Cluster "ice"

General Information

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

Samples

There are in 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`	rl1am	37.1%	76'512
2	**rl2bm**	`run10_sample27`	rl2bm	35.9%	208'026
3	**rl3bm**	`run10_sample28`	rl3bm	37.8%	84'579
4	**rl4am**	`run10_sample29`	rl4am	36.2%	123'193
5	**rl5bm**	`run10_sample30`	rl5bm	36.9%	104'310
6	**rl6bm**	`run10_sample31`	rl6bm	38.1%	39'131
7	**rl7bm**	`run10_sample32`	rl7bm	37.3%	91'538
8	**rl8bm**	`run10_sample33`	rl8bm	37.6%	83'617
9	**bt1am**	`run10_sample34`	bt1am	36.1%	51'981
10	**bt2am**	`run10_sample35`	bt2am	35.0%	109'315
11	**bt3bm**	`run10_sample36`	bt3bm	35.9%	42'208
12	**bt4am**	`run10_sample37`	bt4am	36.1%	76'388
13	**bt5am**	`run10_sample38`	bt5am	36.7%	56'547
14	**bt6am**	`run10_sample39`	bt6am	35.4%	146'343
15	**bt7bm**	`run10_sample40`	bt7bm	38.0%	99'286
16	**bt8am**	`run10_sample41`	bt8am	37.5%	102'750
17	**lb1bm**	`run10_sample42`	lb1bm	35.7%	103'479
18	**lb2am**	`run10_sample43`	lb2am	36.7%	84'447
19	**lb3am**	`run10_sample44`	lb3am	35.9%	67'306
20	**lb4am**	`run10_sample45`	lb4am	36.4%	107'528
21	**lb5am**	`run10_sample46`	lb5am	37.3%	72'728
22	**lb6am**	`run10_sample47`	lb6am	35.8%	128'558
23	**lb7am**	`run10_sample48`	lb7am	36.3%	122'848
24	**lb8am**	`run10_sample49`	lb8am	36.5%	95'158
25	**kt1bm**	`run10_sample50`	kt1bm	37.2%	113'531
26	**kt2bm**	`run10_sample51`	kt2bm	35.7%	139'802
27	**kt3am**	`run10_sample52`	kt3am	38.2%	106'221
28	**kt4am**	`run10_sample53`	kt4am	36.1%	106'803

	Name	Reference	Description	Reads lost	Reads left
29	**kt5bm**	`run10_sample54`	kt5bm	37.1%	83'445
30	**kt6bm**	`run10_sample55`	kt6bm	39.2%	123'679
31	**kt7bm**	`run10_sample56`	kt7bm	38.1%	89'106
32	**kt8bm**	`run10_sample57`	kt8bm	38.1%	80'007
33	**gs1am**	`run10_sample58`	gs1am	37.5%	104'883
34	**sb1bm**	`run10_sample59`	sb1bm	36.0%	168'364
35	**sb2am**	`run10_sample60`	sb2am	36.5%	168'297
36	**sb3bm**	`run10_sample61`	sb3bm	36.9%	139'961
37	**sb4bm**	`run10_sample62`	sb4bm	37.8%	101'383
38	**sb5am**	`run10_sample63`	sb5am	36.2%	114'212
39	**sb6am**	`run10_sample64`	sb6am	38.6%	141'092
40	**sb7am**	`run10_sample65`	sb7am	36.9%	131'382

Table 1. Summary information for all samples.

Processing

This report (and all the analysis) was generated using the ILLUMITAG project at:

http://github.com/limno/illumitag

Version `1.0.0` of the pipeline was used. The exact git hash of the latest commit was:

`e902cd63af4b634a255bb90c228c54ace07017d6`

also refereed to by its tag `submission2-40-ge902cd6-dirty`. This document was generated at `2014-07-31 20:50:02 CEST+0200`.

A brief overview of what happens to the data can be viewed online here:

https://github.com/limno/illumitag/blob/master/documentation/pipeline_outline.pdf?raw=true

The results and all the files generated for this cluster can be found on UPPMAX at:

`/home/lucass/ILLUMITAG/views/projects/ice/cluster/`

Input data

Summing the reads from all the samples, we have 4'189'944 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.

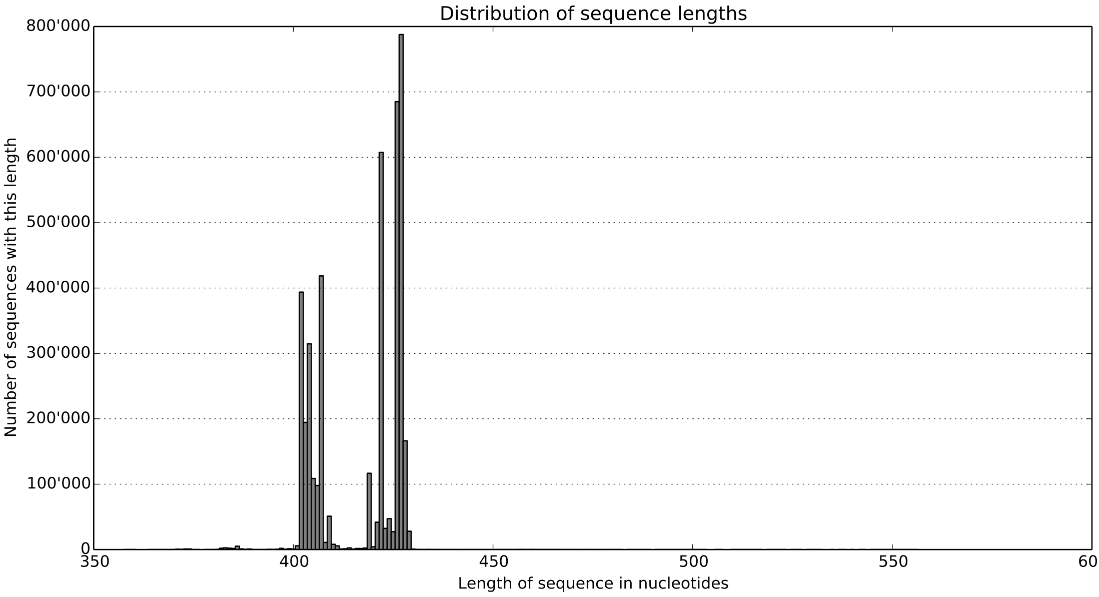

Figure 1. Distribution of sequence lengths at input

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 9'653 OTUs are produced.

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 9'653 OTUs, exactly 9'497 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 9'144 'good' OTUs. As OTUs contain a varying number of sequences in them, we can plot this distribution in figure 2.

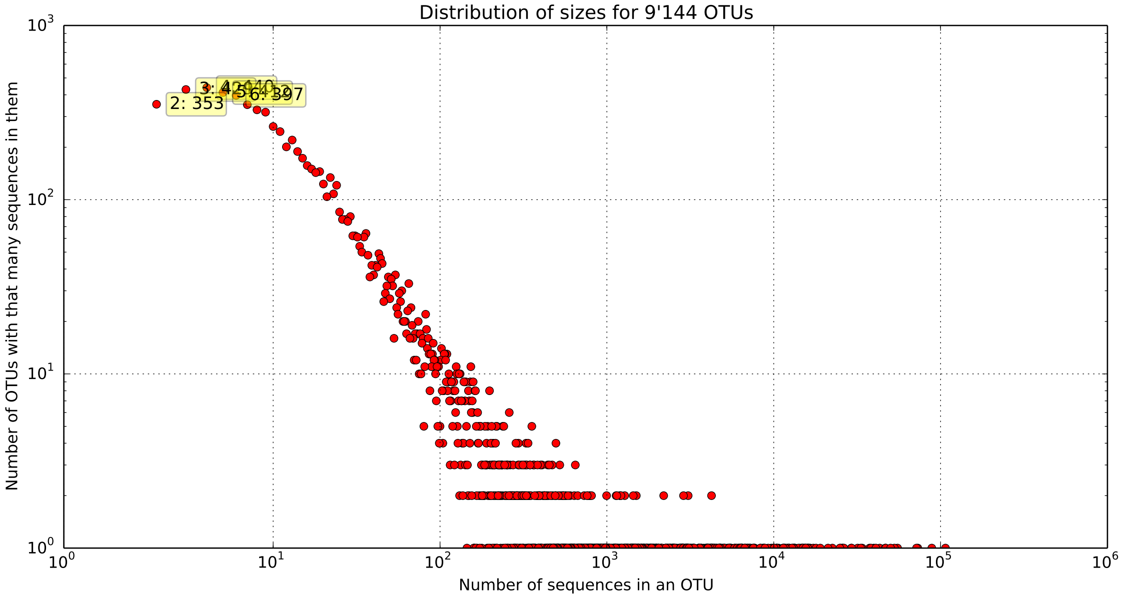

Figure 2. Distribution of OTU sizes

OTU table

Now we check which sample each sequence of each OTU was coming from and make a count table with OTUs as rows (9'144) 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:

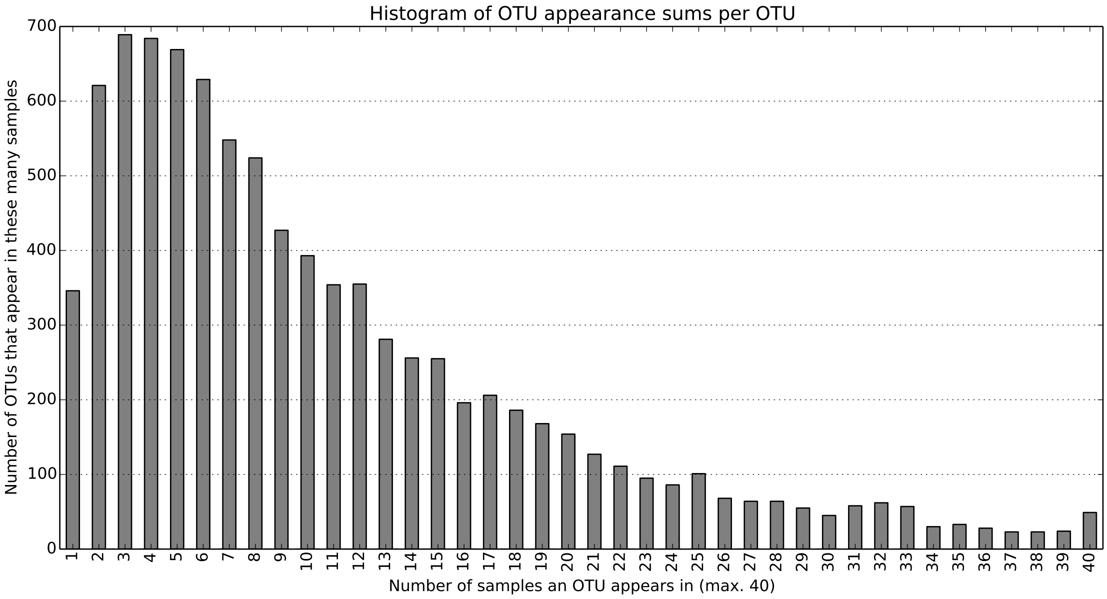

Figure 3. Distribution of OTU presence per OTU

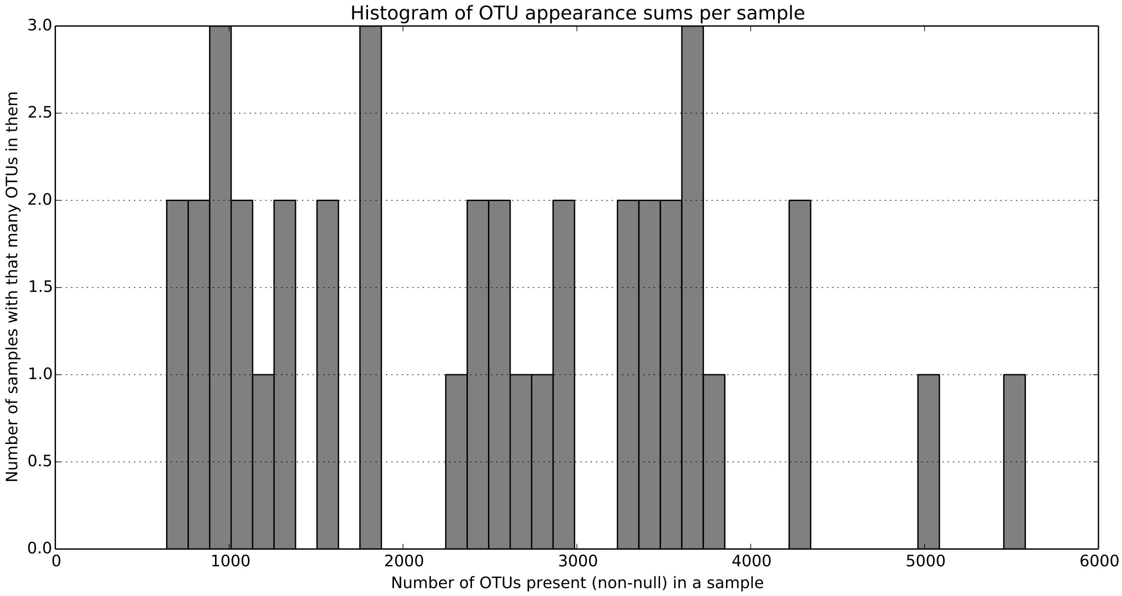

Figure 4. Distribution of OTU presence per sample

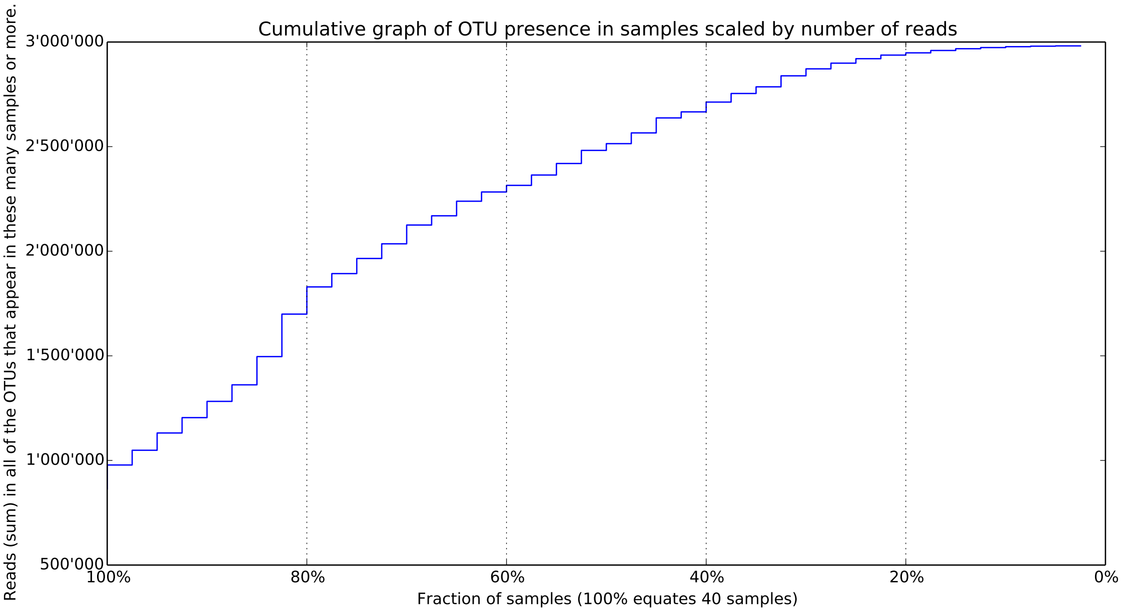

Figure 5. Cumulative number of reads by OTU presence

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 819 named taxa. It's important to consider the difference between an OTU table and a taxa table.

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 sometimes dividing phyla into their composing classes if they are very large (going deeper while still including everything would yield an unreadable graph). This can be seen in figure 6.

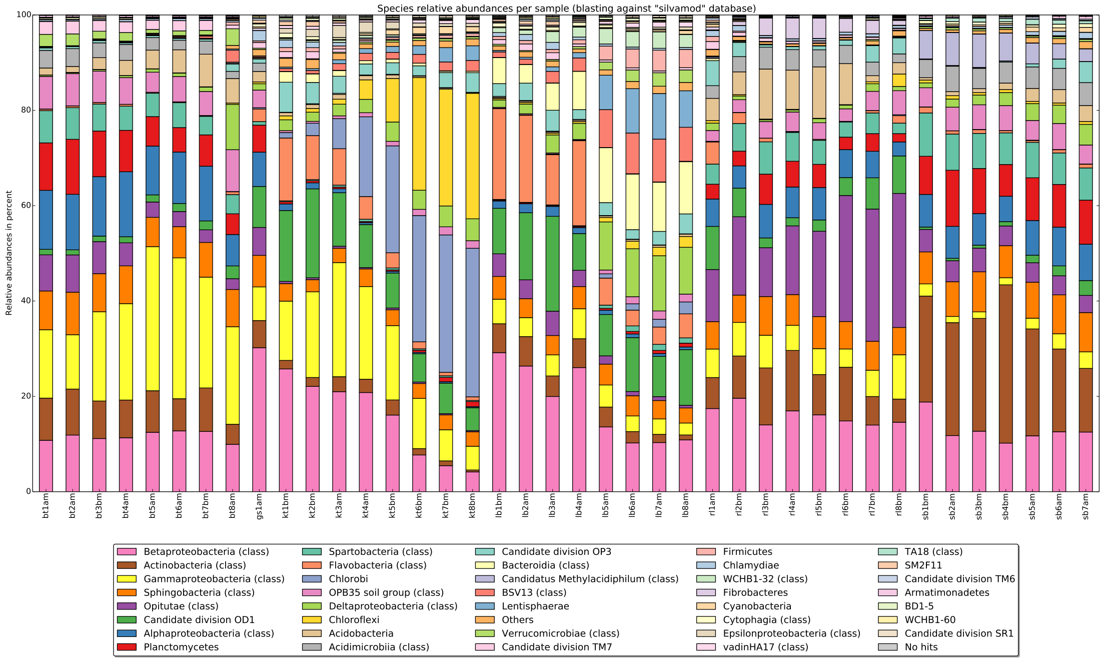

Figure 6. Species relative abundances per sample on the phyla and class levels

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.

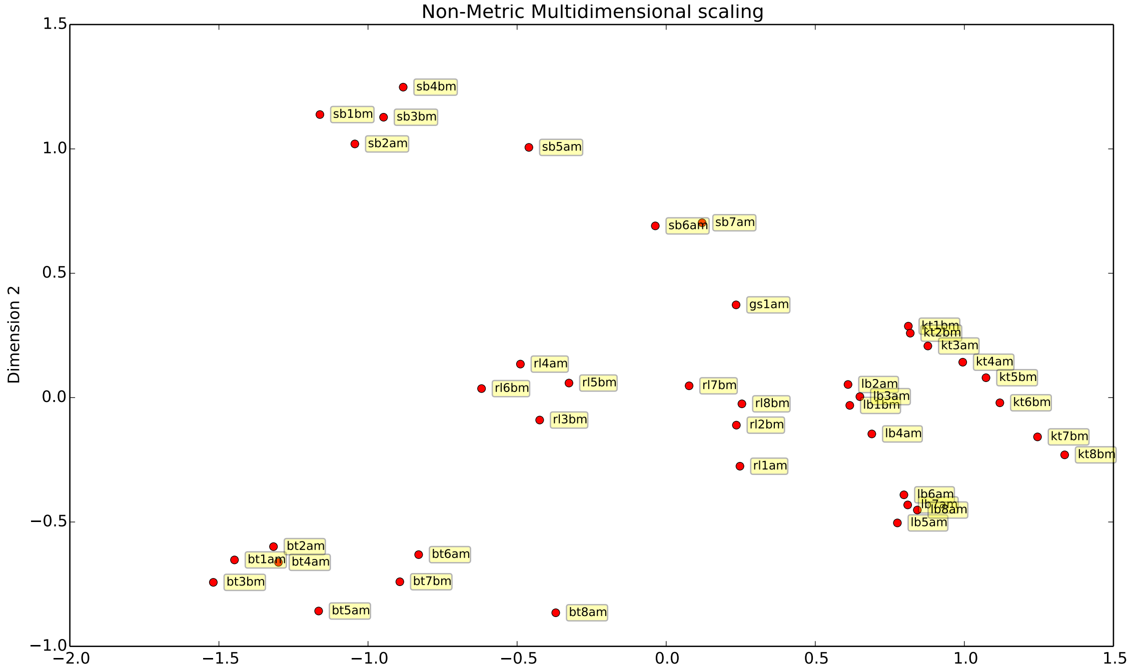

Figure 7. NMDS using the OTU table for 40 samples

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.

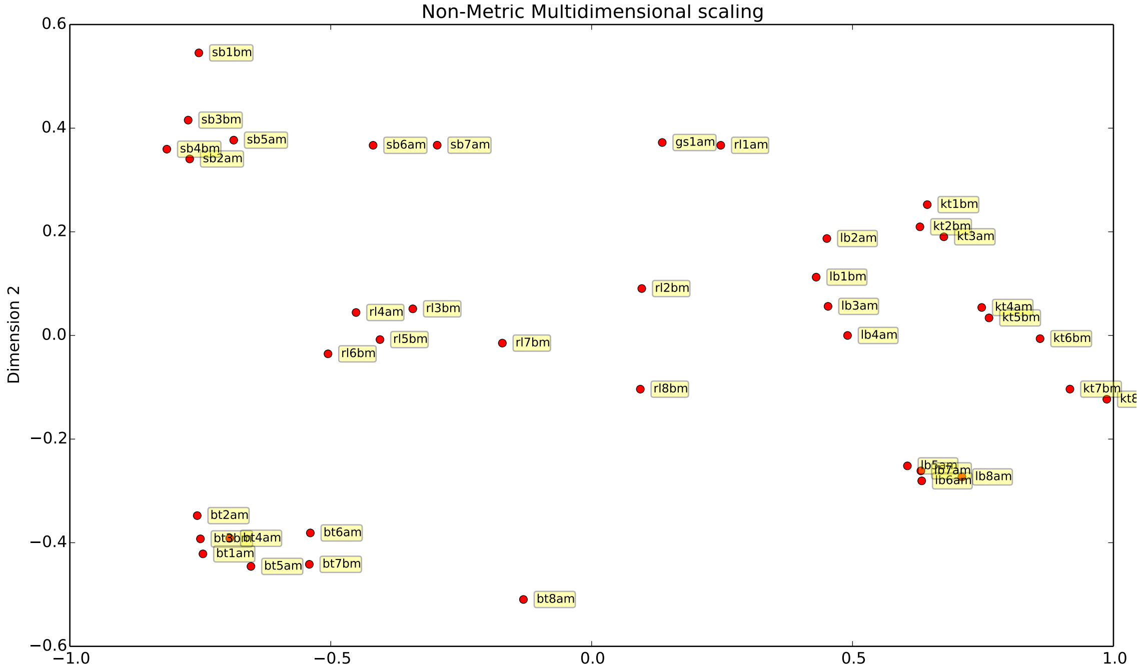

Figure 8. NMDS using the taxa table for 40 samples

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

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).

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