Metagenomics Practical Session

Alexey Larionov – December 2024

Contents

There are several toolsets for metagenomics bioinformatics analysis, including QIIME2, Mothur, UCLUST-family (including USEARCH, UBLAST etc) and a diverse collection of R & Python packages. Our practical session will focus on QIIME2 because it is currently the most popular integral toolset in metagenomics. It includes a comprehensive collection of tools, it is well-documented, it is being actively developed, and it has a large community of users.

Given the timeframe allocated for metagenomics within the course, you are not expected to write metagenomics pipelines yourself. Instead, you will be given a pipeline; your task will be to understand and run this pipeline. During the practical you should review the given scripts, modify them where necessary so that they run on your account on Crescent, execute the scripts, and use their output to generate tables and figures similar to those shown in the handouts.

You will need at least 2GB free disk space available for you on Crescent2 for this practical (you may check the available disk space using ***myquota*** command)

Computational environment

It is expected that during the previous modules you have already learned how to use the Crescent2 HPC and MobaXterm. The Crescent2 HPC already has modules installed for FastQC, MultiQC and QIIME2, which are required for this practical session. However, you still will need to download and configure SRA Tools. Even so the main steps of analysis are performed on Crescent2 cluster, the final (optional) R script can be run on your laptop. Again, it is expected that you have already used R and RStudio during the previous practical sessions.

However, before installing SRA tools, you should think about organising your workspace.

Organising your workspace

During metagenomics practical session, you will use tens of source data files, multiple scripts, and generate tens of output files, logs and plots. It is therefore essential to properly organise your working space to allow successful analysis. You should start by making the ***main project folder*** that will contain ***several sub-folders*** for all your project-related files.

Where should you put the main project folder? I suggest putting it into your home folder.

How to name the working folder? I suggest "metagenomics":

```
…/<user>/metagenomics
```

After creating the main project folder, make five sub-folders in it: for data, tools, scripts, results, resources:

Subfolder name	**Purpose**
data	Will be used to download the source data
scripts	Will be used to keep data analysis scripts and logs
tools	Will be used for installing some tools (e.g. SRA-tools)
results	Will be used to save results
resources	Will be used to keep resources (e.g. the required taxonomy classifier that will be downloaded from QIIME2 data-resources web page)

You may organise your workspace in a different manner (e.g. during the epigenetic session you had slightly different set of folders). However, you MUST organise your working space before you start installing tools, downloading data and resources or running the scripts.

Importantly: do NOT use spaces or special or non-Latin characters in your folders and file names because this will cause errors during the pipeline! It is a general good practice: not to use spaces in files and folder names (you may use underscores instead).

Aim of analysis and Source data

During this practical you will analyse real metagenomics research data generated by a large international project called Nutrient Network (https://nutnet.org/). The project collected many soil samples from different types of grassland from around the world. The goal of the project was to study the effect of Nitrogen and Phosphorus supplementation on soil microbiota. The project led to multiple publications, including the flagship paper in PNAS: https://doi.org/10.1073/pnas.1508382112 (Leff *et al* 2015, PDF provided).

The aim of our analysis will be to compare control soil samples from 3 different types of grassland collected in 3 different countries:

Group code	Num of samples	Type of grassland	Country
frue_ch	6	Pasture	Switzerland
mtca_au	6	Savanna	Australasia
ukul_za	6	Mesic grassland	South Africa

You may find more details about the analysed soil samples in Supplementary Tables 2 and 3 from the above paper (see pages 6 and 7 of the Supplementary Data, PDF also provided):

https://www.pnas.org/doi/suppl/10.1073/pnas.1508382112/suppl_file/pnas.201508382SI.pdf

The sequencing was done following the Earth Microbiome Project protocol, using 515f/806r primers targeting V4 16S region (https://earthmicrobiome.org/protocols-and-standards/16s/).

The raw sequencing results (paired-end FASTQ files) were deposited to NCBI's Sequence Reads Archive (CRA) for public use. You are provided with the list of SRA IDs for the source samples:

frue_ch: SRR1770746, SRR1770744, SRR1770743, SRR1770742, SRR1770738, SRR1770737

mtca_au: SRR1770762, SRR1770760, SRR1770758, SRR1770756, SRR1770753, SRR1770752

ukul_za: SRR1770778, SRR1770775, SRR1770773, SRR1770770, SRR1770768, SRR1770766

These IDs, along with the annotations, are available in file ***samples.txt*** provided with the scripts.

Installing SRA-tools

The Crescent2 cluster already has modules for QIIME2, FastQC and MultiQC. However, you still need to install and configure SRA-tools. This is a toolset for retrieving data from NCBI SRA.

Follow the separate worksheet for SRA-tools installation before continuing with this handout.

Ask help if you have problems with installing or configuring SRA-tools.

Pipeline summary

The pipeline includes a minimal set of steps that should be included in virtually any metagenomic analysis:

- Source data (FASTQ files) import, QC and pre-processing
- "Denoising" or "OTU-clustering", i.e., inferring the "features" = microbial sequences, which could be reported as ASVs or OTUs (our pipeline will use DADA2 to infer ASVs)
- Making the features table (the counts of features per sample)
- Phylogenetic tree (hierarchy and similarity of detected microbial sequences, our pipeline will build a *de-novo* phylogenetic tree)
- Rarefaction (aka normalisation by total features count per sample)
- Assessment of Alpha- and Beta- diversity
- Assessment of microbial taxonomy in the studied samples

Of course, there are many more metagenomics tools and tasks than included in this pipeline. Usually, the other types of analysis follow these initial steps. However, these other steps usually are study-specific and beyond the scope of this introductory course.

After installing the necessary tools and making the project folder and the subfolders, copy the provided scripts (and sample lists) into the “scripts” subfolder.

For clarity, all scripts are provided without the additional header and footer which are required for running scripts on Crescent2. These scripts should NOT be run directly in login node! You should execute scripts on a compute node using qsub. An example of the additional header and footer required for submitting batch jobs to Crescent2 is provided separately in the ***crescent2_batch_job.sh*** file.

Step 1: Downloading source data from SRA

It is assumed that you already created the folder’s structure as discussed in the ***Organising your Workspace*** section. It is also assumed that you have already installed and configured SRA-tools, as described in the separate handout.

If everything is prepared properly, it’s now time to run the scripts!

Review **s01_get_data_crescent.sh** script. Remember that you should add some code at the top and at the bottom of the script to submit it as a batch job to Crescent2 (see examples of the header and footer in **crescent2_batch_job.sh**). Also, remember about updating the path to ***your*** base folder. After preparing the script, submit it to the Crescent queue using a command like the following:

```
qsub s01_get_data_crescent.sh
```

NEVER run you scripts directly on the login node(s) of an HPC cluster! You should receive the Crescent2 job progress updates by e-mail, and you may check your jobs on cluster by using the **myjobsum** command. You may terminate your jobs with the **qdel** command.

Once your script has run you should get 32 FASTQ files in your data folder: 2 files (forward and reverse reads) for each of the 18 soil samples. The download should take **about 5 min**.

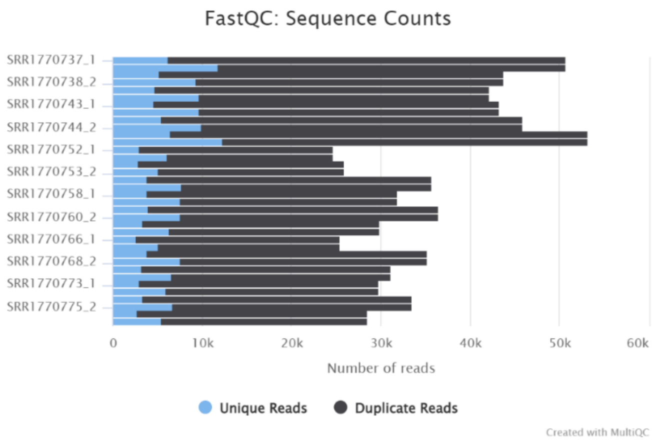
Step 2: FastQC and MultiQC

Review script **s02_qc.sh**. Modify script as required for execution on cluster, update the base folder name. Don’t forget to use FastQC and MultiQC modules.

Then run the script (submit it to the queue).

The script should complete in **about 5 minutes**.

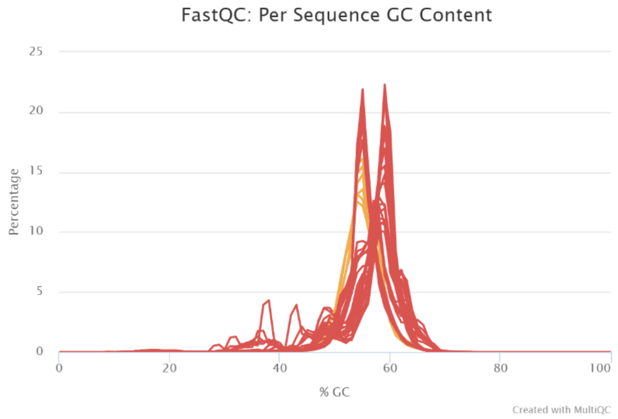
After the run you should see the FastQC and MultiQC output along with your FASTQ files in the **data folder**. Review the QC results (you should already be familiar with FastQC and MultiQC output after the previous modules). To view the HTML files directly on cluster you may use **RightClick - Open with default program** option. Note the features specific for 16S targeted amplicon sequencing. For instance, the MultiQC summary shown below suggests a very high proportion of duplicated reads:



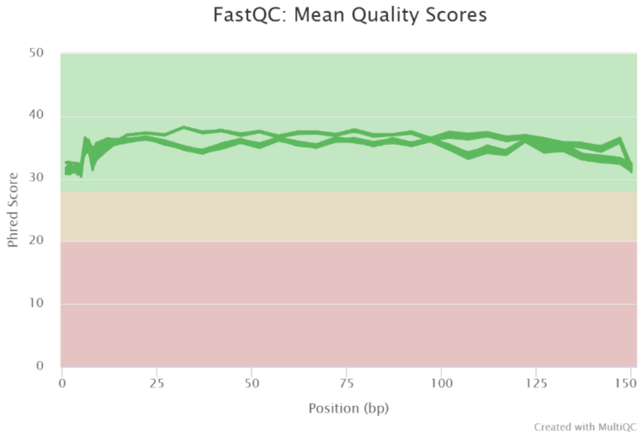
This is confirmed in other FastQC and MultiQC plots relevant to the duplication level. Would you expect high "duplication" in amplicon sequencing data? Why?

The high "duplication rate" was not because of some experimental problem (as you might think looking at the FastQC results). It's just an artefact of the FastQC methodology for duplication detection: because 16S amplicons from different bacterial species are genuinely very similar, they may be mis-interpreted by FastQC as "duplicates". Also, the multiple copies of identical 16S amplicons from multiple bacteria of the same species represent a true biological signal, not the "PCR duplicates".

The MultiQC summary for Per Sequence GC content shows a bimodal distribution:



Would you expect this in the paired-end short amplicon sequencing data? Why?



Is the overall bases quality good? Do you see any failed samples or outliers that should be excluded?

What was the encoding of the FASTQ files? (Hint: look in Basic statistics section in individual sample FastQC reports).

Select and save some plots that might be used in a short report to justify your conclusion about the quality of raw data (e.g. like the plot shown above).

If you are satisfied with the quality of raw data, continue to the next step: importing data into QIIME2 environment.

Step 3: Import data to QIIME2 format, QC and trimming

Importing source data into the QIIME2 file format is the first step in QIIME2 data analysis. Basic information about QIIME2 was provided during the lectures. A detailed tutorial for the latest version of QIIME2 is available here:

https://docs.qiime2.org/2024.10

Briefly, QIIME2 has two internal file types: ***artifacts (.qza)*** and ***visualisations (.qzv)***. Both file types are binary, and as such they can only be processed by QIIME2 itself.

Artifacts are used to store data (e.g. source data imported from FASTQ files, or a feature table, or a distance matrix calculated by QIIME2). Artifacts can be quite large, for instance they may include raw sequences for hundreds of samples.

Visualisations are small files which are intended to be viewed using the ***QIIME2 web viewer***: https://view.qiime2.org/ . The ***QIIME2 web viewer*** allows users to drag-and-drop **.qzv** files to its web-interface. It automatically recognises the file content and generates interactive plot(s) and table(s) corresponding to the file content, for instance:

- interactive bar-plots visualising samples taxonomy data
- box-plots visualising alpha-diversity within groups of samples
- PCOA plot for visualization beta-diversity
- etc.

It is possible to view the content of ***qzv*** files locally, without using the web-viewer. However, it would require additional configuration on HPC, so we will use the web-viewer only.

Along with their actual content all QIIME2 files also include so called "***provenance***" information (a detailed record about how each file had been generated). This information may be visualised using ***QIIME2 web viewer*** too.

QIIME is based on plugin architecture: it includes a small core functionality and multiple ***plugins*** for different analysis tasks. However, from a user's perspective the plugin interfaces are very similar to the core functions.

While QIIME2 provides a very comprehensive collection of tools and visualisations, it hasn't yet embraced everything in metagenomics. One of the very important features of QIIME2 is the ability to export data from its ***artifacts*** to the appropriate platform-independent formats. For instance, if an artefact contains a distance matrix, this matrix can be exported to a plain text file. Then this text can be read by R or Python for downstream analysis. Another example: as of Dec 2023, QIIME2 still didn't have a tool for visualizing phylogenetic trees. Crescent2 provides QIIME2 modules of 2022. So, in step 5 of this practical the tree data will be exported into Newick file format which can later be visualised by many third-party tools. Can you check whether the latest QIIME2 version already includes a tool for visualizing phylogenetic trees? (it's an optional task for section 5 of the practical).

Now, review and update script **s03_q2_import_and_trim.sh**. Don't forget to use the QIIME2 module (use **module spider qiime** to find the module name).

What three steps are included in this script?

How many ***artefacts*** and ***visualizations*** will be generated by this script?

To import data to QIIME2, this script uses files list in **source_files.txt** file.
Review this file and update the locations of your source files.

To run the script submit it to the queue using qsub.

The script should complete in **about 10 minutes**.

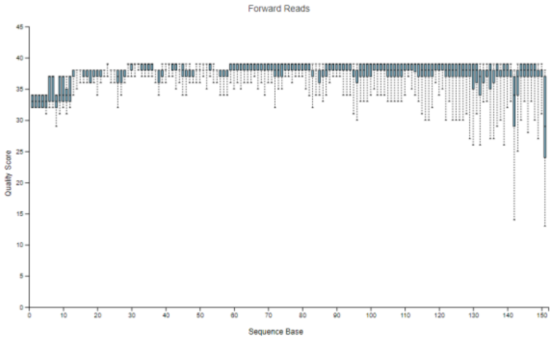
After the run, you should see the newly created ***artifact (.qza)*** and ***visualisation (.qzv)*** files in your **results folder**. Explore the visualisation(s) in the ***QIIME2 web viewer*** (https://view.qiime2.org/) . You need to copy the visualisation **.qzv** file to your PC first.

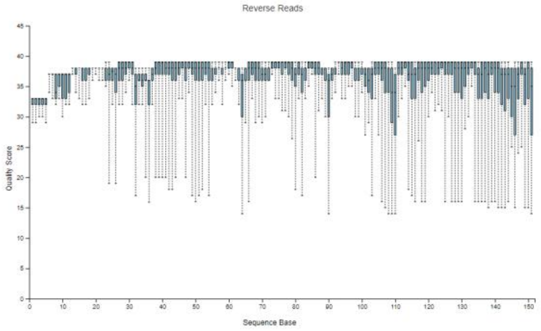
Explore the "Overview" and "Interactive Quality Plot" sections.

What are minimal and maximal numbers of reads per sample?

Unfortunately, QIIME2 does not allow you to download some of its interactive plots.

Instead, save screenshots of the Quality Plots for forward and reverse reads:





Do QIIME2 plots confirm the good quality of reads in the dataset?

Is there need in any further quality trimming in this dataset?

Step 4: Denoising with DADA2

Review script **s04_q2_denoise.sh**.

This is the longest step in this pipeline: it may take **15-20 min** when you use 12 threads. It implements DADA2 denoising and some additional steps (as you will see when you explore the results of this step).

How many ***artefacts*** and ***visualizations*** will be generated by this script?

Update and run the script (submit it to the queue as a batch job).

After script completion you should see the newly created ***artifacts*** and ***visualisations*** in your **results folder**. Explore the visualisations in the ***QIIME2 web viewer*** (https://view.qiime2.org/). You need to copy the **.qzv** files to your PC first.

On the next page you may see an example of the summary table provided in the **s04_stats_dada2.qzv** visualisation. It shows that in addition to the denoising (i.e. detecting ASVs) this step also performed filtering, merging the paired reads and removing chimeric reads. Can you explain what was done at each of these steps?

What is the total number of samples and features (=ASVs) in the feature table?
(hint: visualize **s04_table_dada2.qzv**).

Can you find how to see the actual sequences of ASVs?

Step 5: Phylogenetic tree of detected microbial sequences

During this step you will calculate how similar the detected ASVs between each other are. Importantly, you are not yet going to look for similarity using already known hierarchy of microbial taxa. The taxonomy assignment will be done only in the Step 10. The Phylogenetic analysis only looks at the distances (similarity/dissimilarity) between the ASVs sequences (*de-novo* phylogeny), independent of their taxonomies.

Of course, detecting similarity between sequences and building the Phylogenetic tree is based on the (multiple) alignments of the detected ASVs. Again, like in the previous scripts, QIIME2 provides a convenient wrapper that binds all the required steps together: the multiple alignment + distance calculations + clustering + preparing data for trees + additional technical steps (like masking). Again, a detailed discussion of each step that is selected by QIIME2 for Phylogenetic analysis is beyond the scope of this module. However, like for all the other QIIME2 tools, you may read the details in the web-tutorial (e.g. https://docs.qiime2.org/2022.11/tutorials/phylogeny for your version of QIIME2).

Phylogenetic trees are very important for many of the downstream steps of the metagenomics analysis. For instance, you will see that a Phylogenetic tree is required as input for Rarefaction analysis or for calculating Diversity metrics in Steps 6 and 7 of this pipeline. However, somehow QIIME2 v.2022.11. does not provide a visualisation for Phylogenetic trees. So, to plot the phylogenetic tree, you will export it into a platform-independent tree file format (called Newick), which will be used for visualization by third-party tools (I recommend using NCBI Tree Viewer).

Review script **s05_q2_phylogenetic_tree.sh**. What input does it take? What ***artifacts*** does it produce? Is there any ***visualisation*** produced at this step? Note the QIIME2 code used to export tree data.

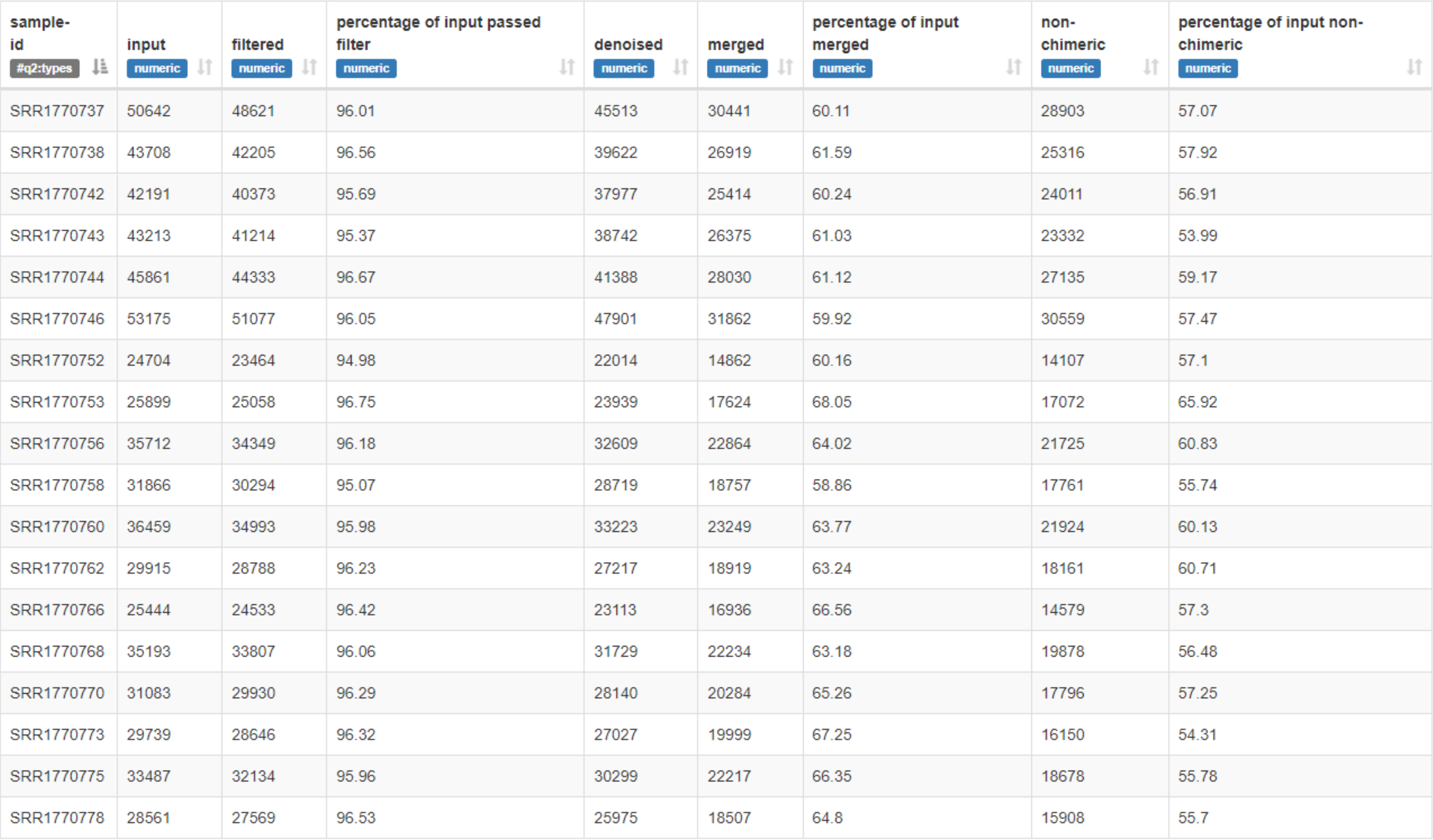
Update and run the script (submit it to the queue using qsub).

The script should complete in **about 10 minutes**.

After completion of the script, you should see the newly created ***artifacts*** in your **results folder**. You should also see a sub-folder in the **results folder**: it contains the tree data exported in the Newick file format. Visualize the tree exported in Newick file format using NCBI Tree Viewer:

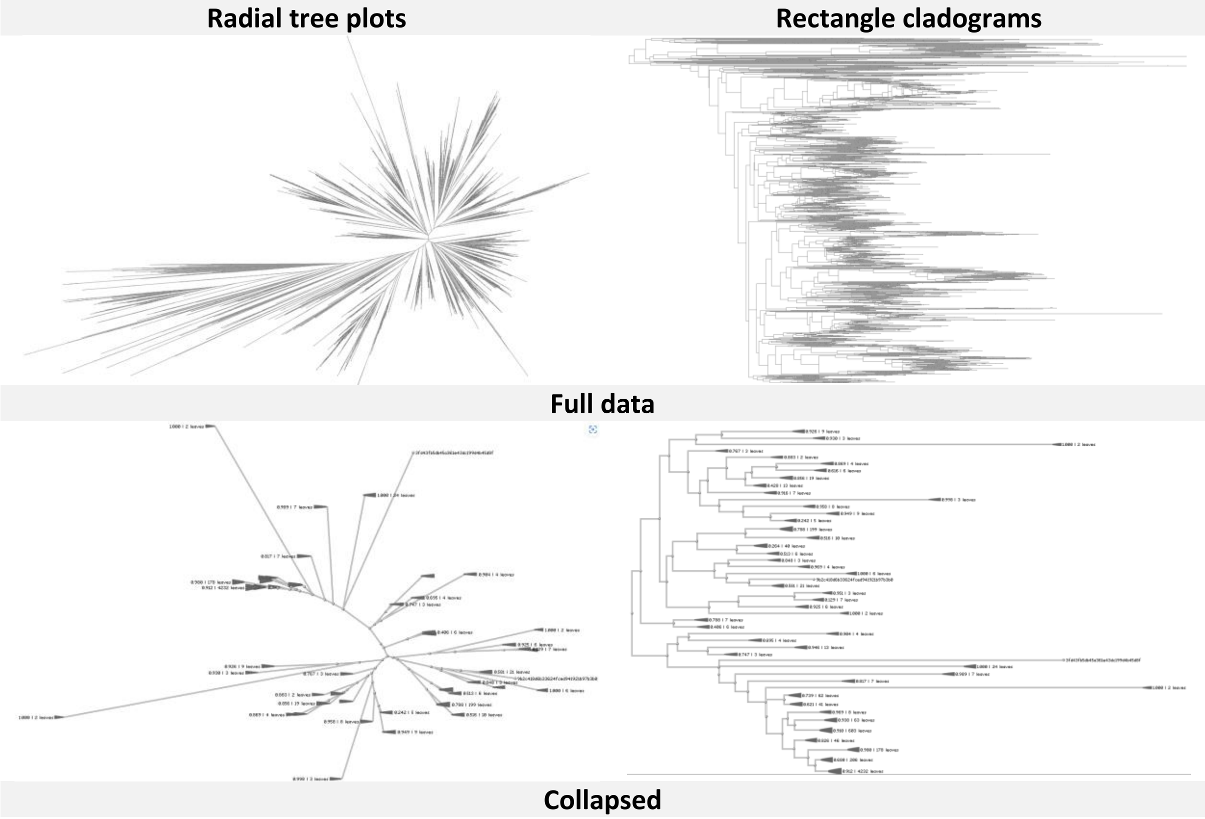
https://www.ncbi.nlm.nih.gov/projects/treeview/tv.html?appname=ncbi_tviewer&renderer=radial&openuploaddialog

Of course, you need to copy **tree.nwk** file to your PC first.



sample-id	input	filtered	percentage of input passed filter	denoised	merged	percentage of input merged	non-chimeric	percentage of input non-chimeric
#q2:types	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric
SRR1770737	50642	48621	96.01	45513	30441	60.11	28903	57.07
SRR1770738	43708	42205	96.56	39622	26919	61.59	25316	57.92
SRR1770742	42191	40373	95.69	37977	25414	60.24	24011	56.91
SRR1770743	43213	41214	95.37	38742	26375	61.03	23332	53.99
SRR1770744	45861	44333	96.67	41388	28030	61.12	27135	59.17
SRR1770746	53175	51077	96.05	47901	31862	59.92	30559	57.47
SRR1770752	24704	23464	94.98	22014	14862	60.16	14107	57.1
SRR1770753	25899	25058	96.75	23939	17624	68.05	17072	65.92
SRR1770756	35712	34349	96.18	32609	22864	64.02	21725	60.83
SRR1770758	31866	30294	95.07	28719	18757	58.86	17761	55.74
SRR1770760	36459	34993	95.98	33223	23249	63.77	21924	60.13
SRR1770762	29915	28788	96.23	27217	18919	63.24	18161	60.71
SRR1770766	25444	24533	96.42	23113	16936	66.56	14579	57.3
SRR1770768	35193	33807	96.06	31729	22234	63.18	19878	56.48
SRR1770770	31083	29930	96.29	28140	20284	65.26	17796	57.25
SRR1770773	29739	28646	96.32	27027	19999	67.25	16150	54.31
SRR1770775	33487	32134	95.96	30299	22217	66.35	18678	55.78
SRR1770778	28561	27569	96.53	25975	18507	64.8	15908	55.7

Explore different options of the viewer to generate different plot layouts for the tree data, e.g.:



Radial tree plots

Rectangle cladograms

Full data

Collapsed

Download or make a screenshot of a Phylogenetic tree plot.

Step 6: Rarefaction

If you reached this step before the end of Thursday then you work faster than expected: you may take a (short!) break 😊

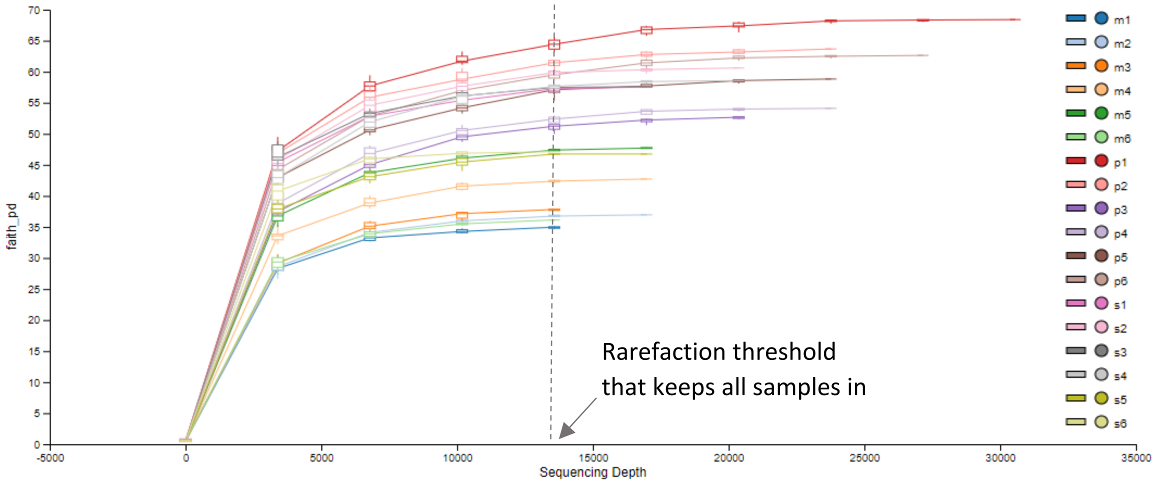
Rarefaction is a method used to normalize different samples by the number of detected features. It could be used, for instance, when comparing Diversity metrics between samples. There are pros- and contras- for using **Rarefaction** for normalization (as mentioned in the lecture). Importantly, rarefaction curves may also be used to evaluate whether the depth of sequencing was sufficient: the rarefaction curve for a sample should reach plateau, if the depth was sufficient for this sample (Why?).

Review, edit and run script **s06a_q2_rarefaction_plot.sh**. As usual, you will need to update the base folder, and make other edits required to submit the script as a batch job on Crescent2 cluster. The script should complete in about **3 minutes**.

Visualise **s06a_alpha_rarefaction.qzv** in in the ***QIIME2 web viewer*** (https://view.qiime2.org/) . You need to copy the file to your PC first.

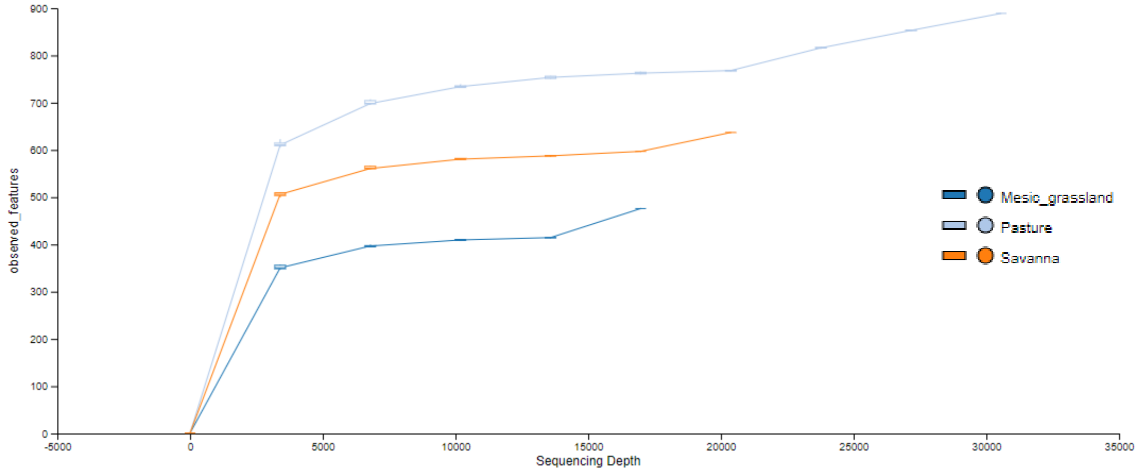
Explore different plotting options (see screenshots examples on the next page). Look at the **Provenance information** provided in this visualisation.

Have all samples plateaued? Take a screenshot of a representative **Rarefaction plot** (also, you may download the **CSV file** with data).



Rarefaction threshold that keeps all samples in

Rarefaction curves per sample (faith_pd)



Rarefaction curves per type of grassland (observed features)

Select the rarefaction threshold to keep all samples (use summary table from step 4, page 9).

After selecting the rarefaction threshold, apply it by running script **s06b_q2_apply_rarefaction.sh**. As usual, update the script to run it on the Crescent2 cluster. The script run should take < **1 minute**.

Step 7: Diversity metrics calculation

After we selected the rarefaction threshold, we may calculate diversity metrics to assess and compare **alpha- and beta- diversity** in our samples and sample groups.

Review script **s07_q2_calculate_diversity_metrics.sh** .

As earlier, QIIME2 wraps calculating multiple metrics into a single command. You will explore and plot some of these metrics later. Note that the syntax used in this script takes the rarefaction threshold (`--p-sampling-depth`) and the ***non-rarefied*** feature table as inputs. Also, the Phylogenetic tree is included into inputs to calculate the metrics involving phylogenetic information. Just for illustration, the script exports some of the calculated metrics to text file formats, so you may explore how the alpha- and beta- diversity metrics look numerically (or use in R or Python scripts).

Edit and run script **s07_q2_calculate_diversity_metrics.sh**. The script run should take about **10 minutes**. After the run you should see a new sub-folder in your **Results folder**. The new sub-folder will contain numerous **artefacts**, **visualizations** and further sub-folders with some distance metric data exported as text files.

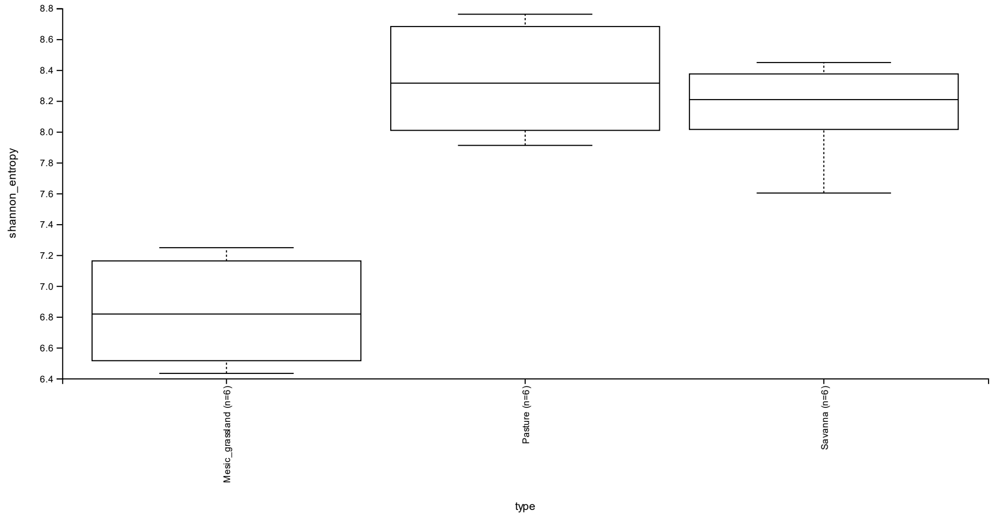
If you explore the text files, you may note that the **Alpha-diversity metrics** (e.g. Shannon or Faith-pd metrics) are represented as **vectors**: one number per sample. In contrast, the **Beta-diversity** (e.g. Unifrac or Jaccard distances) are represented by **distance matrices** containing pair-wise distances between each pair of samples.

Step 8: Make Alpha-diversity box-plots

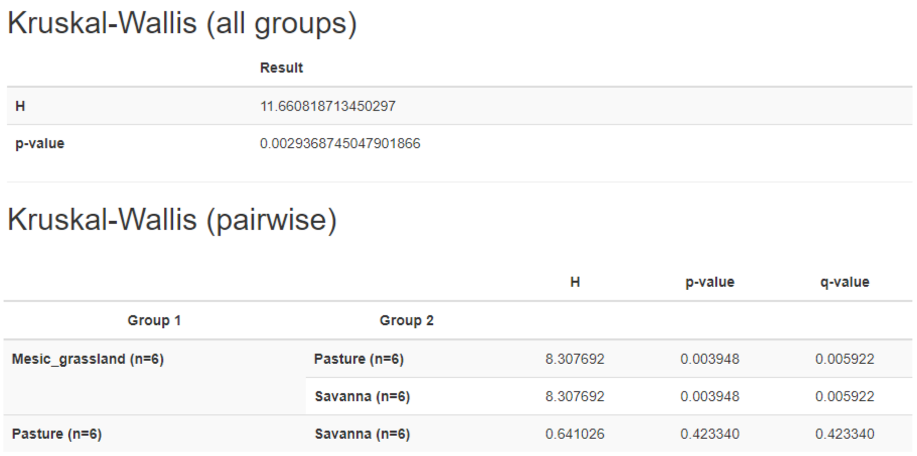
It is not very easy to evaluate vectors and matrices visually in their numerical form. Of course, we need to plot the results to evaluate them.

Review, edit and run script **s08_q2_alpha_diversity_box_plots.sh**.
The script should complete in **< 1 minute.**

After the run you should see some new ***visualisations*** in your **Results folder**. Exploring these ***visualisations*** in the ***QIIME2 web viewer*** (https://view.qiime2.org/), you may see plots like this:



It's much easier to comprehend than the numeric vectors or matrices! Also, QIIME2 applies appropriate statistical methods to compare diversity between the studied groups:



Kruskal-Wallis (all groups)

	Result
H	11.660818713450297
p-value	0.0029368745047901866

Kruskal-Wallis (pairwise)

		H	p-value	q-value
Group 1	**Group 2**			
Mesic_grassland (n=6)	**Pasture (n=6)**	8.307692	0.003948	0.005922
	Savanna (n=6)	8.307692	0.003948	0.005922
Pasture (n=6)	**Savanna (n=6)**	0.641026	0.423340	0.423340

Is Alpha-diversity different between the studied groups? What group is significantly different from the others? Are the results similar for the different Alpha-diversity metrics? Download SVG plot(s) for Alpha-diversity and the text file(s) with the statistical analysis results from the ***QIIME2 web viewer***.

Step 9: Make PCOA plot for Beta-diversity

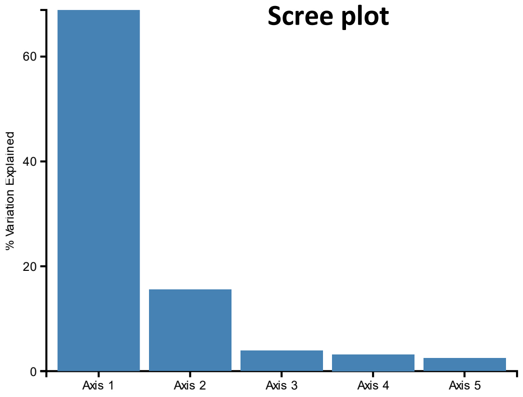
In the previous step you evaluated **Alpha-diversity**, visualising the diversity of features detected **within each individual sample**. In contrast, the **Beta-diversity** compares multiple **samples to each other**: how strong are the differences between the samples? are there groups of similar samples in the dataset? Arguably, the Beta-diversity assessment in metagenomics may resemble such techniques as Hierarchical Clustering or PCA in transcriptomics or metabolomics (which you already studied during the previous modules). However, because the metagenomic data are sparce and compositional, it requires different statistical techniques to produce similar visualisations. Thus, instead of applying Principal Component Analysis (PCA) for projecting samples on 2D/3D space, metagenomics uses Principal Coordinates Analysis (PCoA) for the same purpose. While the technical details of implementation are different, the conceptual assessment of the plots is similar. In metagenomics (and ecology) this sort of plots is often called “ordination”.

Review, edit and run script **s09_q2_beta_diversity_pcoa_plot.sh**. The script should complete in **about 5 minutes.**

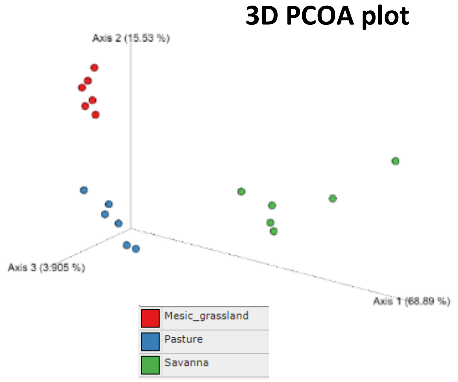
After the run you should see some new ***visualisations*** in your **Results folder**.

Explore these ***visualisations*** in the ***QIIME2 web viewer*** (https://view.qiime2.org/): use the “Color” tab to colour samples according to desired annotation; use the “Axes” tab to change the black background to white; use mouse left-press+ movement to explore 3D view; use right-click to download the PCOA plot; use the “Axes” tab to download Scree plot.

You may obtain plots like this:



Scree plot



3D PCOA plot

Are the samples from different groups similar? What is shown by the Scree plot? (some of you may still remember it from the Statistics module :) Are the results similar for different distance metrics?

Step 10: Taxonomy bar-plots

Now, when we know how different the microbiomes are between the studied types of grassland (or how different they are between the different studied sites :) one minor question remains unanswered: what are the actual microbes present in the studied samples?

To assign a taxonomy to each detected ASV, we need to compare the ASV sequences detected in our dataset with already known microbial sequences from the microbial taxonomy databases.

Some popular taxonomy-aware bioinformatics databases of 16S microbial sequences include

- RDP: https://www.lcsciences.com/documents/sample_data/16S_sequencing/src/html/top1.html
- SILVA: https://www.arb-silva.de/
- GreenGenes2 https://www.nature.com/articles/s41587-023-01845-1

Most of all, of course, the assigned taxonomy will depend on the selected database. GreenGene2 is the currently preferred one. For each database, there should be an algorithm to compare our ASVs with the sequences present in the database. It is likely that for many ASVs there may be no ***exact*** matches to any known microbial sequence in the reference database. In this case, the algorithm cannot assign an exact microbial ***species*** to our sequence. However, even if there is no ***exact*** match, the classification algorithm may still find similarity to some higher level of taxonomy, such as ***genus*** or even ***phyla***. Also, the classification algorithms (also called “classifiers”) allocate a certain probability to the assigned taxonomy.

Again, QIIME2 can do it all for us in a single command: see script **s10_q2_taxonomy_barplot.sh** .

There are many complex steps and resources behind the apparent simplicity of this QIIME2 implementation. The most important resource required for classification, of course, is the appropriately trained classifier.

QIIME2 provides tools to train bespoke classifiers starting directly from the microbial taxonomy databases (e.g. see https://docs.qiime2.org/2022.11/tutorials/feature-classifier). However, QIIME2 also provides pre-trained classifiers for several typical 16 targeting designs. Our source data were generated using 515F/806R primers targeting V4 region of 16S gene. For quite a while, this was one of the most commonly used designs of 16S metagenomic experiments (e.g. it was used by the Earth Microbiome project). So, the respective pre-trained classifier can be downloaded from the QIIME2 resources at https://docs.qiime2.org/2022.11/data-resources .

Before running **s10_q2_taxonomy_barplot.sh** script you should download this classifier to your **Resources folder**, as shown below. Because this download requires virtually no resources, you may execute the above command directly on the login node.

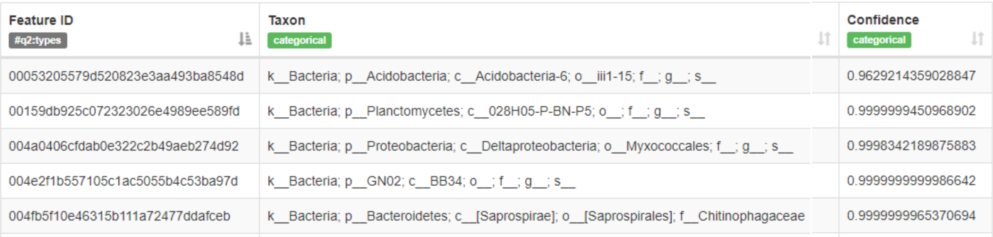
```
cd "${resources_folder}"

wget https://data.qiime2.org/2022.8/common/gg-13-8-99-515-806-nb-classifier.qza
```

The classifier is ~27MB, and it can be used with many versions of QIIME2 (including the versions installed on the Crescent2 cluster). Please note that we use this classifier here for teaching purposes. For real research in 2024, you would use pre-trained ***GreenGene2*** classifier (or even train a bespoke new classifier reflecting your PCR design and using the latest version of the taxonomy databases).

Now, edit and run **s10_q2_taxonomy_barplot.sh** script. The script should complete within **about 10 minutes**. It should add one ***artefact*** and two ***visualizations*** to your Results folder. Explore the produced ***visualisations*** in the ***QIIME2 web viewer*** (https://view.qiime2.org/):

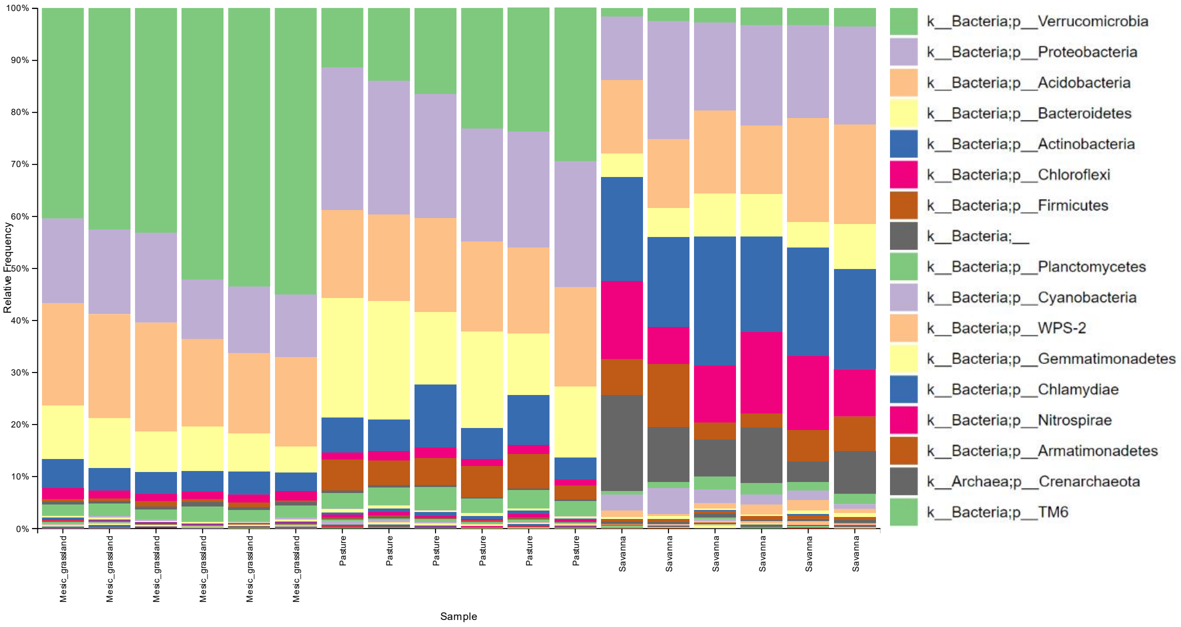
The **s10_taxonomy.qzv** visualization file contains information about taxonomic assignment of each ASV detected in the samples:



Feature ID #q2:types	Taxon categorical	Confidence categorical
00053205579d520823e3aa493ba8548d	k__Bacteria; p__Acidobacteria; c__Acidobacteria-6; o__iii1-15; f__; g__; s__	0.9629214359028847
00159db925c072323026e4989ee589fd	k__Bacteria; p__Planctomycetes; c__028H05-P-BN-P5; o__; f__; g__; s__	0.9999999450968902
004a0406cfdab0e322c2b49aeb274d92	k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__Myxococcales; f__; g__; s__	0.9998342189875883
004e2f1b557105c1ac5055b4c53ba97d	k__Bacteria; p__GN02; c__BB34; o__; f__; g__; s__	0.9999999999986642
004fb5f10e46315b111a72477ddafceb	k__Bacteria; p__Bacteroidetes; c__[Saprospirae]; o__[Saprospirales]; f__Chitinophagaceae	0.9999999965370694

Scroll down and see whether any ASV was identified to Genus or Species level?

The **s10_taxa_bar_plot.qzv** file summarises taxonomic groups per sample into bar-plots. Explore the plotting options for this type of data in ***QIIME2 web viewer***. Generate a plot similar to one below:



Is there clear visual distinction in taxonomies between the studied groups? What type of the grassland contains the higher proportion of Actinobacteria? What is the dominant fila in Mesic grassland? Download the plot(s) and CSV text file(s) from ***QIIME2 web viewer***, which could be used to describe and illustrate taxonomic composition of the studied soil samples.

Step 11: From QIIME2 to R (optional)

The visual assessment of the PCoA plot or the taxonomy bar-plots above clearly suggest significant differences in the microbial composition of the soil samples between the studied sites. However, we still have not performed the formal significance tests for this (like we did when we compared alpha-diversity metrics). QIIME2 has tools for performing the appropriate statistical tests for Beta-diversity or Taxonomy too (e.g. https://docs.qiime2.org/2022.8/plugins/available/diversity/beta-group-significance/ or https://docs.qiime2.org/2022.8/plugins/available/composition/ancom/).

However, many interesting tools and algorithms are not yet adopted by QIME2. One of the ways to apply external tools to QIIME2-prepared data is to export the data from QIIME2 as was illustrated in the Step 5. Because of QIIME2 popularity however, there are also tools in other computational environments which can directly import QIIME2 ***.qza*** files.

This last part of the practical shows an example of importing QIIME2 data into R environment, plotting samples' Hierarchical Clustering Dendrogram, and performing a specialised statistical test (PERMANOVA) to estimate the significance of the differences that we visually observed earlier in PCoA plot.

You may now switch from Crescent2 cluster to RStudio on your own laptop. Copy the following files (prepared during your QIIME2 analysis) to a location accessible by RStudio:
s05_rooted_tree.qza, **s06b_rarefied_table.qza**, **s10_taxonomy.qza** and **samples.txt**.

Review the provided **s11_q2_to_R.html** file. Use it to prepare and run your own R markdown script.

First, the script will require installing several R packages: **qiime2R**, **phyloseq**, **vegan** and **dendextend**. Note that **qiime2R** is installed from GitHub by **devtools** (another R package that you may need to install). To avoid unnecessary installation issues, you may say "**No**" if you are asked whether you wish to install newer versions of some packages already available on your system, or if you are asked whether you wish to compile from source instead of using the pre-compiled versions.

Phyloseq is a very popular metagenomics R package that provides functions to "to import, store, analyze, and graphically display complex phylogenetic sequencing data" that has already been processed to the level of OTUs/ASVs: https://joey711.github.io/phyloseq . Some Phyloseq functions may duplicate the QIIME2 functionality. The provided script illustrates how to import QIIME2 data into Phyloseq format using **qiime2R** package.

The script should calculate PERMANOVA test using **vegan::adonis** function. **Vegan** is another popular metagenomics R package. It provides methods for ordination (aka PCA/PCoA) and for diversity assessment: https://github.com/vegandevs/vegan . QIIME2 plugins with similar functionality often are just wrappers around the **vegan** functions.

The script should also apply base R functions to plot a **Hierarchical Clustering Dendrogram** of samples based on the distance matrix.

Conclusion

This practical guided you through a complex bioinformatics pipeline:

- You started from installing and configuring SRA-Toolkit
- You retrieved the metagenomics dataset from NCBI SRA
- You performed QC using standard genomic QC tools (FastQC and MultiQC)
- You have completed a multi-step metagenomic analysis in QIIME2 environment, including data import, assessment, denoising (with DADA2), performing Phylogenetic, Diversity and Taxonomy analyses of the dataset
- You explored the methods to connect QIIME2 pipeline to other computational environments like R or the 3rd party tools for plotting phylogenetic trees.

In short, this practical session introduced you to QIIME2: the most popular current metagenomics integrated toolset. Not least, you practiced to organise scripts, logs, data, results and resources in a complex project that involved tens of input and output files; and you got a hands-on experience with analysis of real-life soil metagenomic data.

Well done!

You have completed the Metagenomics Practical Session.