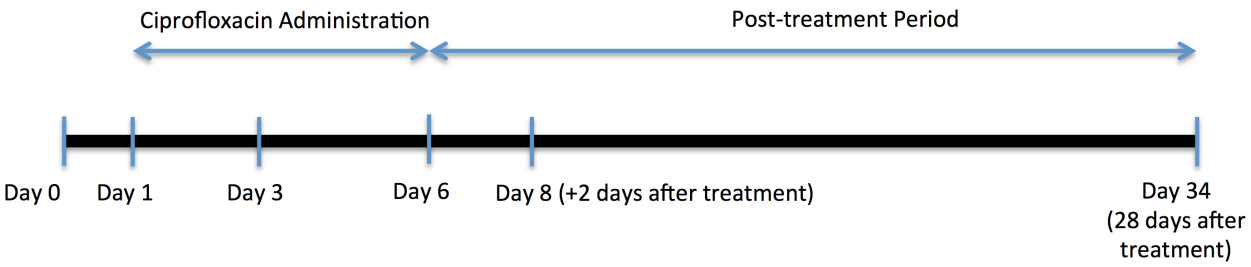
Daniel H. Huson 2-Aug- 2021 University of Tübingen

DIAMOND+MEGAN TUTORIAL

PART 1: SHORT READS

We will look at Illumina sequences from some stool samples from an antibiotics treatment pilot study (Willmann et al. Antibiotic Selection Pressure Determination through Sequence-Based Metagenomics. Antimicrob Agents Chemother. 2015 59(12):7335-45. doi: 10.1128/AAC.01504-15.)

Two volunteers, Alice and Bob:



- 2 x 6 stool samples
- Shotgun sequencing
- ~60 million reads per sample (101 bp per read)
- ~800 million reads in total

DATA

In this tutorial: 1 million reads per sample:

```
Alice00-1mio.fq.gz Alice08-1mio.fq.gz Bob03-1mio.fq.gz

Alice01-1mio.fq.gz Alice34-1mio.fq.gz Bob06-1mio.fq.gz

Alice03-1mio.fq.gz Bob00-1mio.fq.gz   Bob08-1mio.fq.gz

Alice06-1mio.fq.gz Bob01-1mio.fq.gz   Bob34-1mio.fq.gz
```

PREPROCESSING: DIAMOND ALIGNMENT AGAINST NCBI-NR:

We align all reads against the NCBI-nr protein database:

```
diamond blastx -d nr -q Alice00-1mio.fq.gz -o Alice00-1mio.daa -f 100
```

```
diamond blastx -d nr -q Alice01-1mio.fq.gz -o Alice01-1mio.daa -f 100
```

…

PREPROCESSING: MEGANIZATION

We run the meganizer tool to index all reads and alignments, and to bin the reads to taxonomic and functional classes:

```
MEGAN/tools/daa-meganizer -i Alice00-1mio.daa -mdb megan-map-Jan2021.db
```

```
MEGAN/tools/daa-meganizer -i Alice01-1mio.daa -mdb megan-map-Jan2021.db
```

…

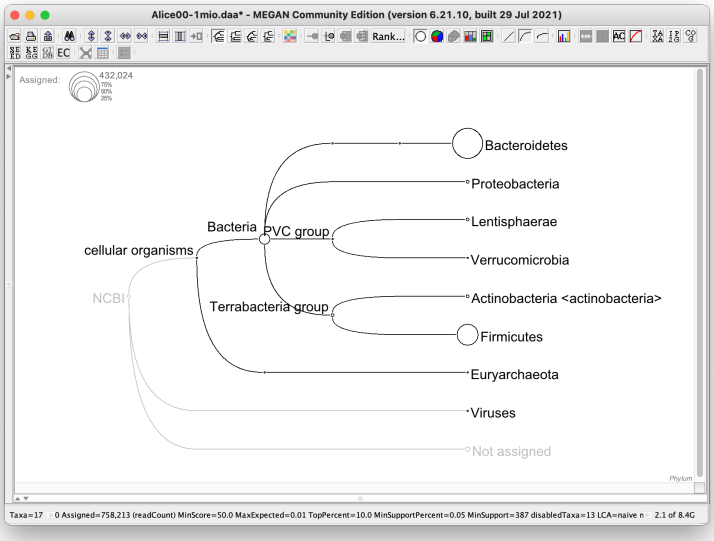
INTERACTIVE ANALYSIS USING MEGAN

We can open the meganized DAA files in MEGAN to perform interactive exploration and analysis.

BASIC TAXONOMIC ANALYSIS

Use the File->Open menu item to open the file `Alice00-1mio.daa`

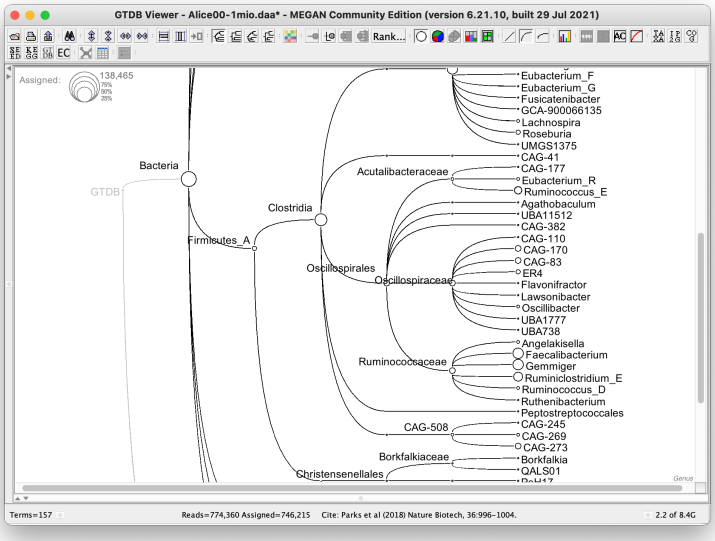
The file will open in the main viewer, which displays the binning of reads to the NCBI-taxonomy. You can interactively collapse and expand nodes in the tree to show more or fewer details of the taxonomy.



Questions A:

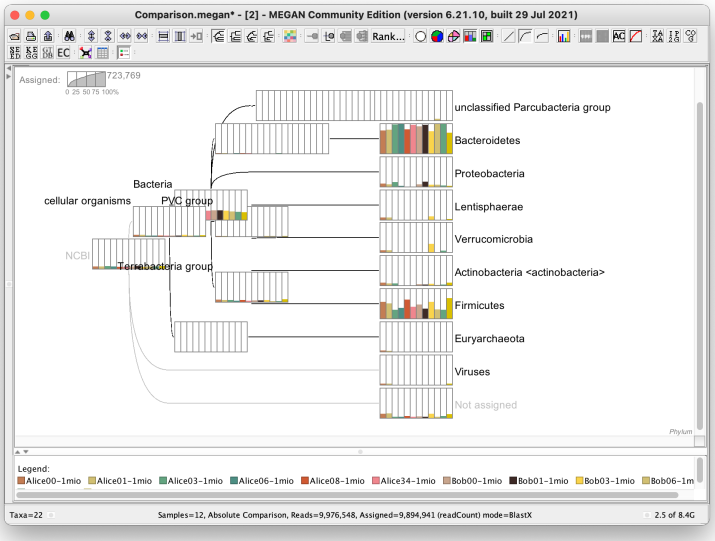
- Which phyla obtain the highest reads counts?
- Which species obtain the highest read counts?
- If you select a node in the viewer, you will see two numbers: Assigned and Summed, what do these mean?

Open the GTDB viewer. This shows taxonomic binning to bacteria and archaea, based on the GTDB taxonomy. Answer the same questions using this viewer.

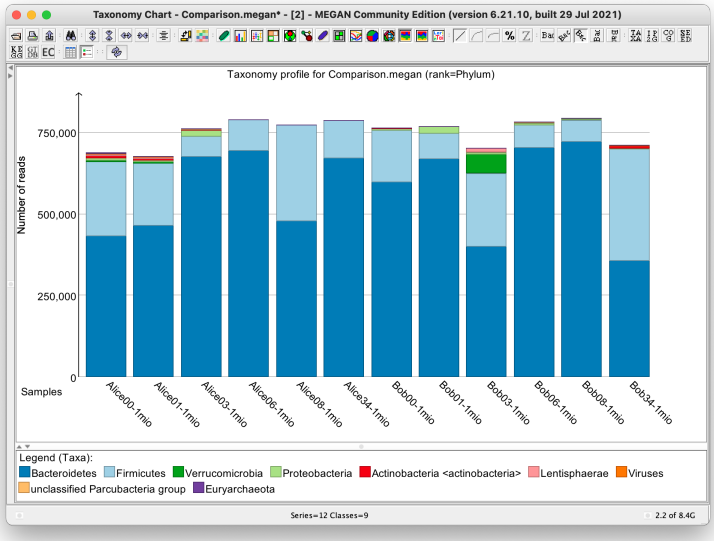


COMPARATIVE TAXONOMIC ANALYSIS

Use the File->Compare menu item to open all 12 files together in a *comparative document.*



Collapse the taxonomy view at Phylum rank. Open a stacked bar chart to compare the phylum-level counts for among all samples.

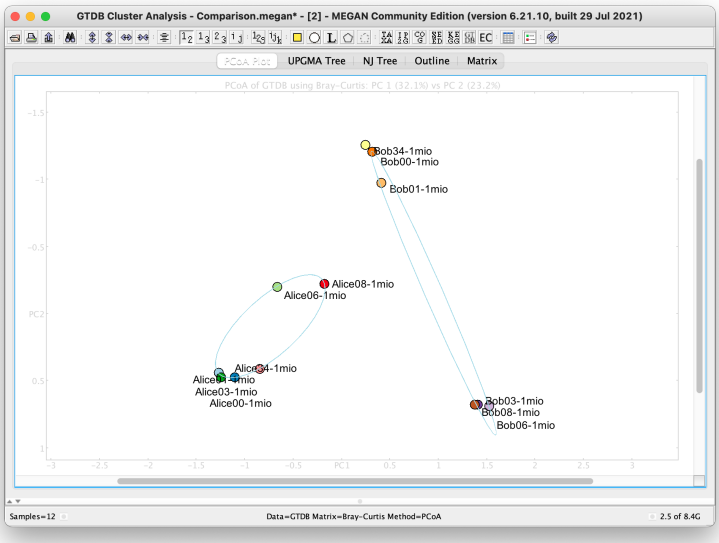


Questions B:

- Do all twelve samples have the same ranking of Phyla by read count?
- Is this also reflected in the GTDB-based analysis?
- Are the NCBI and GTDB binnings compatible with each other?
- Explore the Options-> Project Assignments to Rank… option, what problem does it address?

PCoA analysis:

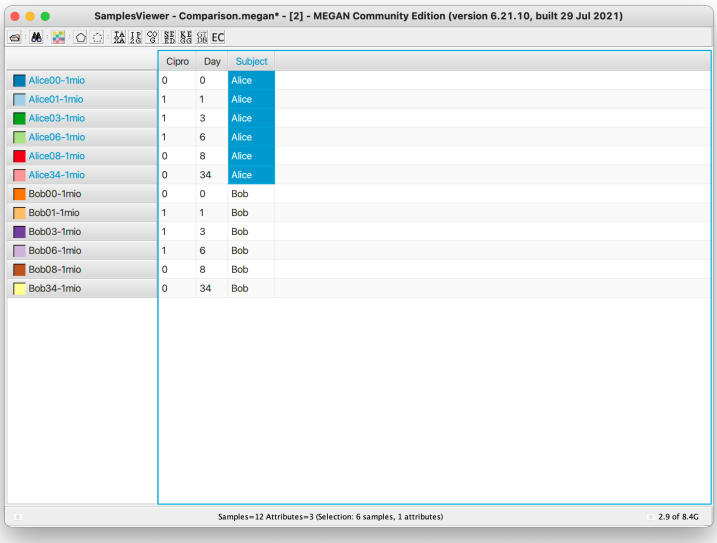
Uncollapse the GTDB viewer to the species level and then compute a PCoA plot.

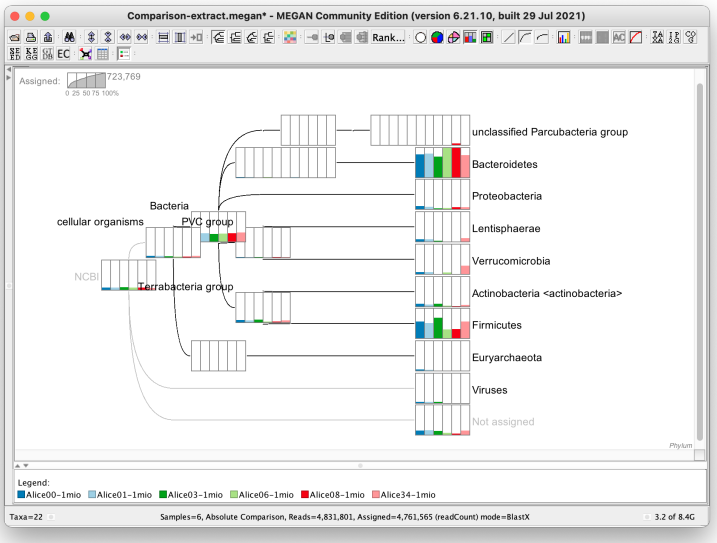


Questions C:

- What does the PCoA plot tell us about the changes in the gut microbiome of Alice (or Bob) as they go through the study?
- Use the “group by attribute” menu item to highlight the trajectories of Alice and Bob.

The comparison document that we have created contains six Alice samples and six Bob samples. Open the Samples Viewer, select the six Alice samples and use the Samples->Extract Samples… menu item to extract the six Alice samples into a new comparison document called Comparison-Alice.megan.



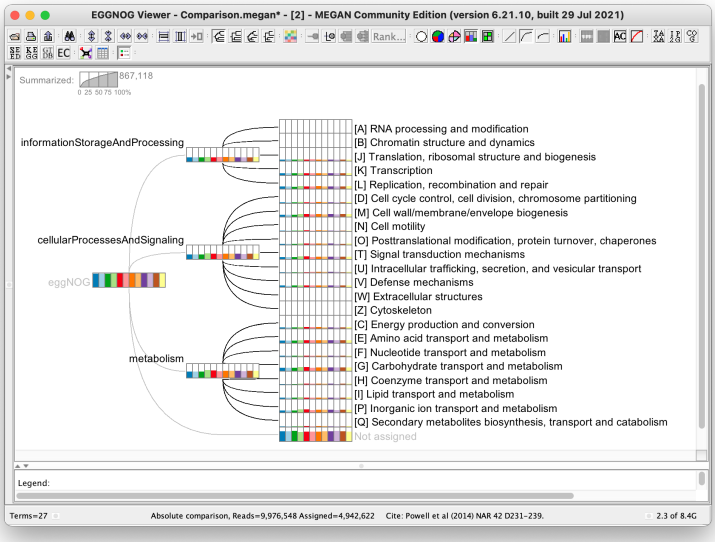


Questions D:

- Which of the two taxonomy viewers, the NCBI viewer or the GTDB viewer, displays a better circular trajectory of the six Alice samples in a PCoA plot based on the species level?

BASIC FUNCTIONAL ANALYSIS

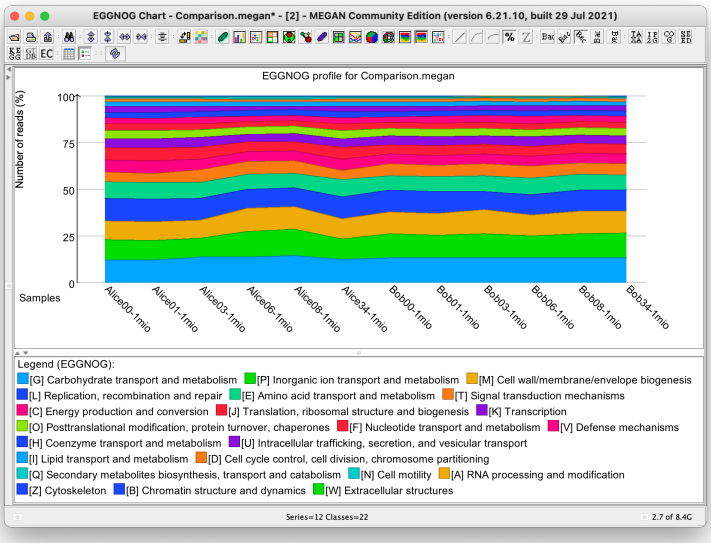
The COG/eggNOG classification uses letters A-Z to classify different types of gene functions. It is well-known that the reads counts for high-level functional categories are quite similar across samples.



Questions E:

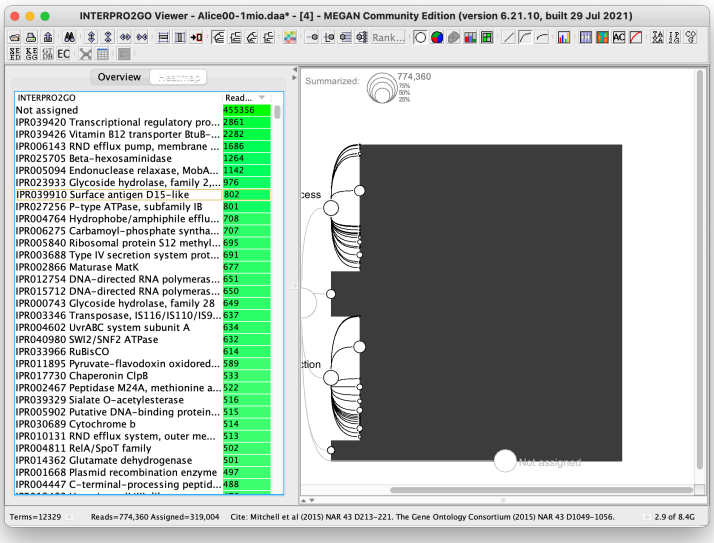
- Produce a stacked-line chart that shows how similar the assignments to the A-Z classes of COG/eggNOG are across the 12 samples.
- Look at the PCoA plot for this data for Alice samples on their own, do we still see the trajectory?

Open the file Alice00-1mio.daa. Explore the InterPro2GO functional classification. Uncollapse the classification down to the leaves. Then use the Heatmap in the left part of the window to list InterPro families by decreasing number of assigned reads.



Questions F:

- If you have downloaded the full short read files, then open Alice00-1mio.daa. For an InterPro family with a high assignment rate, do the reads align “uniformly” across the reference genes or do they stack at some domain? Use the Alignment Viewer to explore this question.



PART 2: LONG READS

We will look at a set of MinION sequences from an enrichment bioreactor targeting polyphosphate accumulating organisms (Arumugam, K., Bağcı, C., Bessarab, I. *et al.* Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data. *Microbiome* **7,** 61 (2019). https://doi.org/10.1186/s40168-019-0665-y).

The original dataset consists of 695,000 long reads with an average length of 9 kb. These were assembled into 1702 contigs of average length 61 kb.

PREPROCESSING: LONG-READ ASSEMBLY

```
unicycler -l reads.fq.gz -o reads_asm -t 16 --keep 3

gzip < reads_asm/assembly.fasta > assembly.fa.gz
```

PREPROCESSING: DIAMOND ALIGNMENT AGAINST NCBI-NR

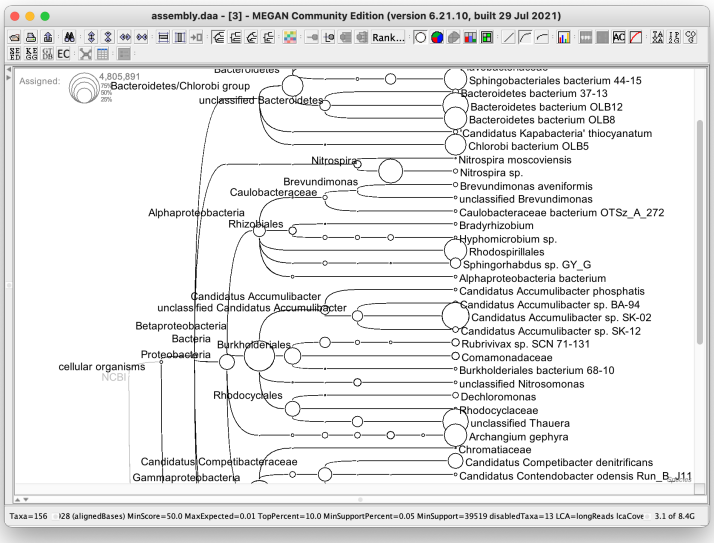
```
diamond blastx -d nr -q assembly.fa.gz -o assembly.daa -f 100
                                       -F 15 --range-culling --top 10
```

PREPROCESSING: MEGANIZATION

```
daa-meganizer -i assembly.daa -mdb megan-map-Jan2021.db -l
```

BASIC TAXONOMIC ANALYSIS

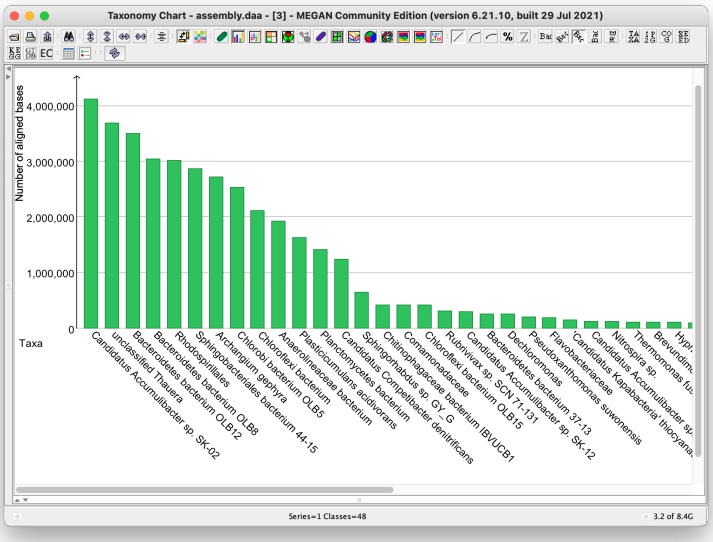
Open the file called assembly.daa.



In long read analysis, by default, MEGAN reports the number of “aligned bases” assigned to a bin rather than the number of reads assigned to a bin.

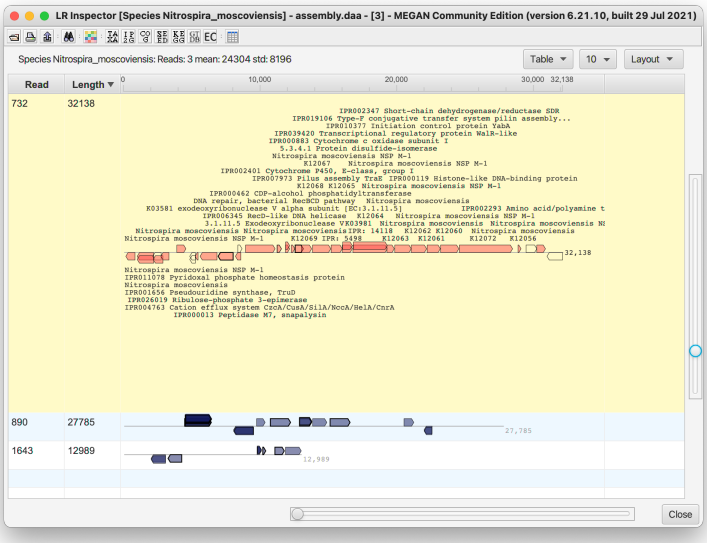
Questions G:

- For which organisms does MEGAN report genome-size numbers of aligned reads?



BASIC FUNCTIONAL ANNOTATION

While MEGAN bins long reads and contigs to functional nodes, this is not particularly useful. A more useful analysis is to show functional and taxonomic assignments along the reads and contigs. To explore this, open the long-reads inspector on one of the taxonomic nodes, namely *Nitrospira moscoviensis.* Use the Layout button to add annotations and use the button and right sliders to scale the representation.



Questions H:

- Do the genes along the contigs show a consistent taxonomic assignment?
- Which genes are present on the contigs?

The annotations calculated by MEGAN can be exported in GFF3 format.

Long reads and, to a lesser extent, long-read assemblies, contain erroneous insertions and deletions. These cause problems for tools that use translated-alignment (rather than frame-shift aware alignment), such as CheckM or Prokka. To significantly improve the performance of such programs on long reads, MEGAN can perform frame-shift error correction on sequences that can then be exported.