PRACTICAL: Visualization of data & comparative metagenomics

Yesterday and the day before yesterday you have been working with a single metagenomic sample. In this exercise you should compare the taxonomic profiles from the same sample, and from different samples and visualise the comparison using different tools. In addition to the sample you already have become familiar with, you will be working with two new samples. Both samples, "*Muddy*" and "*Sandy*" are real metagenomic samples obtained from the marine sediment of the Barents sea. The original datasets were down sampled from approximately 4 million PE reads to 0,5 million PE reads to ease the analysis.

We have performed all the preprocessing steps prior to the assembly and the assembly of the metagenomes. We have also performed taxonomic analysis on the samples. The data is available here `/practical/6/`

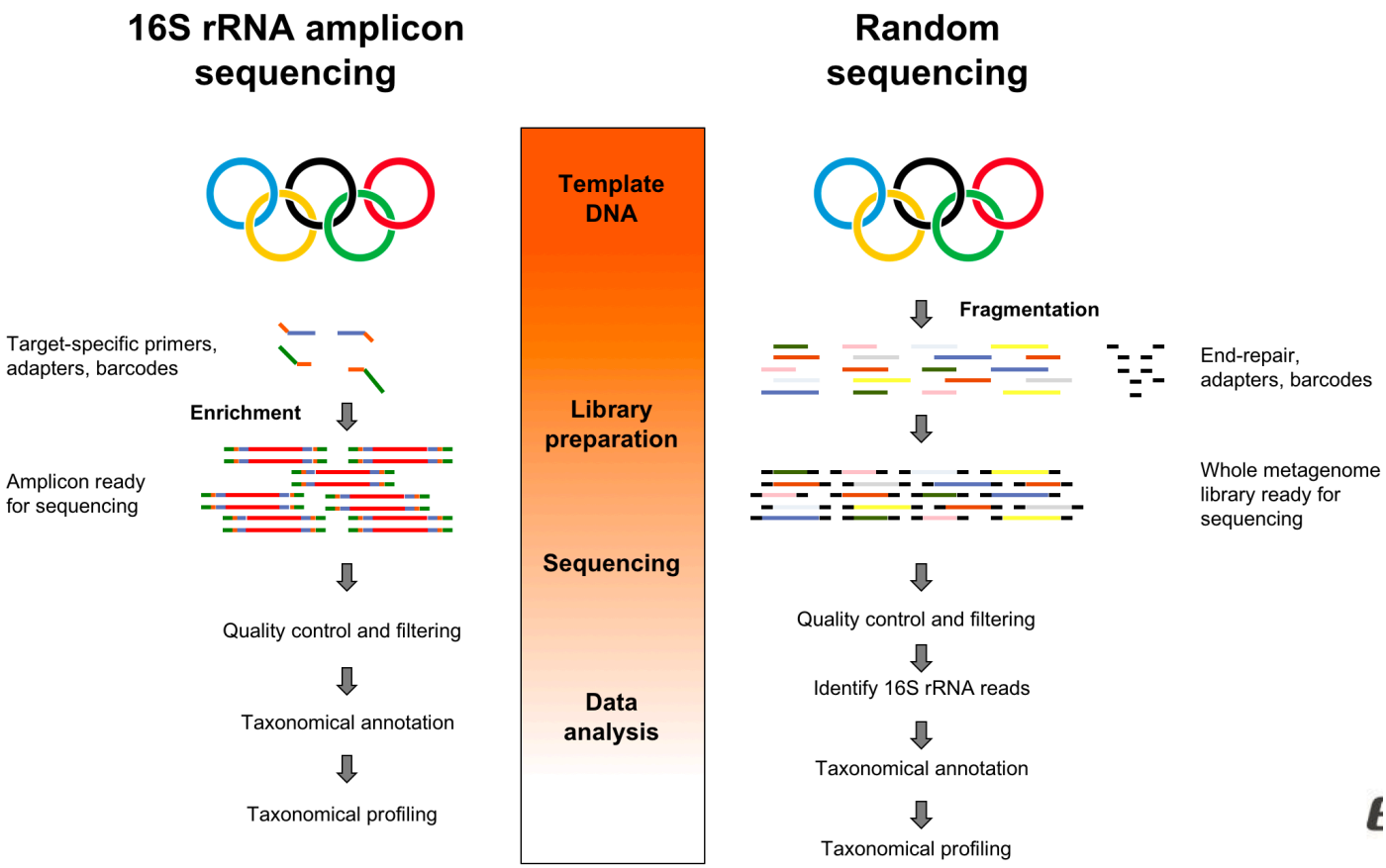
Later in the exercise, you will also look at the effect of using different methods to produce a taxonomic profile of a sample, by comparing the taxonomic profiles from the same samples obtained with different methods.

This exercise consists of six parts:

1. Compare taxonomic profiles
2. Compare assemblies
3. Prepare data for **MEGAN**
4. Get familiar with **MEGAN**
5. **MEGAN** compare samples
6. **MEGAN** compare multiple samples

1] Compare taxonomic profiles from the same sample

In this part of the exercise you should compare two taxonomic profiles from the same sample. The first is produced from amplicon sequencing of 16S rRNA. The second is produced from random shotgun sequences.



💡 Ideally, sample comparisons should be preformed and visualized in the same tool. However, your taxonomic profiles have been generated using very different tools and the results are not easily comparable in one tool. We are not saying it is impossible, but this requires that the taxonomic profiles are created a bit different than how you have done it.

You will compare the taxonomic profile from the **QIIME 2** analysis of the 16S amplicon data against the taxonomic profile from the **Kaiju** analysis of the corresponding shotgun data from the same sample (which we have prerun). This will be a crude way of comparing the taxonomic profiles, but it will hopefully show that the profiles will differ between shotgun and amplicon data.

First you will generate a taxonomic profile from the amplicon data of the sample named 44.S44. You have already done this for multiple samples including this, but now you will extract only this single sample and make a taxonomic barplot.

Identifier-based filtering is used to retain only a user-specified list of samples or features based on their identifiers (IDs) in a **QIIME 2** metadata file. In the directory `~/practical/6/compare/amplicon` we have

generated a list of samples that should be kept (only 44.S44) and named the list `samples-to-keep.tsv`. Here you will also find the same files as you were working with in the visualization of amplicon data exercise on Monday (`/~practical/1/prerun_all`).

I] Run the filter-samples method with the `samples-to-keep.tsv`. The resulting table will contain only the sample whose ID are listed in `samples-to-keep.tsv`:

```
qiime feature-table filter-samples --i-table table_all.qza --m-metadata-file sample

Output artifacts: 44.S44-table.qza
```

II] Generate taxonomic barplot of sample 44.S44:

```
qiime taxa barplot --i-table 44.S44-table.qza --i-taxonomy taxonomy_all.qza --m-met

Output visualizations: 44.S44-taxa-bar-plots.qzv
```

III] View the taxonomic composition of sample 44.S44 with the interactive bar plot:

```
qiime tools view 44.S44-taxa-bar-plots.qzv
```

IV] Set the taxonomic level to 2 (phylum) and leave the bar plot open in a web browser.

We have prerun Kaiju on the shotgun data from the same sample as you run the 16S amplicon analysis on. The data is located here: `~/practical/6/compare/shotgun`

V] The **Kaiju** output was generated using the following command. ALETERNATIVELY you can run the analysis yourself:

```
kaiju -t /net/software/databases/kaijudb/nodes.dmp -f /net/software/databases/kaiju
```

VI] Convert the output file into the proper **Krona Tools** format using **kaiju2krona**:

```
kaiju2krona -t /net/software/databases/kaijudb/nodes.dmp -n /net/software/databases
```

VII] The file `muddy_kaiju_mar.krona` can then be imported into **Krona** and converted into an HTML file using Krona's **ktImportText** program:

```
ktImportText -o muddy_kaiju_krona_mar.html muddy_kaiju_mar.krona
```

VIII] Open the HTML file containing the taxonomic profile generated using **Kaiju** in a web browser.

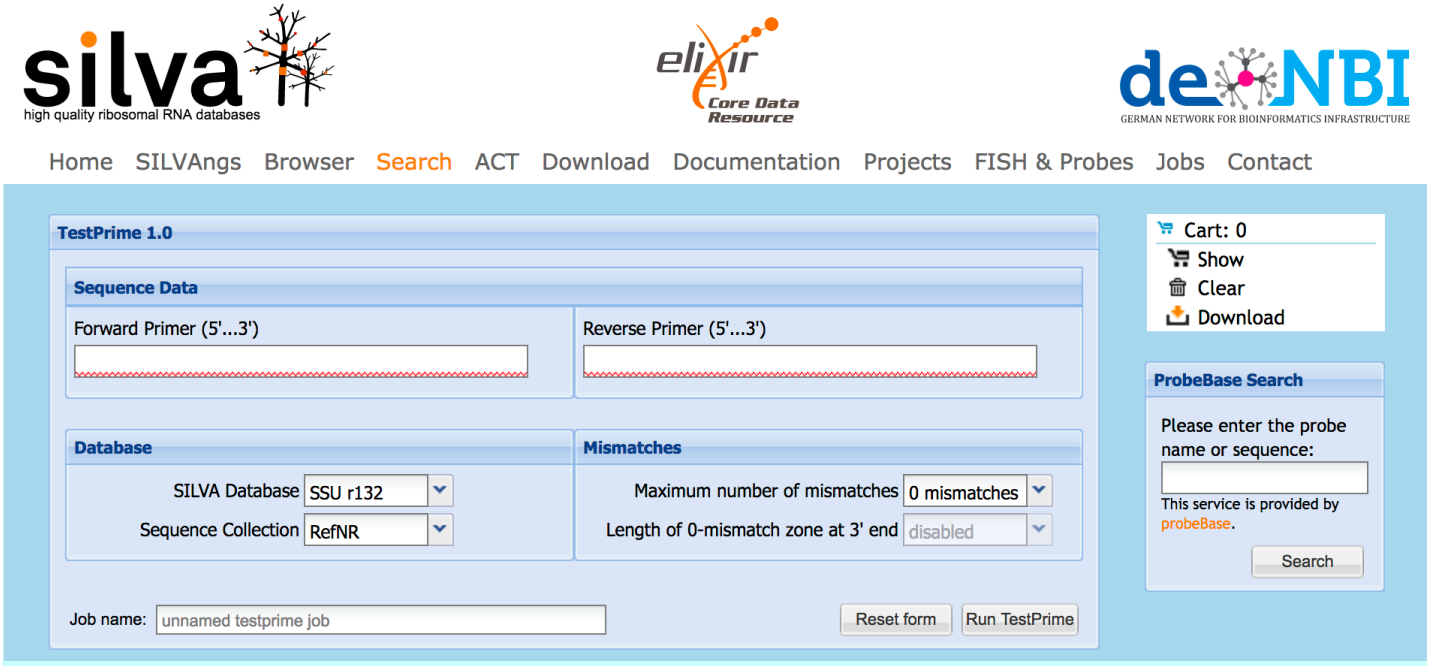
💡 **NOTE:** The taxonomic levels and naming of them varies between SILVA taxonomy (QIIME 2) and NCBI taxonomy (RefSeq). For example:

Taxonomy	Level 1	Level 2	Level 3	Level 4	....
SILVA	Bacteria	Actinobacteria	....	....	....
AW31	Bacteria	Terrabacteria group	Actinobacteria	....	....

❓ Can you identify any taxonomic group that is similar between the two samples?

▶ **Solution** - Click to expand

💡 **SILVA TestPrime** allows you to evaluate the performance of primer pairs by running an in silico PCR on the SILVA databases. From the results of the PCR, **s** computes coverages for each taxonomic group in all of the taxonomies offered by SILVA. These coverages can then be inspected in our taxonomy browser, making it easy to quickly identify strengths and weaknesses of a particular pair of primers.



The output from **SILVA TestPrime** will contain this type of information:

```
taxonomy    coverage    specificity
Bacteria;Acidobacteria;Fischerbacteria; 100.0   25.7
Bacteria;Acidobacteria;GBS-L1-B05;  0.0 0
Bacteria;Firmicutes;    88.2    29.5
Bacteria;Firmicutes;Bacilli;    88.4    26.9
Bacteria;Firmicutes;Bacilli;Bacillales; 88.0    26.4
```

From this you can deduse that if there were any Acidobacteria GBS-L1-B05 in the sample, they would most like not be detected using these amplification primers.

Espen Robertsen made a similar comparison between 16S amplicon and shotgun data. below are some observations he made:

Taxonomy		Amplicon		Shotgun	
Level	Taxa	Fraction	Primer Coverage	Fraction	Comp.ab
Order	Enterobacteriales	0	79.6	0.05	0
Phylum	Verrucomicrobia	0.0001	2.5	0.008	0.01
Class	Anaerolineae	0.0003	28.1	0.01	0.02
Phylum	Planctomycetes	0.008	0.7	0.12	0.06
Phylum	Lentisphaerae	0.001	5.1	0.01	0.1
Class	Betaproteobacteria	0.006	78.5	0.04	0.16
Phylum	Chloroflexi	0.01	42.7	0.03	0.36
Phylum	Gemmatimonadetes	0.02	85.4	0.01	0.47
Phylum	Candidate Div WS3	0.02	68.4	0.09	0.55
Phylum	Proteobacteria	0.71	79.4	0.44	0.62
Phylum	Acidobacteria	0.06	80.8	0.04	0.73
Phylum	Actinobacteria	0.02	75.6	0.01	0.74
Phylum	Cand. Div BRC1	0.005	10.1	0.007	0.76
Phylum	Nitrospirae	0.01	82.4	0.02	0.94
Phylum	Bacteroidetes	0.05	79.7	0.05	0.98

❓ Can you identify any taxonomic group that is differing between the two samples?

▶ **Solution** - Click to expand

Progress tracker

Part 1 finished

2] Compare assemblies - methods and samples

We have prerun the assembly of the two environmental samples using **MEGAHIT**. The final assembly files

(contig files) and the assembly you generated the first day of the course are located here: `~/practical/6/quast`

Here is a little trick: If you just want a quick overview of the statistics of the assemblies, just run **QUAST** instead of **MetaQUAST**. By default, **QUAST** does not try to identify taxon and match your input against reference genomes like **MetaQUAST** does.

I] Go to the directory: `~/practical/6/quast` and run **QUAST** to compare the three assemblies:

```
quast.py *.fasta
```

II] Open the QUAST report in a web browser

❓ Which assembly is the largest and how many bp is it?

▶ **Solution** - Click to expand

❓ Are you likely to find many complete genes in the sandy and muddy sample?

▶ **Solution** - Click to expand

Progress tracker

Part 2 finished

3] Prepare data for MEGAN

MEGAN is a comprehensive toolbox for analyzing microbiome data and comes with a user-friendly interface. **MEGAN** can perform both taxonomic and functional analysis and visualise the results in various plots and charts. You can read more about the tool here. The main application of **MEGAN** is to parse and analyze the result of a **BLAST** comparison of a set of reads against one or more reference databases, typically using **BLASTN** to compare against NCBI-NT. **MEGAN** also supports import of data from other programs in a delimiter-separated format (using comma's or tabs). An example of such file is:

Taxon	#reads
Specie A	100
Specie B	400
Specie C	50

The best way to create a taxon summary from a **Kaiju** output file that include all taxonomic levels, is to generate a file a tab-separated text file using **kaiju2krona**. You made this conversion several times yesterday.

I] Go to the directory: `~/practical/6/megan`. The directory should contain four files. Look at the first part of the file `sample_kaiju.out`

```
head sample_kaiju.out
```

II] Convert the **Kaiju** output to a tab-separated text file using **kaiju2krona**:

```
kaiju2krona -t /net/software/databases/kaijudb/nodes.dmp -n /net/software/databases
```

III] Open the output from **kaiju2krona**. It should look similar to this (only showing three lines):

```
20  root    cellular organisms  Bacteria    Proteobacteria  Gammaproteobacteria Aer
69020   root    cellular organisms  Bacteria    Proteobacteria  Gammaproteobacteria
253 root    cellular organisms  Bacteria    Proteobacteria  Gammaproteobacteria Ent
```

IV] Convert the **Kaiju** output to a **MEGAN** readable input:

```
sort -n -r sample_kaiju.krona | awk '($1 > 100)' | awk -F '\t' 'BEGIN{FS=OFS="\t"}{

# sort will sort the taxons by counts with the most abundant at the top
# awk remove all coloumns except the first and the last and reorder the two remaini
```

V] Open the converted output from **kaiju2krona** either in a text editor or using **head**.

It should look similar to this (only showing 10 lines) with the first column displaying taxon, and the second showing how manu sequence reads that mapped to this taxon. Note that the taxons are on various taxonomic levels (genus, species, etc.):

```
Photobacterium  339466
Aliivibrio  18606
Vibrio  9117
Vibrionaceae    6197
Moritella   4812
Gammaproteobacteria 4192
Vibrio harveyi group    4059
Shewanella  3639
Photobacterium damselae 3497
Photobacterium phosphoreum  2490
```

❓ Count the number of lines (taxons) in the file

▶ **Solution** - Click to expand

💡 **Reduce number of taxa.** It is very normal to reduce the number of taxa in a taxonomic profile, by filtering out the least abundant taxa. This can be done for example by removing all taxa with less than 1 % abundance, or by removing all taxa with less reads than a certain number we set as a threshold. Reducing the number of taxa will also make viewing in **MEGAN** easier.

VI] Open `sample_kaiju_megan.tab` in a text editor and remove all taxons with less counts than 500 sequence reads. Name the reduced file `sample_kaiju_megan_500.tab`

VII] Repeat the above steps (II] - VI]) on the **Kaiju** output files from the "*Sandy*" and "*Muddy*" samples.

In addition to these three samples, we have prerun the above steps on the results from taxonomic profile of the 16S rRNA you did yesterday. The "meganized" file is named `sample_16S_megan.tab` .

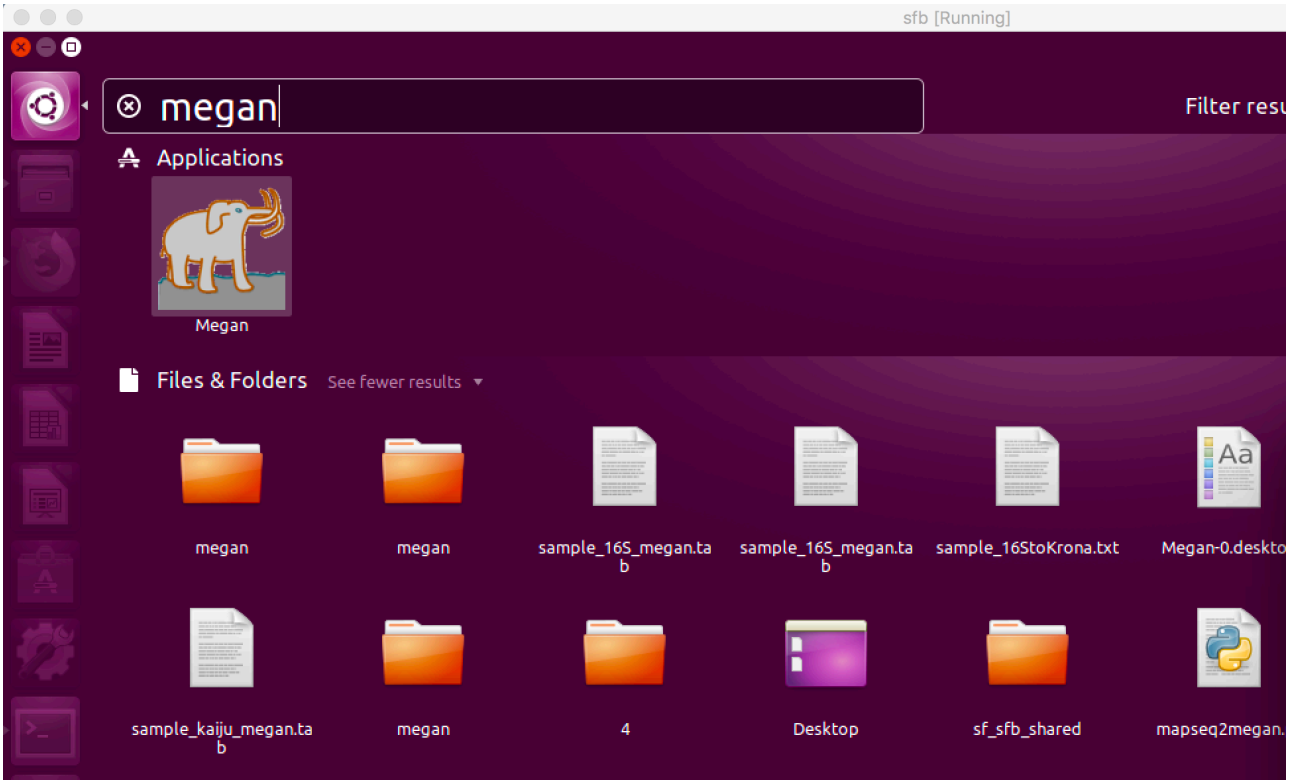
Progress tracker

Part 3 finished

4] Get familiar with MEGAN

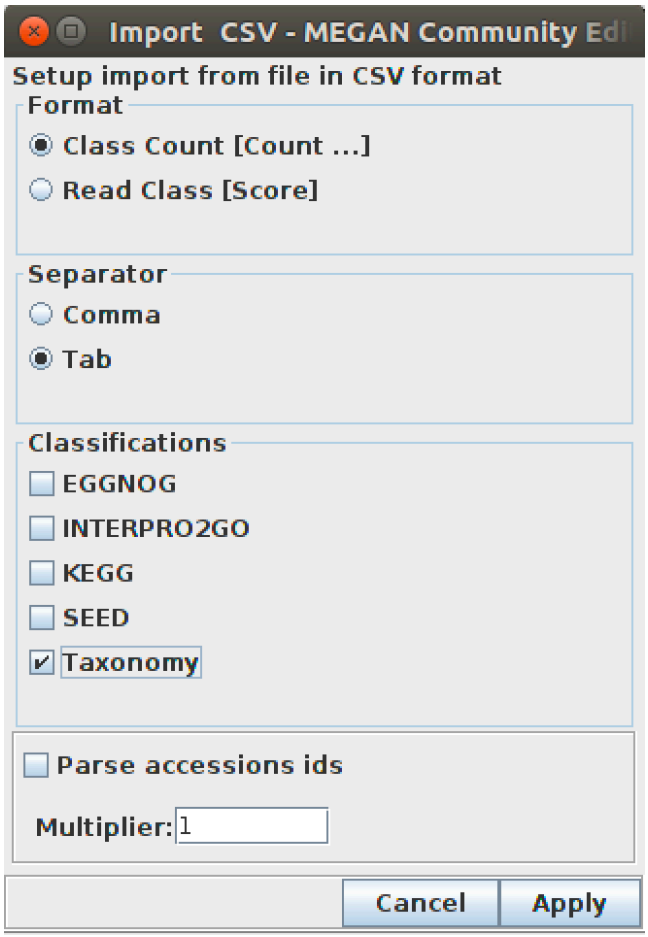
The complete **MEGAN** manual can be found here.

I] Open **MEGAN** either by clicking the "**Search your computer**" icon on the top left of your desktop and type `megan` in the search field, or you can start **MEGAN** by double clicking the short cut on the desktop.

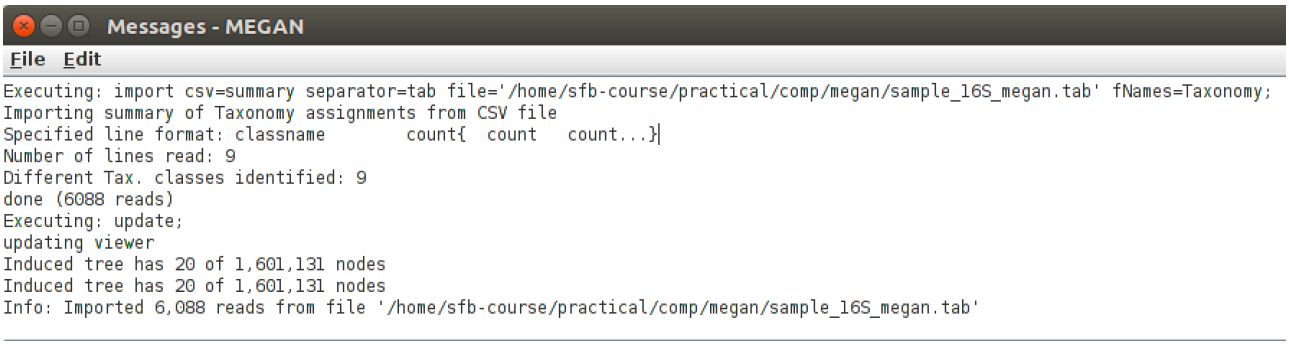


Once **MEGAN** starts, two new windows pops up: one containing a tree structure and one Message window. Import the "meganized" file `sample_16S_megan.tab` (located in `~/practical/6/megan`). **MEGAN** do not keep track of sample names inside the tab-delimited, but by importing them individually we can keep track of the samples by their file names.

II] In **MEGAN**, click on the "**File**" menu in the tree window, select Import and then select CSV format. A file browser window appears. Browse to and select `sample_16S_megan.tab`. A new window asking you to define the import format will appear. Tick off "**Taxonomy**" under "**Classification**" and press "**Apply**".



A new window with the taxonomic assignments on a hierarchical tree structure will appear. In addition there is some information about the file you just imported in the "**Message window**":

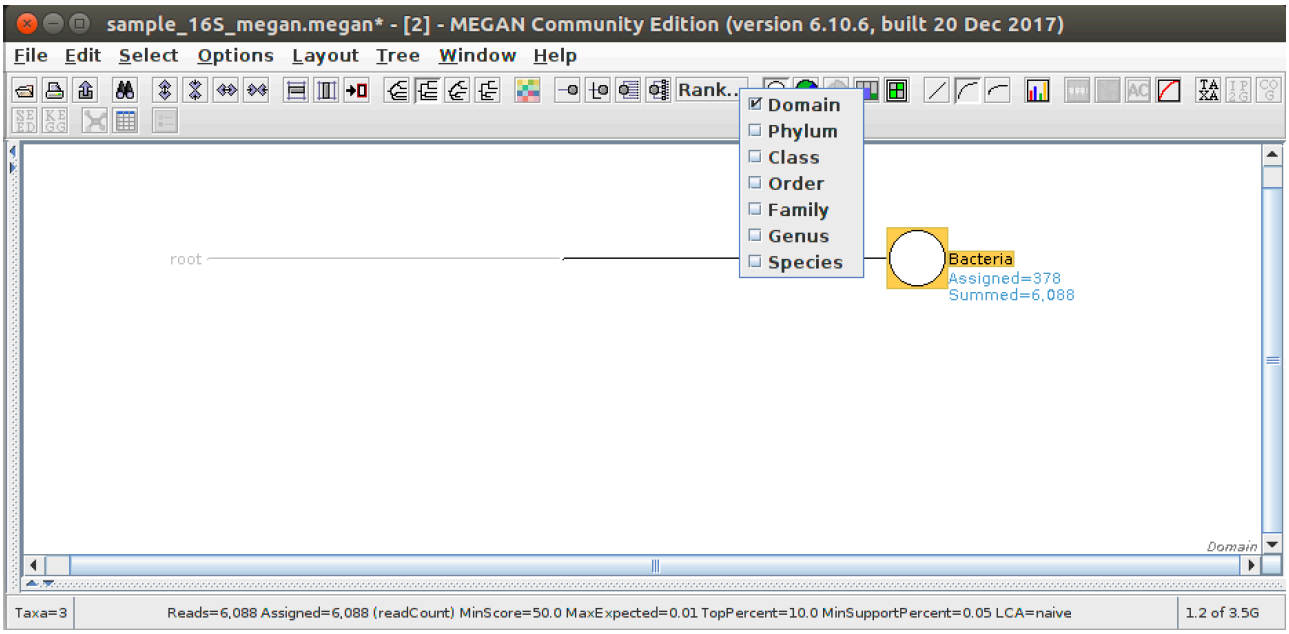


As you can see, there is 9 taxa present in the sample consisting of 6088 (16S rRNA) sequence reads.

III] Click on the circle representing Bacteria. Two numbers will appear: Assigned means how many reads that were classified to the Bacteria taxon - that is on domain level; Summed means how many reads in total in the sample that are assigned to this AND deeper hierarchical levels of this domain.

❓ Can you try to explain why these reads are not placed further down in the taxonomic hierarchy (at family or genus level)?

▶ **Solution** - Click to expand



On the top of the hierarchical tree structure window, there are many options for what you can do with the data. You can mouse over all the symbols at the top to see their functions.

IV] Press the "**Rank**" button, and collapse the tree to the Family level.

❓ What does the different sizes of the circles represent?

▶ **Solution** - Click to expand

❓ How many reads are in total assigned to the family *Vibrionaceae*?

▶ **Solution** - Click to expand

V] Click on the "**Rank**" button and select Species. **MEGAN** will try to expand the taxonomic tree all the way to species level.

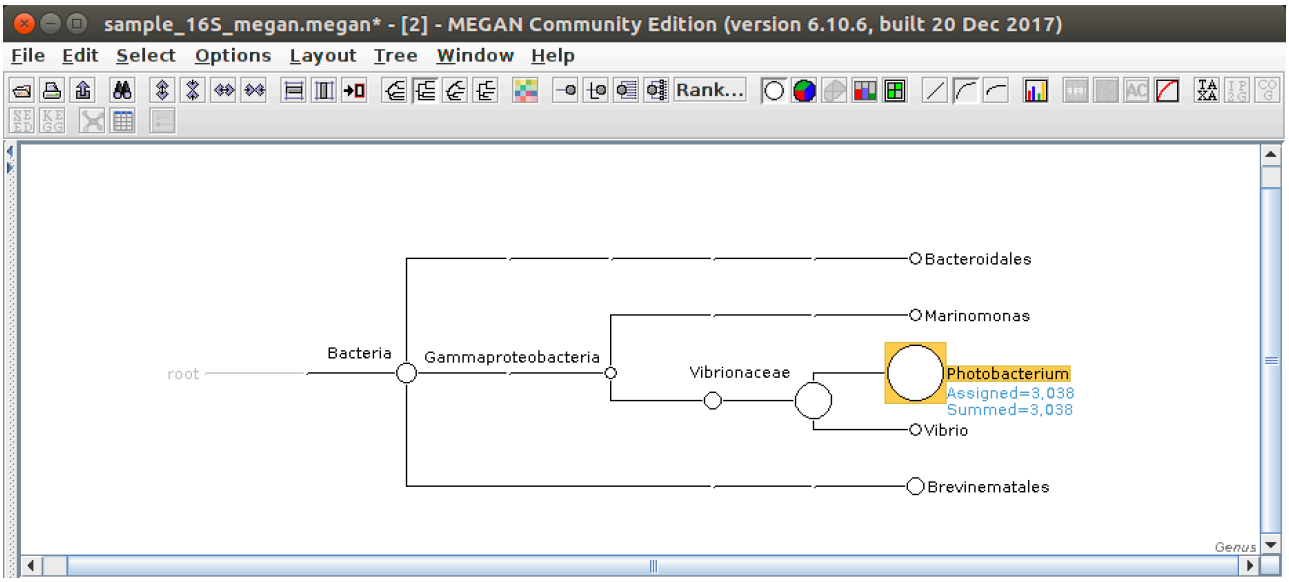
❓ What is the lowest taxonomic level of this data, and why do you think there are no leaves on species level?

▶ **Solution** - Click to expand

VI] Click on the circle representing *Photobacterium*.

❓ Why is the number of assigned reads identical to the number of summed reads?

▶ **Solution** - Click to expand

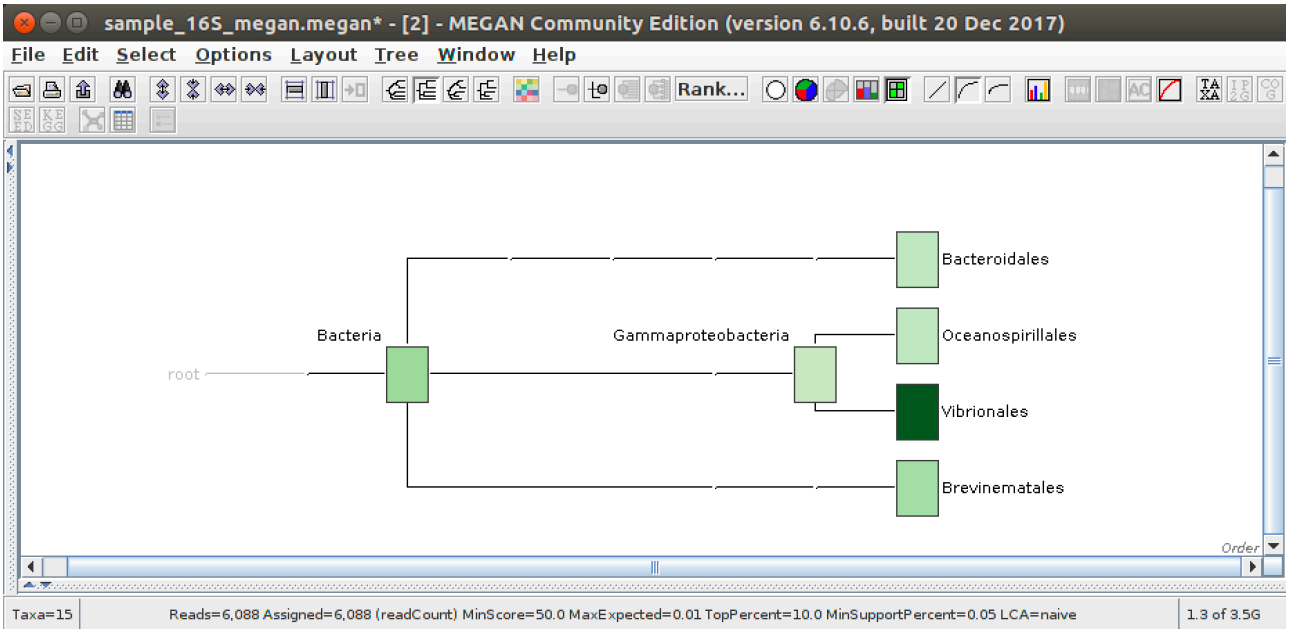


MEGAN has an interactive GUI, so you can select several nodes in the window by pressing **Shift + mouse** button. Alternatively, go to the "**Select**" menu.

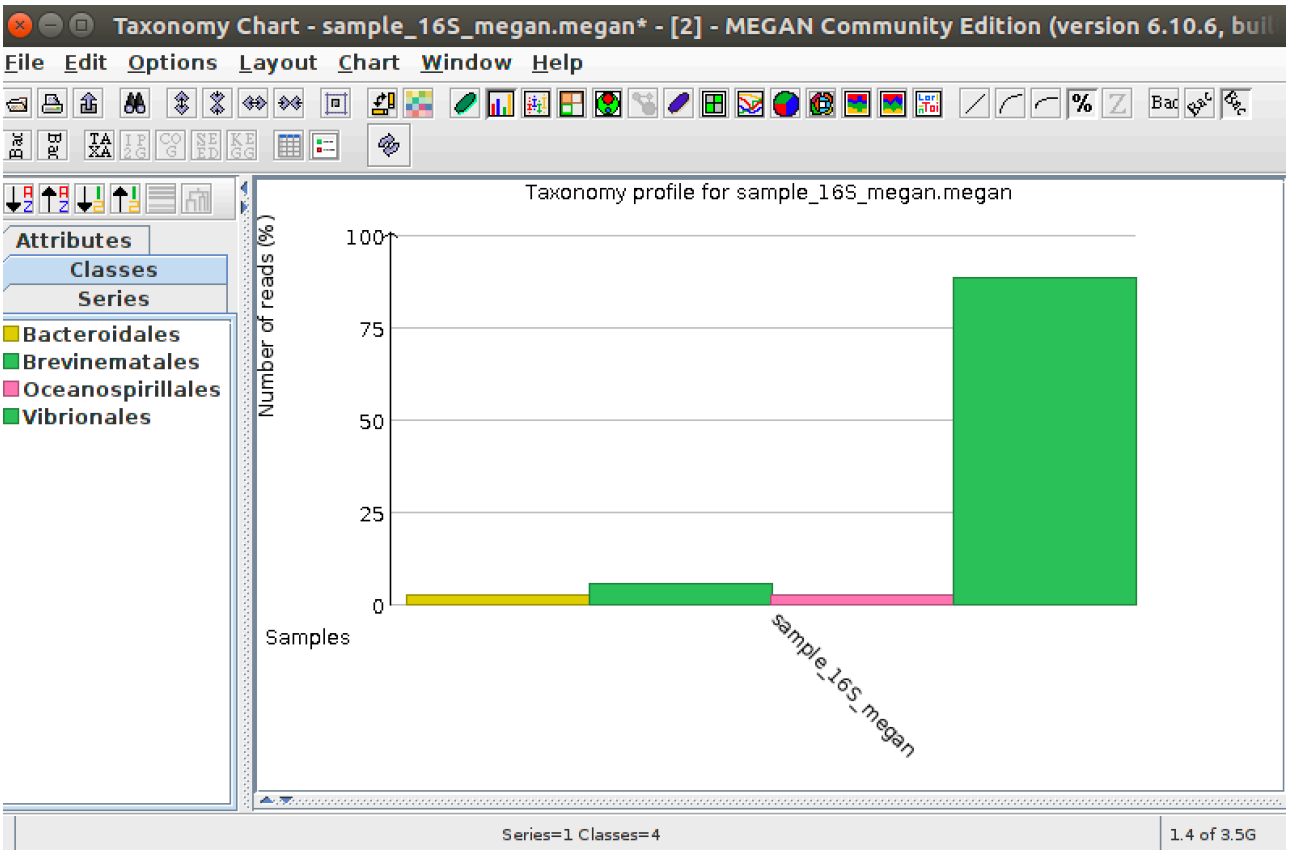
VII] Either way you choose, select all leaves at order level (there are no taxa at species level), and click on the "**Draw data as heatmaps**".

? What does the darker green colour mean?

▶ **Solution** - Click to expand



VIII] Click on the symbol "**Show charts**" and select "**Show bar charts**". In the new window showing the distribution of reads at the level of the sample that appears, click on the "**Classes**" tab in the left part of the window.



IX] Try to change the appearance of the taxonomic profile by selecting some of the other chart views on the top, for example "**Bubble chart**"

❓ Approximately how many percent of the reads belong to *Vibrionales*? Hint: Press the % sign at the top.

▶ **Solution** - Click to expand

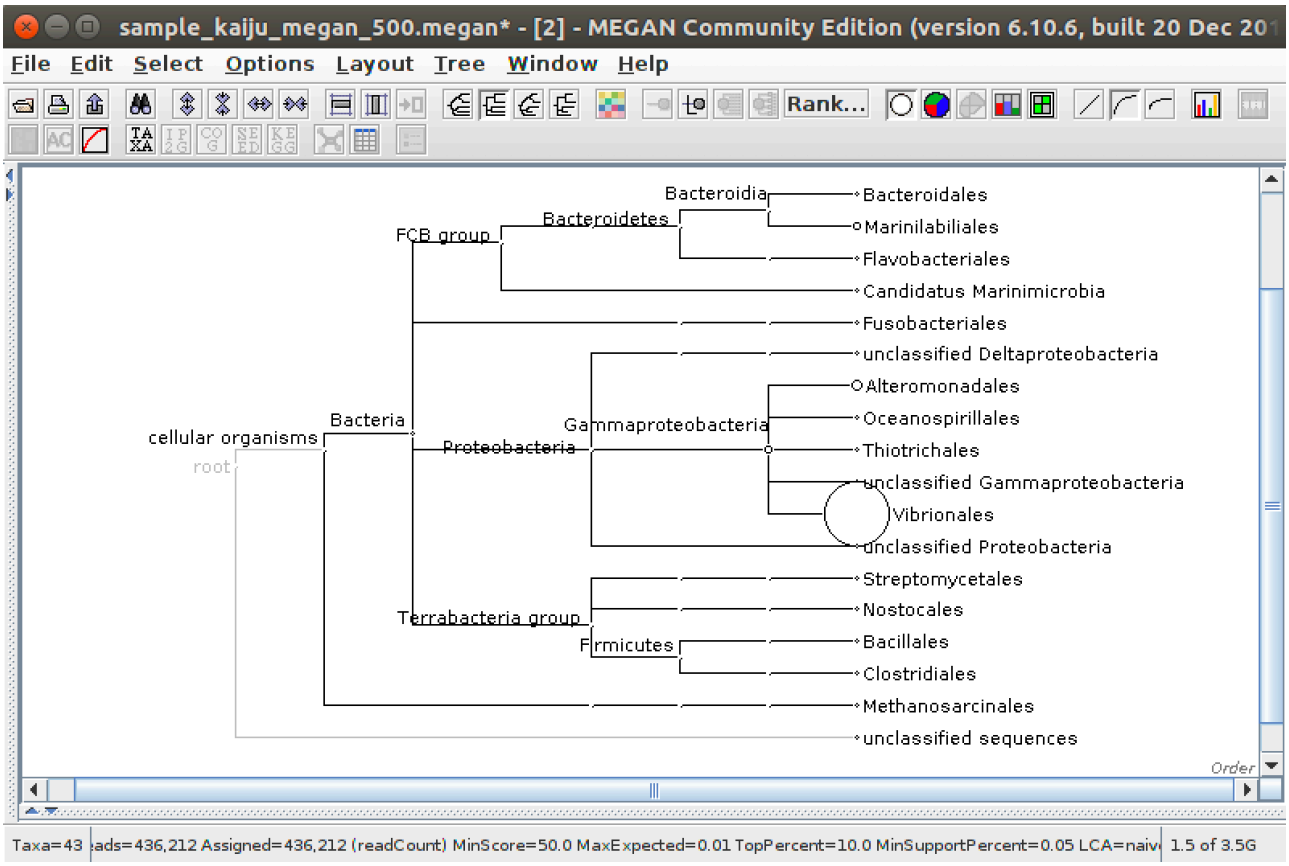
X] Make a word cloud of the taxonomic profile before you close this window (leave the tree window open).

Progress tracker

Part 4 finished

5] MEGAN compare samples

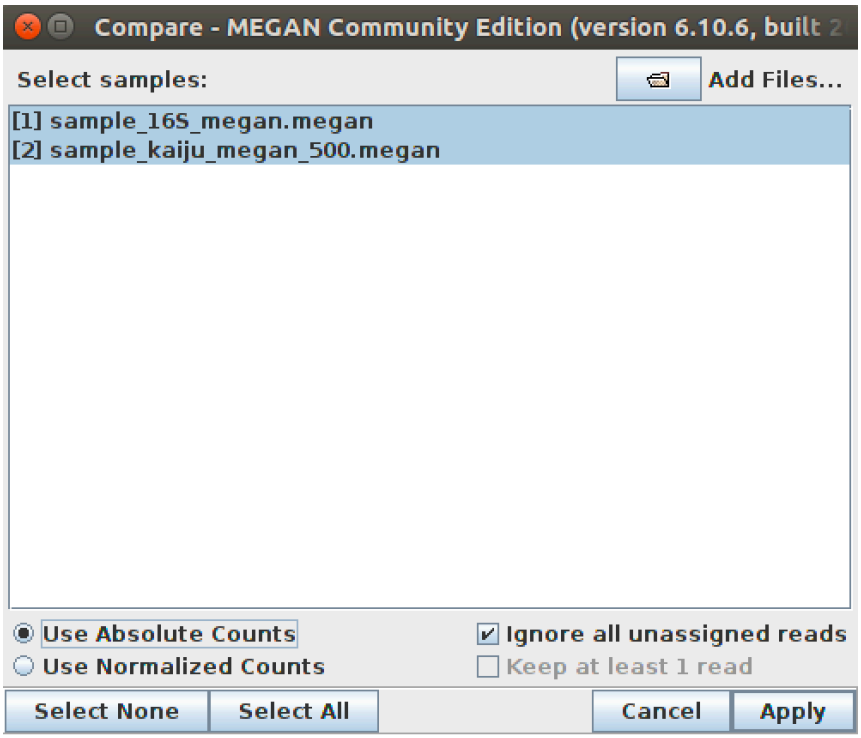
I] Click on the "**File**" menu in the open tree window, and repeat the import of the file `sample_kaiju_megan_500.tab` like you did previously.



II] Press the "**Rank**" button, and collapse the tree to the Order level. Select all leaves and press the "**Draw data as heatmaps**" button.

You should now have two hierarchical tree structure windows open from the same sample, one for the taxonomic profile generated using the 16S rRNA method and one generated by **Kaiju**. Now we can use **MEGAN** to compare the samples against each other.

III] Click on the "**File**" menu and select "**Compare**", select both samples. Select "**Use Absolute Counts**", and press Apply.



A new window with a hierarchical tree structure will open, only this time it will contain both samples.

IV] Press the "**Rank**" button, and collapse the tree to the Order level and press the "**Show chart**" button and select "**Show bar chart**".

Initially, it may look like there is no taxonomic information in the 16S sample.

V] Increase the size of this window and select the *Vibrionales* order in the left panel.

❓ How many reads are assigned to *Vibrionales* for each of the samples?

▶ **Solution** - Click to expand

❓ Approximately how many percent of the reads are assigned to *Vibrionales* for each of the samples?

▶ **Solution** - Click to expand

VI] Try to change the appearance of the taxonomic profile by selecting some of the other chart views on the top.

VII] Finally, make a word cloud of the taxonomic profiles before you close the comparison window.

VIII] Repeat the above to import the two files `sandy_kaiju_megan_500.tab` and `muddy_kaiju_megan_500.tab` .

IX] Click on the "**File menu**" and select "**Compare**", select both samples. Select "**Use Normalized Counts**", and press "**Apply**".

X] Press the "**Rank**" button, and collapse the tree to the Order level and press the "**Show chart**" button and

select "**Show bar chart**".

VI] Change the appearance of the taxonomic profiles to the view that best describes which samples are most alike.

❓ Which sample are most similar to each other?

▶ **Solution** - Click to expand

Progress tracker

Part 5 finished

6] MEGAN compare multiple samples

In the final part of this exercise, you are given a completely different dataset. These samples are not of marine origin, but isolated from the gut of premature human infants. The microbial diversity in the gut of premature infants is very low, and therefore more suitable for this exercise. In the directory `~/practical/6/multiple_megan/` you will find two files that you will be working with.

You are given eight samples, where four infants have been treated with probiotics, and the other four have not been treated - Hence two groups: **Treated** and **Untreated**. In addition, the different individuals have been given a combination of different antibiotic (AB) treatments, some have been delivered by caesarean section, etc. All the metadata about the mode of delivery and treatment is found in the file `multiple_metadata.txt`.

I] Open the file containing all metadata in a text editor. The samples are ordered in rows, while the metadata information is ordered in columns.

❓ How many different types of metadata do you find pr sample?

▶ **Solution** - Click to expand

We have already performed taxonomic profiling of the samples using the 16S rRNA approach on shotgun metagenomic data you have done previously in this course. The `multiple_megan.txt` is a tab separated file with the taxonomic profiles of 8 samples.

II] Open the file containing the taxonomic profiles in a text editor. The samples are ordered in columns, while the taxonomic counts are ordered in columns.

❓ How many different types of taxons do you find in total?

▶ **Solution** - Click to expand

III] Open the file `multiple_megan.txt` in **MEGAN** like previously (File -> Import -> Text (CSV format) -

> Taxonomy)

IV] Click the "**Draw data as heatmaps**" button. You will see the relative distribution of counts for all taxons in all eigth samples in the tree window.

V] Try to swith the tree view between Family and Genus level.

❓ What do you think is the reason for that a large proportion of the reads in the *Enterobacteriaceae* family are not possible to classify lower in the taxonomic hierarchy?

▶ **Solution** - Click to expand

Because of this, we suggest that you perform the analysis on Family level.

VI] Click on the symbol "**Show charts**" and select "**Show bar charts**".

In the new window showing the distribution of reads at the Family level of the sample that appears.

❓ What does the different coulored bars represent?

▶ **Solution** - Click to expand

VII] Alternate the view between "**Classes**" and "**Series**" clicking the tabs in the left part of the window.

VIII] Double click on either a taxon or a sample and see that the data from this particular sample/taxon is hidden.

Leave this window open.

The metadata

I] Import the metadata through the "**File**" tab. If asked to clear existing metadata, choose "**Yes**". A new window called "**Sample viewer**" will appear.

❓ How many individuals were born by caesarean section (sectio)?

▶ **Solution** - Click to expand

❓ How many individuals were the antibiotic (AB) Vancomycin (vanco)?

▶ **Solution** - Click to expand

❓ Which individual received the longest duration of antibiotic (AB) treatment?

▶ **Solution** - Click to expand

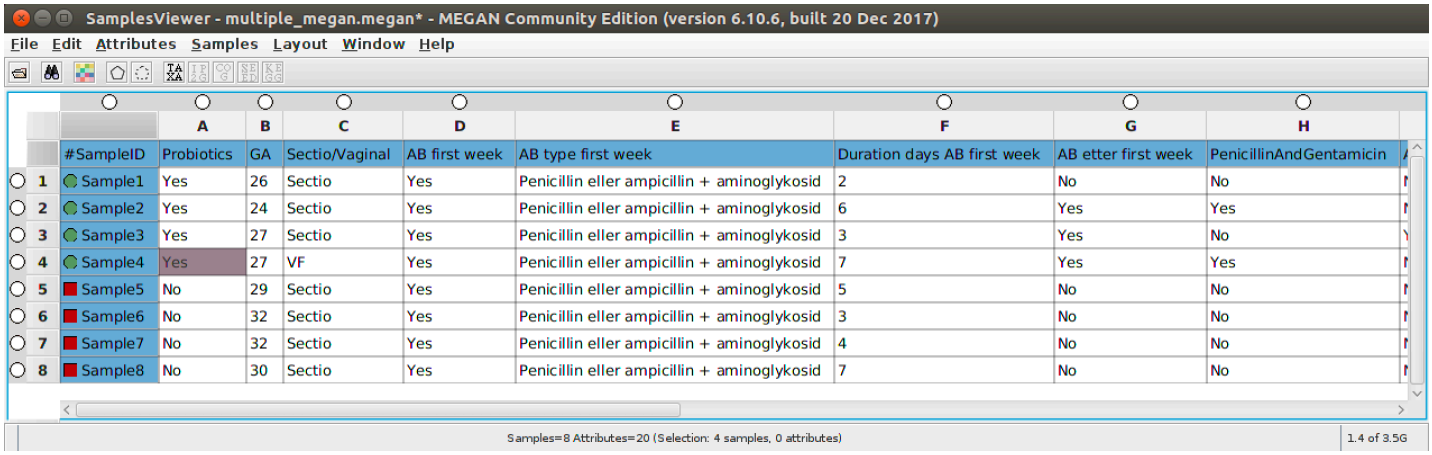
💡 **Note:** You can change and manipulate the metadata using the sample viewer in order to better view

your data after analysis. For example, you can label samples that belong to the same group (Treated and Untreated) similarly. Sample 1-4 belong to the probiotic Treated group and 5-8 belong to the Untreated group (see column A - "Probiotics"). In addition to this, you can add more colums with information if you want to.

II] Click on the circle symbol above the "**SampleID**" column and sort the samples increasingly.

III] Select Samples 1-4 and click on the circle symbol to the left of one of the selected samples. From the drop-down menu, select "**Set shape**" and choose "**oval**" for this group. Click on the circle symbol to the left again and select the colour drop-down menu at the very buttom of the list and choose a green colour.

IV] Select Samples 5-8 and repeat the above, except choose "**rectangle**" shape and a red colour for this group.



		A	B	C	D	E	F	G	H
	#SampleID	Probiotics	GA	Sectio/Vaginal	AB first week	AB type first week	Duration days AB first week	AB etter first week	PenicillinAndGentamicin
1	Sample1	Yes	26	Sectio	Yes	Penicillin eller ampicillin + aminoglykosid	2	No	No
2	Sample2	Yes	24	Sectio	Yes	Penicillin eller ampicillin + aminoglykosid	6	Yes	Yes
3	Sample3	Yes	27	Sectio	Yes	Penicillin eller ampicillin + aminoglykosid	3	Yes	No
4	Sample4	Yes	27	VF	Yes	Penicillin eller ampicillin + aminoglykosid	7	Yes	Yes
5	Sample5	No	29	Sectio	Yes	Penicillin eller ampicillin + aminoglykosid	5	No	No
6	Sample6	No	32	Sectio	Yes	Penicillin eller ampicillin + aminoglykosid	3	No	No
7	Sample7	No	32	Sectio	Yes	Penicillin eller ampicillin + aminoglykosid	4	No	No
8	Sample8	No	30	Sectio	Yes	Penicillin eller ampicillin + aminoglykosid	7	No	No

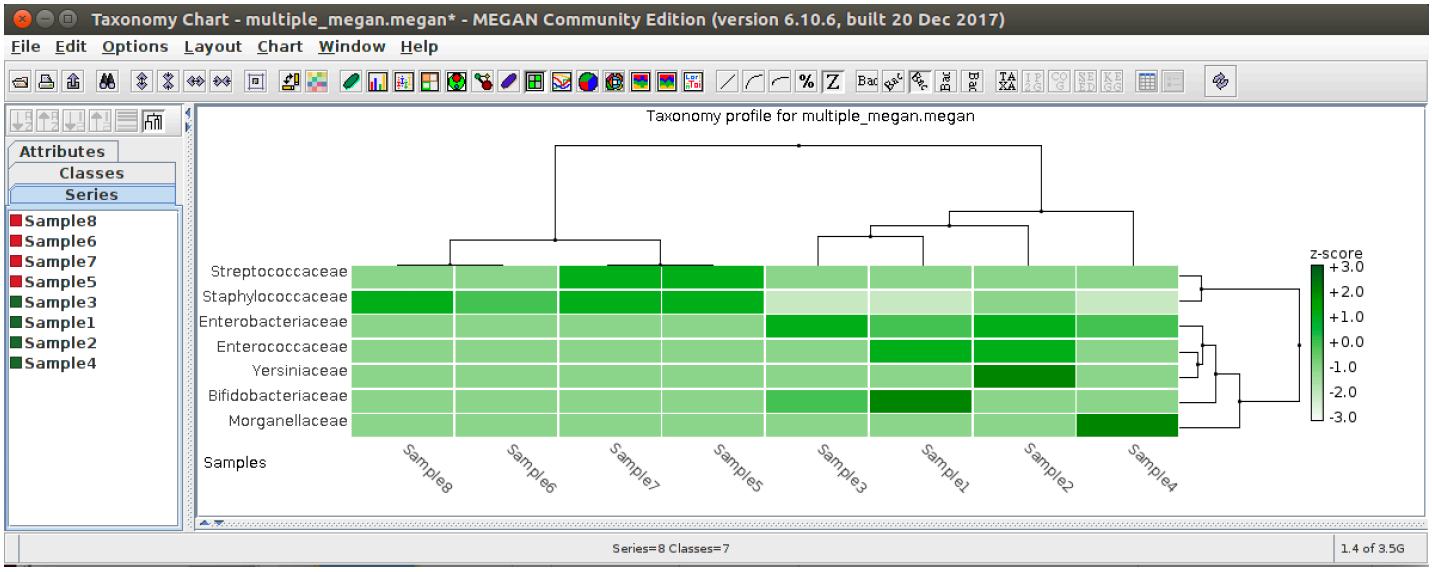
---

Return to the bar charts window. The colours for each sample should now have changed to red and green for the Untreated and Treated samples, respectively.

I] In the "**Series**" tab, sort the list of sample (entries) alphabetically by clicking the sort symbol in the upper left corner of the bar charts window.

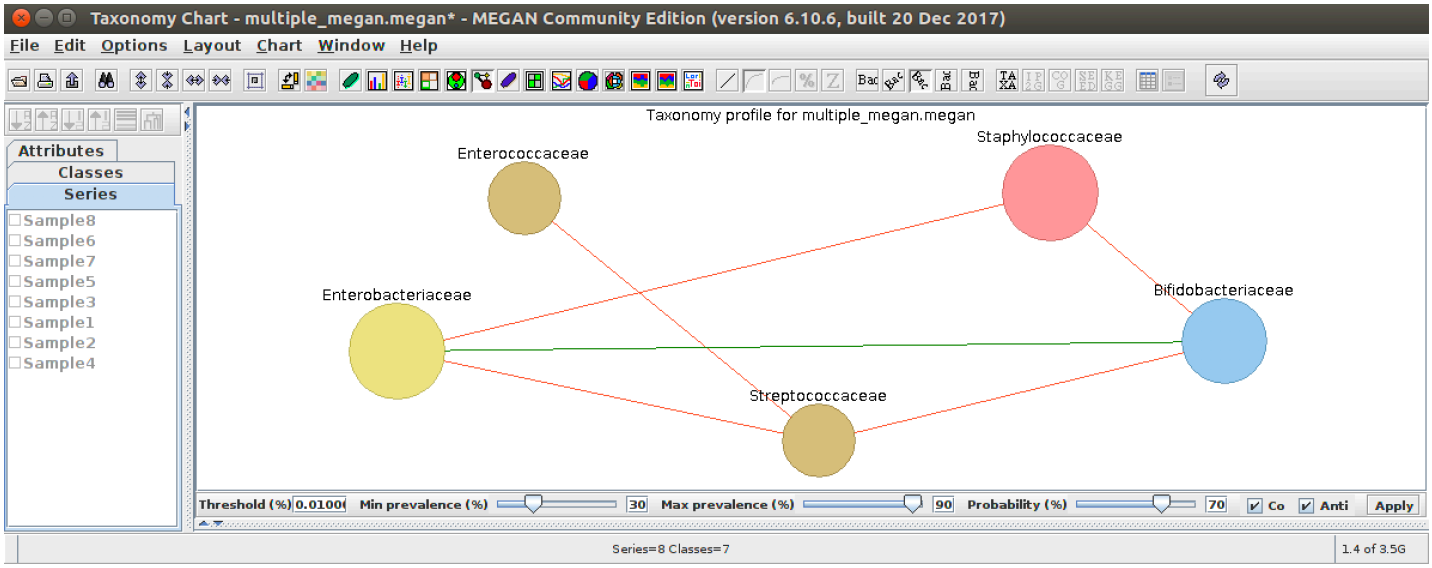
II] Click on the "**Transpose**" button so that the data are viewed by sample rather by taxon.

III] Change to heatmap view and click on the "**Turn clustering on or off**" (tree symbol upper left corner - ask if you don't find it).



In this chart you can see the relative abundance of families between samples, and also how the different samples cluster together.

III] Change to the "**Co occurence**" plot and adjust the min prevalence bar to 30% and press "**Apply**".



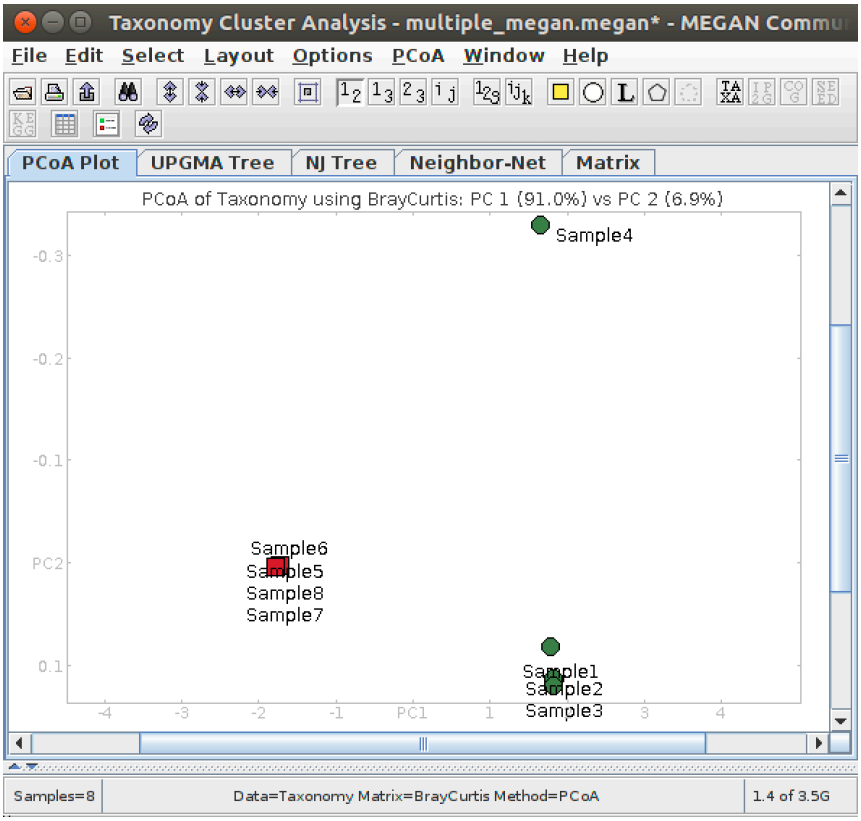
This plot provide a graphic visualization of potential relationships between the occurrence of different taxonomic groups in the different samples. For example, in your plot you will see that the family *Enterococcaceae* and *Streptococcaceae* appear together (co-occurre) in minimum 30% of the samples, while *Enterococcaceae* does not co-occurre with any other family.

Return to the tree view window.

I] In the "**Window**" menu, select "**Cluster analysis**".

A new window showing a PCoA (principle coordinates analysis) plot will appear. A PCoA plot is a method to explore and to visualize similarities or dissimilarities of data. The plot is calculated based on a cluster

analysis of the taxonomic profiles of each sample.



💡 **Note:** The Cluster Analysis viewer provides methods for comparing multiple samples. The Cluster Analysis viewer allows one to compute a distance matrix on the set of samples, based on for example their taxonomic profiles. The calculated distances are displayed as a PCoA plot, a hierarchical clustering (UPGMA tree), an unrooted tree (Neighbor-Joining tree) or an unrooted split network (Neighbor-net). You can find more information about this in the manual.

II] Switch between the different way to view the result from the cluster analysis.

❓ Does the trees separate the samples into branches as expected from the metadata?

▶ **Solution** - Click to expand

Alpha diversity is describing species richness (number of taxa) within a single microbial ecosystem. **How many different microbial species could be detected in one sample?**

Beta diversity describes the diversity in microbial community between different environments (difference in taxonomic abundance profiles from different samples). **How different is the microbial composition in one environment compared to another?**. Bray-curtis is a statistic method used to quantify the Beta diversity of different samples, by calculating the compositional dissimilarity between two different sites based on counts at each site. The higher Bray-curtis distance, the more diverse are the samples.

III] Click on the "**Matrix**" tab.

❓ What are the most dissimilar samples realtive to Sample 1?

▶ **Solution** - Click to expand

IV] Go back to viewing the PCoA plot. Press the "**PCoA**" menu on the top of the window, select "**Show Biplot**".

❓ What do you think these arrows and taxa represents?

▶ **Solution** - Click to expand

❓ Is there anything in the metadata that may give an indication of why Sample 4 does not cluster tightly with the other probiotic treated samples?

▶ **Solution** - Click to expand

OPTIONAL I: If you have time at the end, go back to the tree view window and try to calculate the core biome of the 8 samples. "**s**" in this setting means which taxa all of them share?

OPTIONAL II: Make a word cloud of the taxonomic profiles.

Progress tracker

Complete

That was the end of the this practical - Good job 👍