presence per OTU

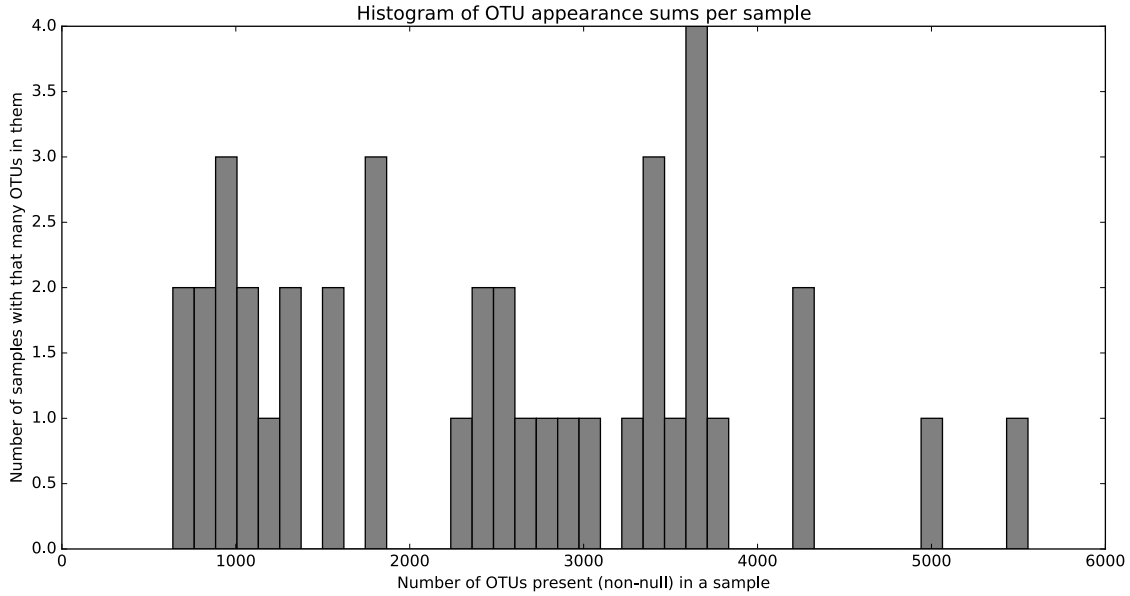

Figure 4. Distribution of OTU presence per sample

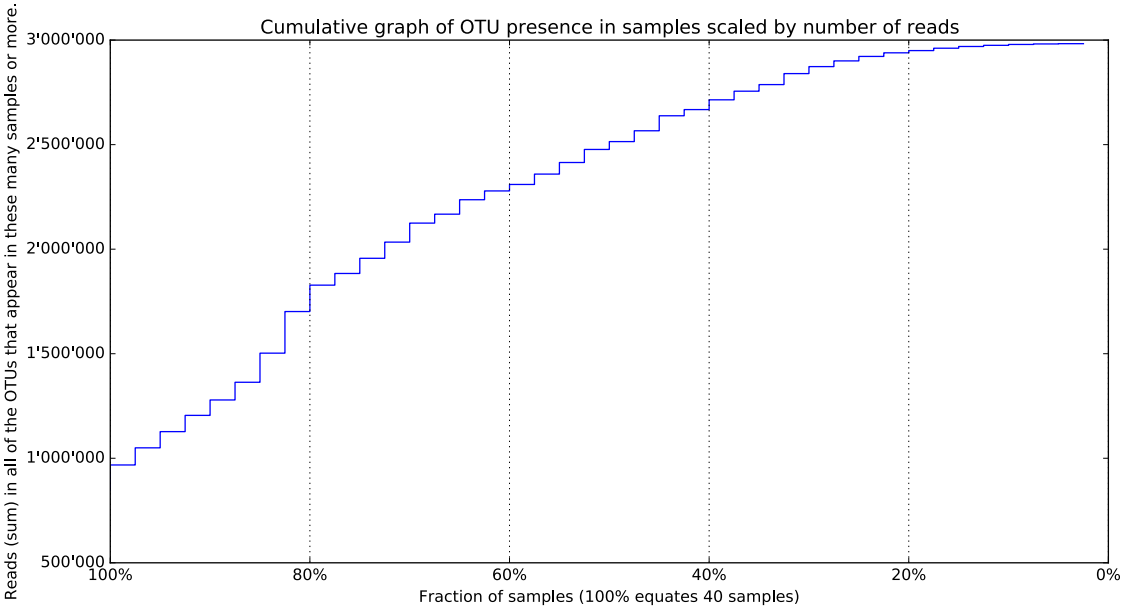

Figure 5. Cumulative number of reads by OTU presence

1.8 Taxa table

If we modify the rows of our table to become taxonomic names instead of OTUs, some rows will have the same affiliations and will be merged together by summation. This produces the taxa table which has 40 samples and 795 named taxa. It's important to consider the difference between an OTU table and a taxa table.

1.9 Composition

At this point, one of the most obvious graphs to produce is a bar-chart detailing the composition in terms of taxonomy of every one of our samples. To keep things simple we will only consider the 'phyla' taxonomic level and only divide phyla into their composing classes when they contain a very large proportion of reads (going deeper while still including everything would yield an unreadable graph). This can be seen in figure 6.

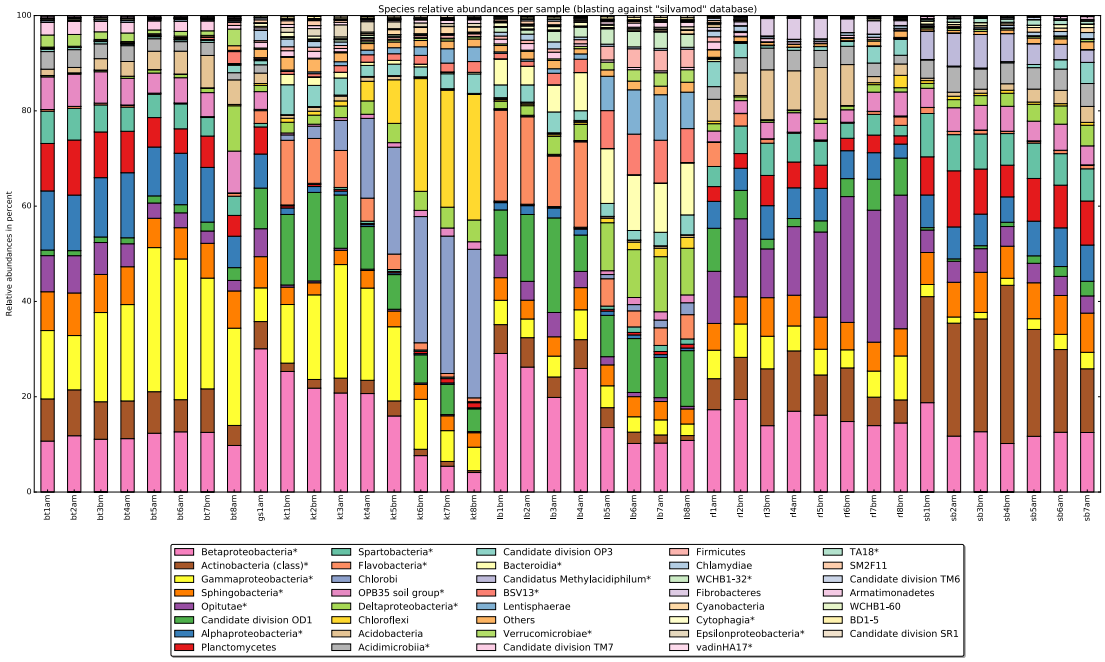

Figure 6. Species relative abundances per sample on the phyla and class levels

---

1.10 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" one to place them on an ordination plot. This can be seen in figure 7.

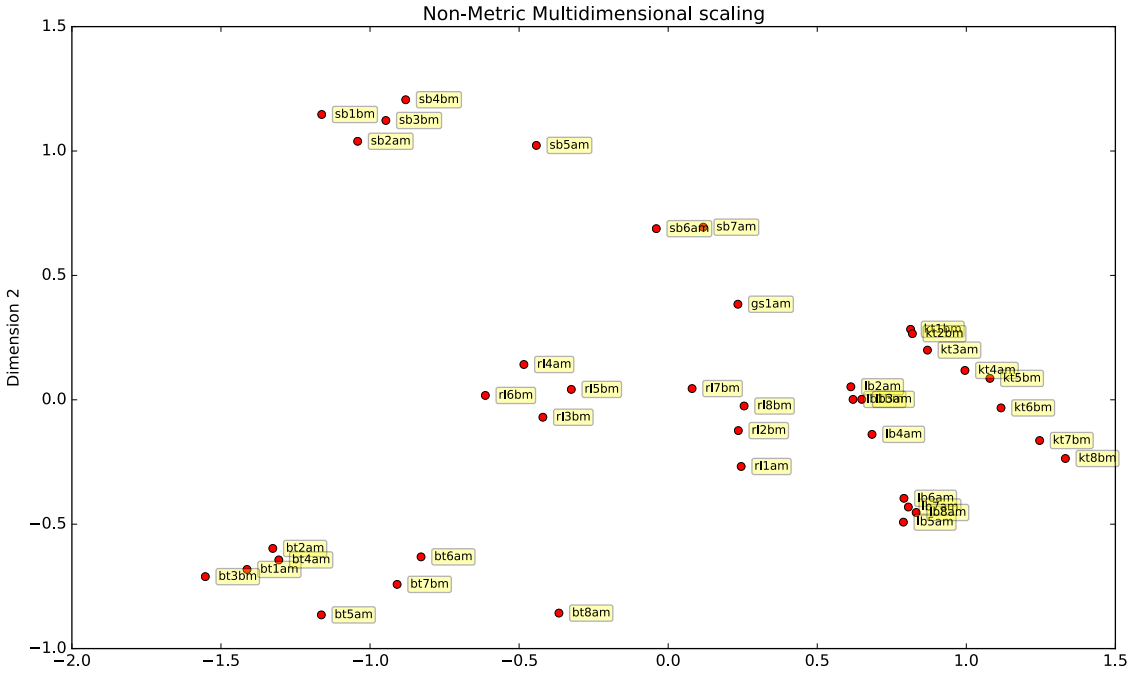

Figure 7. NMDS using the OTU table for 40 samples

These kind of graphs have a random component to them and can be easily influenced by one or two differently looking samples. If one uses the taxa table instead, already one gets a different result as seen in figure 8.

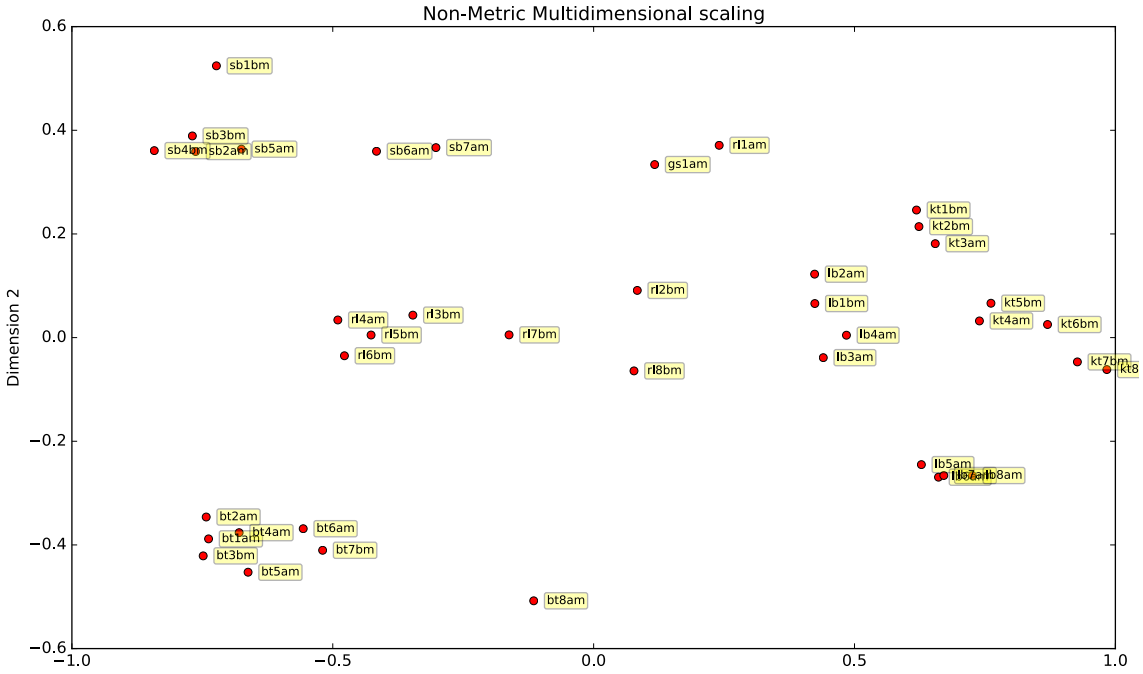

Figure 8. NMDS using the taxa table for 40 samples

One can also make NMDS plots with more complicated distance measures such as phylogenetic ones. More about that later.

1.11 Distances

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

1.12 Environmental tags

Relying on the same kind of databases and their meta-data, we can try to infer a typical environmental tag to each sequence. This, in turn, enables us to assign a linear combination of environmental tags to each sample and to the cluster as a whole. This method is also implemented in the pipeline (results on demand):

http://environments.hcmr.gr/seqenv.html