1 Cluster “under_ice”

1.1 General Information

This is the cluster named “under_ice”. It contains 39 samples. It corresponds to project code ‘under_ice’ (‘Jämtland lakes under ice’)

1.2 Samples

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

#	Name	Description	Reads lost	Reads left
1	**rl1**	Ice RL1Am	50.8%	59’838
2	**rl2**	Ice RL2Bm	48.7%	166’676
3	**rl3**	Ice RL3Bm	52.2%	64’920
4	**rl4**	Ice RL4Am	49.3%	97’820
5	**rl5**	Ice RL5Bm	50.3%	82’089
6	**rl6**	Ice RL6Bm	52.2%	30’200
7	**rl7**	Ice RL7Bm	50.7%	72’002
8	**rl8**	Ice RL8Bm	50.7%	66’017
9	**bt1**	Ice BT1Am	49.9%	40’808
10	**bt2**	Ice BT2Am	47.8%	87’755
11	**bt3**	Ice BT3Bm	48.8%	33’725
12	**bt4**	Ice BT4Am	49.8%	59’956
13	**bt5**	Ice BT5Am	50.4%	44’323
14	**bt6**	Ice BT6Am	48.4%	116’957
15	**bt7**	Ice BT7Bm	52.8%	75’679
16	**bt8**	Ice BT8Am	50.8%	80’880
17	**lb1**	Ice LB1Bm	49.3%	81’628
18	**lb2**	Ice LB2Am	51.0%	65’441
19	**lb3**	Ice LB3Am	49.7%	52’826
20	**lb4**	Ice LB4Am	50.0%	84’634
21	**lb5**	Ice LB5Am	51.1%	56’779
22	**lb6**	Ice LB6Am	49.3%	101’548
23	**lb7**	Ice LB7Am	49.9%	96’545
24	**lb8**	Ice LB8Am	50.7%	73’916
25	**kt1**	Ice KT1Bm	51.5%	87’763
26	**kt2**	Ice KT2Bm	49.4%	109’910
27	**kt3**	Ice KT3Am	53.0%	80’770
28	**kt4**	Ice KT4Am	49.9%	83’729

#	Name	Description	Reads lost	Reads left
29	**kt5**	Ice KT5Bm	51.1%	64'818
30	**kt6**	Ice KT6Bm	54.2%	93'202
31	**kt7**	Ice KT7Bm	53.0%	67'672
32	**kt8**	Ice KT8Bm	53.5%	60'165
33	**sb1**	Ice SB1Bm	50.1%	131'108
34	**sb2**	Ice SB2Am	50.9%	130'299
35	**sb3**	Ice SB3Bm	51.3%	107'981
36	**sb4**	Ice SB4Bm	52.5%	77'429
37	**sb5**	Ice SB5Am	50.1%	89'302
38	**sb6**	Ice SB6Am	53.6%	106'618
39	**sb7**	Ice SB7Am	50.7%	102'588

Table 1. Summary information for all samples.

1.3 Processing

- This report (and all the analysis) was generated using the SIFES project at: http://xapple.github.io/sifes/
- A more detailed peer reviewed article has been published in PLoS ONE describing parts of this method.
- Version `2.0.1` of the pipeline was used.
- This document was generated at `2016-11-11 23:20:30 CET+0100`.

1.4 Input data

Summing the reads from all the samples, we have 3'186'316 sequences to work on. Before starting the analysis we can look at the length distribution pattern that these reads form in figure 1.

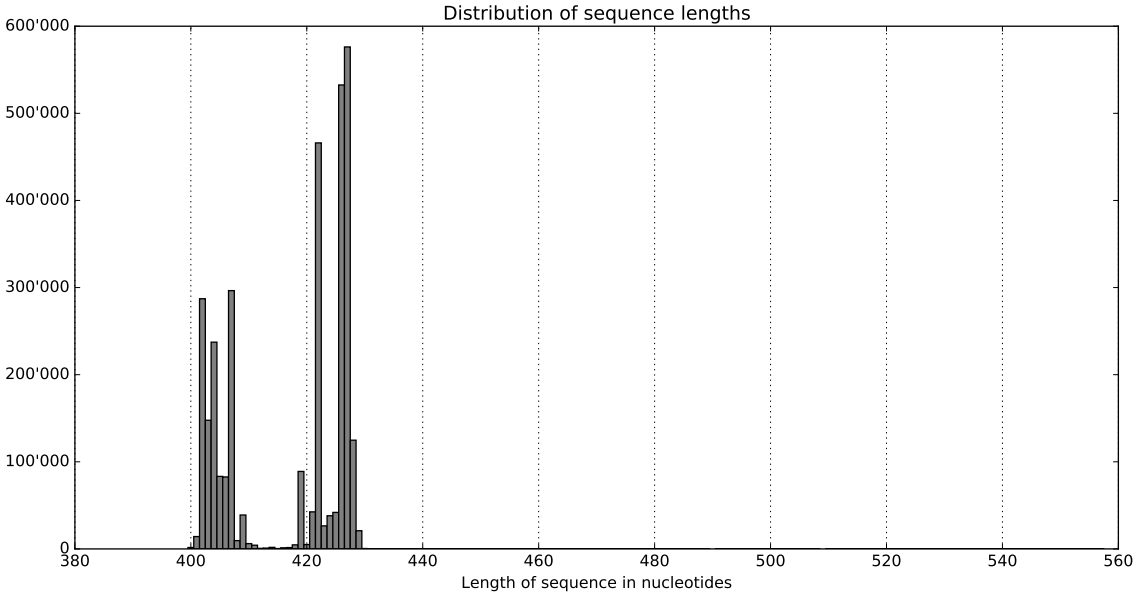

Figure 1. Distribution of sequence lengths at input

1.5 Clustering

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. Therefor, we place 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 (v8.1.1861_i86linux64). The publication is available at:

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

The similarity threshold chosen is 3.0%. Exactly 8'320 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 'Mothur Version 1.37.4' method and the non-redundant, no-gaps Silva version 123 database.

Out of our 8'320 OTUs, some are totally unclassified while others have predictions at different positions in the tree of life. The proportion of classified OTUs is summarized below:

#	Rank	Classified	Unclassified
1	**Domain**	8316	4

#	Rank	Classified	Unclassified
2	**Phylum**	7854	466
3	**Class**	3906	4414
4	**Order**	3208	5112
5	**Family**	2389	5931
6	**Genus**	1519	6801
7	**Species**	0	8320

Table 2. Summary information for all samples.

1.7 OTU filtering

At this point we are going to remove some OTUs. All those pertaining to any of the following phyla are discarded: Plastid and Mitochondrion. This leaves us with 8'320 'good' OTUs. As OTUs contain a varying number of sequences in them, we can plot this distribution in figure 2.

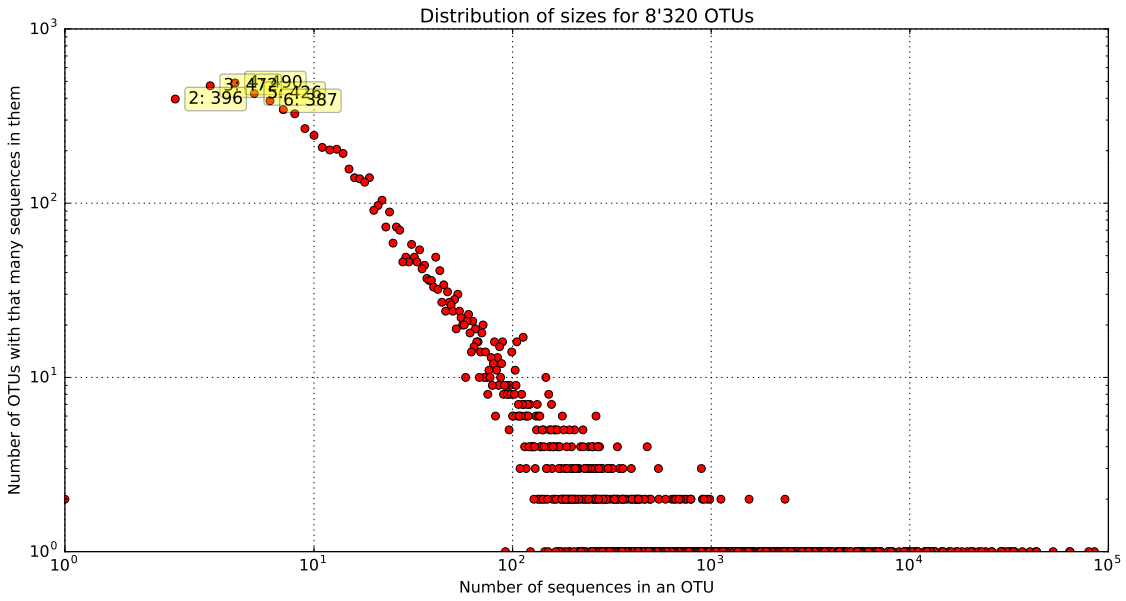

Figure 2. Distribution of OTU sizes

1.8 OTU table

Now we can take our good OTUs and pick them apart, producing a table with OTUs as rows (8'320) and samples as columns (39). Each cell tells us how many sequences are participating in the given OTU originating from the given 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:

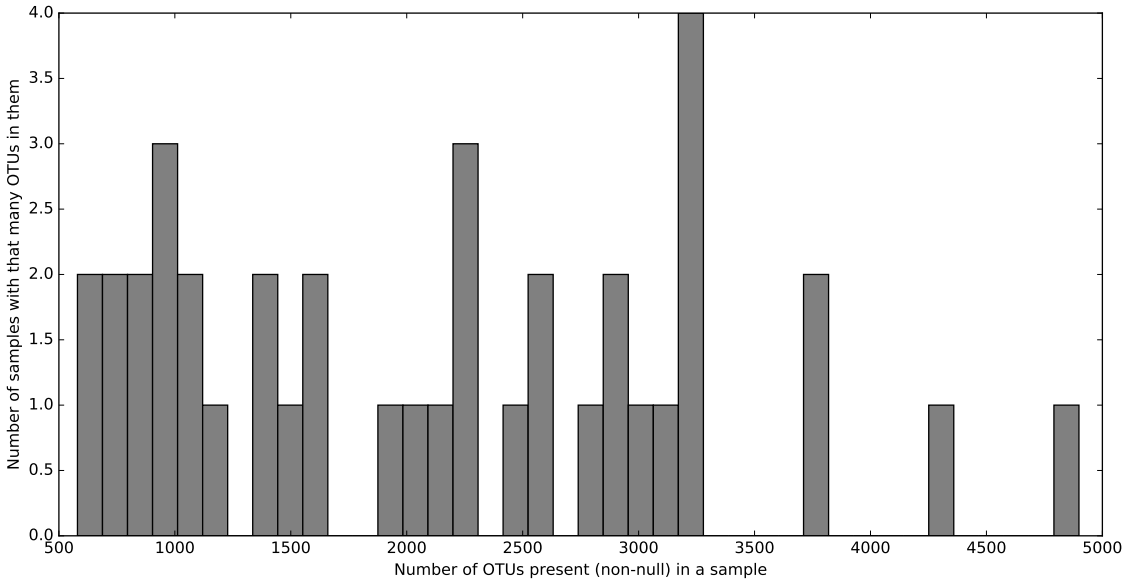

Figure 3. Distribution of OTU presence per OTU

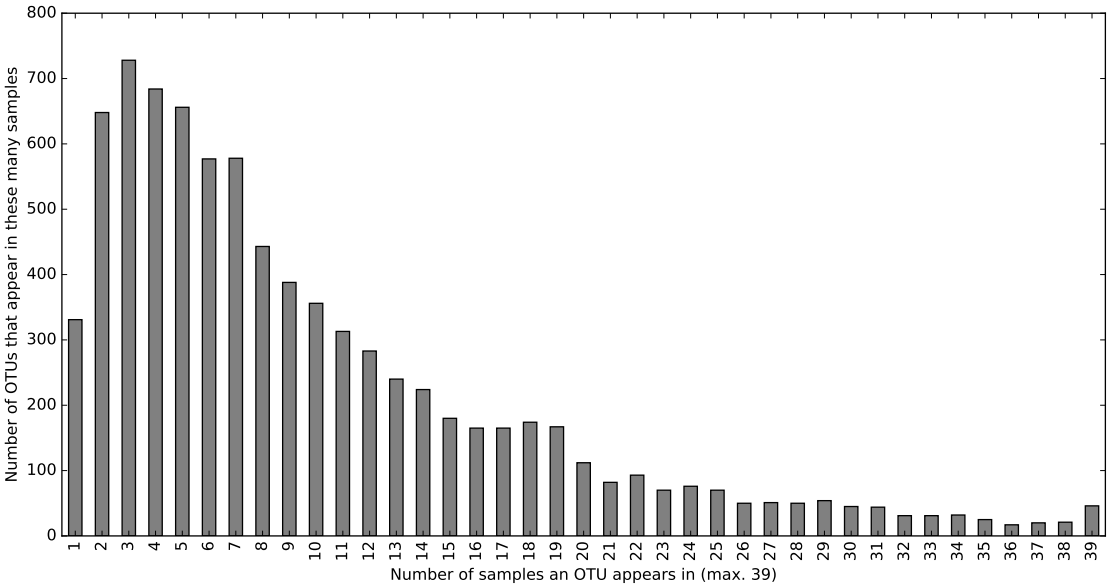

Figure 4. Distribution of OTU presence per sample

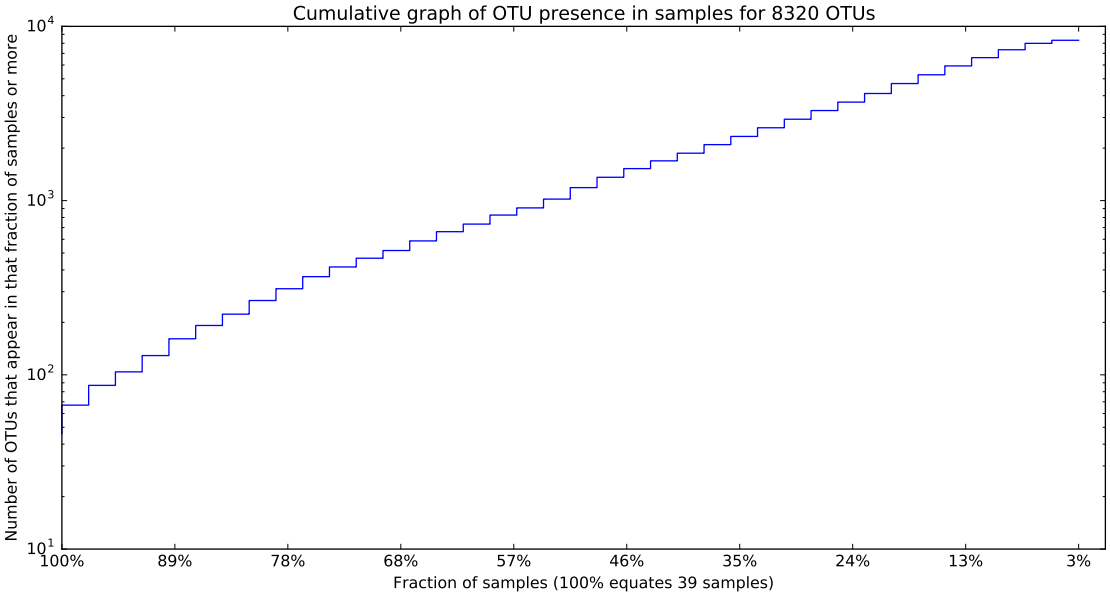

Figure 5. Cumulative number of reads by OTU presence

1.9 Taxa tables

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 procedure enables us to create taxa tables, which resemble OTU table somewhat. Such names can be made at several levels. It's important to consider the difference between an OTU table and a taxa table.

1.10 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. Once again, this can be done at several levels or ranks of classification ranging from Domain to Species. At levels that are too deep such visualization become too crowded and unreadable. This of course depends on the complexity of the samples. Here is piloted the 'phylum', 'class' and 'order' taxonomic levels in figures 6, 7 and 8:

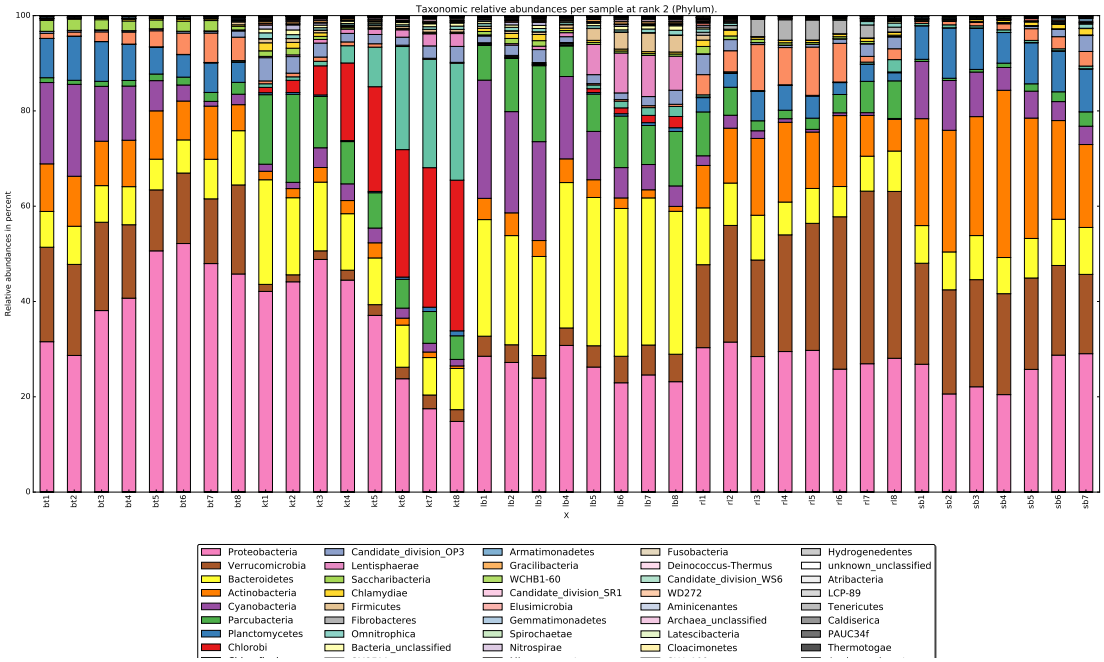

Figure 6. Relative abundances per sample on the phyla level

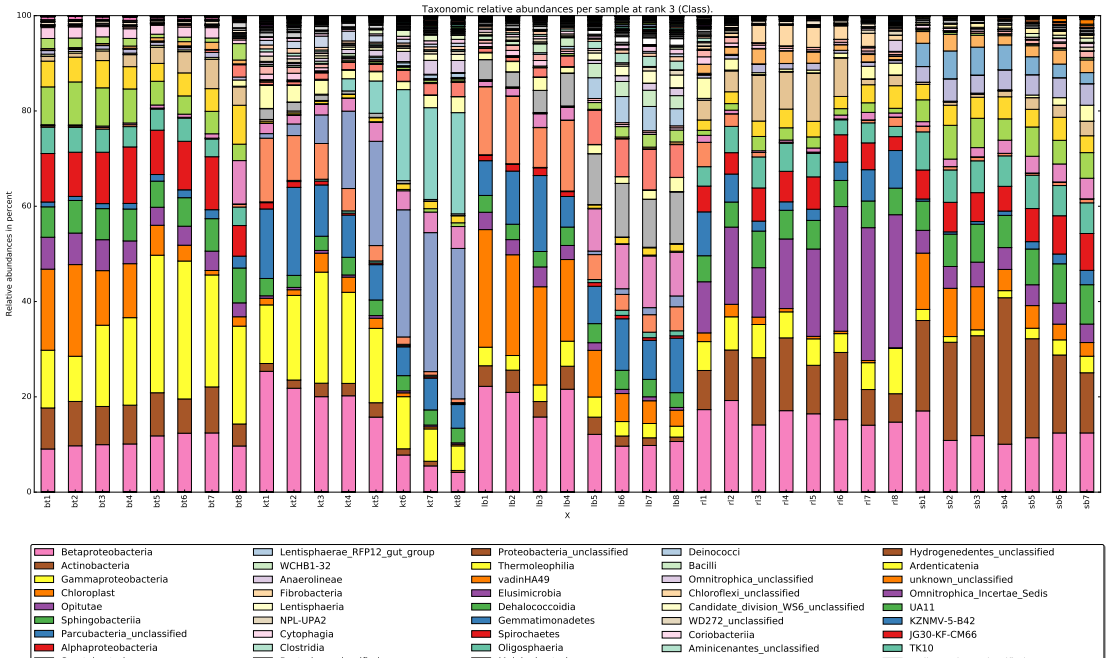

Figure 7. Relative abundances per sample on the class level

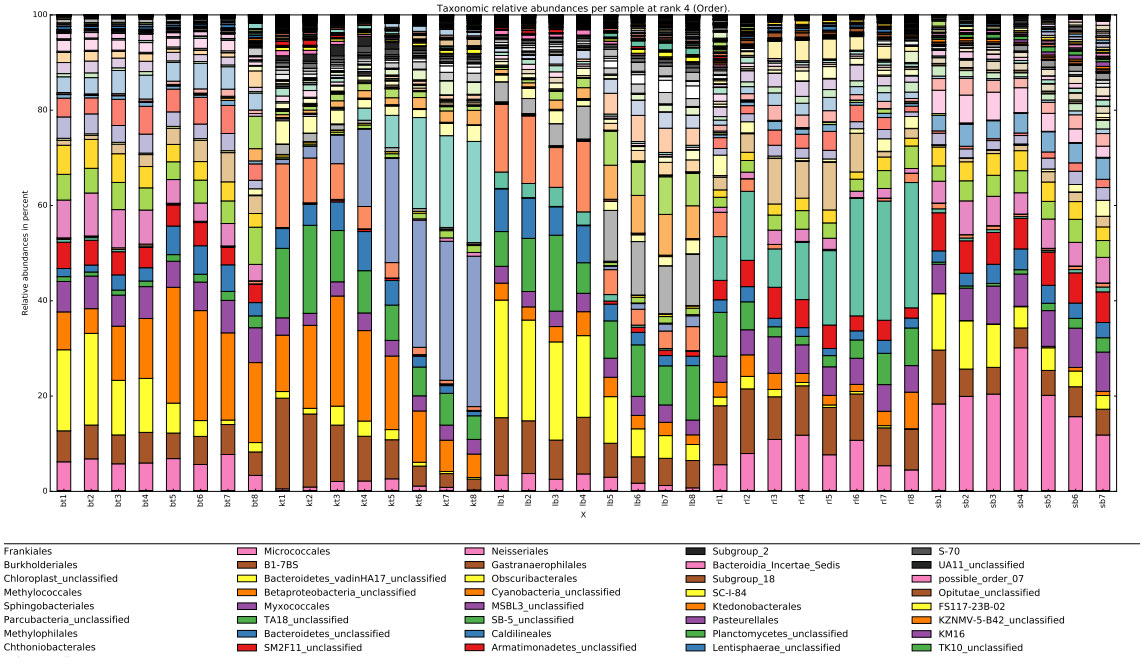

Figure 8. Relative abundances per sample on the order level

1.11 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 9.

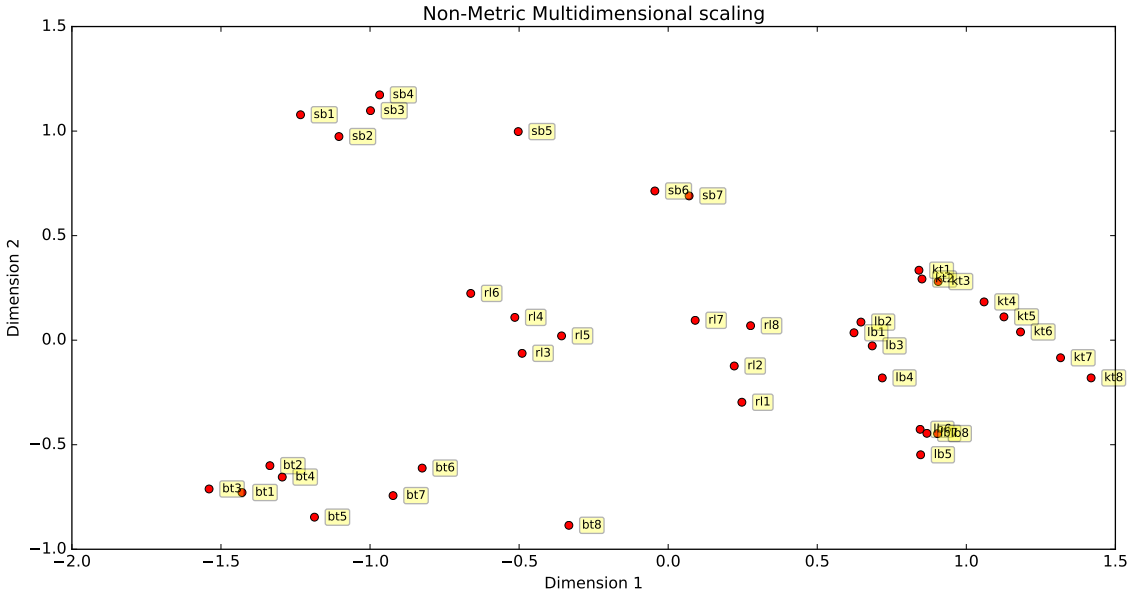

Figure 9. NMDS using the OTU table for 39 samples

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

1.12 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).

1.13 Alpha diversity

For each individual sample, we can compute several diversity estimators. More details on this procedure are available in each individual sample report. Here, a summary table is provided where the OTU table was downsampled (randomly rarefied) to 20'532 counts so that the different diversity estimates can be compared across samples.

#	Name	Chao1	Ace	Shannon	Simpson
1	**rl1**	3517.56	3688.83	8.37772	0.984376
2	**rl2**	3043.32	3498.55	7.12146	0.964135
3	**rl3**	2009.38	1946.26	6.49987	0.972029
4	**rl4**	1688.92	1855.26	6.31235	0.967856
5	**rl5**	2127.13	2041.82	6.04639	0.954659
6	**rl6**	2166.46	2438.44	5.80504	0.92341
7	**rl7**	2618.64	2695.95	6.55822	0.929449
8	**rl8**	2830.67	2804.78	6.6432	0.923175
9	**bt1**	1033	969.846	6.33921	0.977044
10	**bt2**	966	892.162	6.21784	0.97287
11	**bt3**	990.887	999.149	6.37018	0.978073
12	**bt4**	1085.01	1029.95	6.42862	0.977923
13	**bt5**	1221.74	1197.41	5.94876	0.950062
14	**bt6**	1192.38	1210.39	5.98446	0.952448
15	**bt7**	1183.9	1180	6.1481	0.964091
16	**bt8**	1667.94	1686.86	6.87542	0.975409
17	**lb1**	3200.01	3559.63	6.67039	0.964271
18	**lb2**	3460.33	3763.87	7.08327	0.967405
19	**lb3**	3664.06	3906.36	7.57044	0.976456
20	**lb4**	3207.5	3579.14	6.73399	0.970263
21	**lb5**	2965.22	3397.27	7.35811	0.979748
22	**lb6**	3255.34	3329.59	7.69068	0.981112
23	**lb7**	2956.45	2991.22	7.63968	0.982811
24	**lb8**	3255.5	3330.21	7.86685	0.981324
25	**kt1**	4448.72	4699.4	8.08775	0.974521
26	**kt2**	4828.04	4974.8	8.12403	0.972681
27	**kt3**	4006.23	4412.49	7.37829	0.967947
28	**kt4**	3149.08	3424.64	6.84278	0.963451
29	**kt5**	3230.25	3243.76	6.7037	0.952435
30	**kt6**	2673.77	2817.59	6.15066	0.923691
31	**kt7**	2681.2	2552.44	6.13152	0.914131
32	**kt8**	2432.93	2447.7	5.86469	0.902005
33	**sb1**	962.027	847.673	6.6651	0.982275
34	**sb2**	753.875	808.453	6.54975	0.98009
35	**sb3**	885.429	891.788	6.57592	0.979867

#	Name	Chao1	Ace	Shannon	Simpson
36	**sb4**	929.444	986.14	6.39365	0.97201
37	**sb5**	1354.26	1524.01	6.91616	0.981539
38	**sb6**	2221.67	2265.03	7.37414	0.98634
39	**sb7**	2344.13	2293.38	7.73675	0.989413

Table 3. Summary of diversity estimates for all samples.

1.14 Environmental tags

Relying on different kinds of databases and their meta-data, we can try to infer and assign 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 available upon request:

https://github.com/xapple/seqenv