TASK METHODOLOGY

Overall, I have used different software while solving this task. As I have very little experience with programming I relied on AI tools (mostly GEMINI) to help me with the task. I have installed the Virtual Studio Code, Python3, R, R Studio and FastQC. Generally, I used Virtual Studio Code and used the add-ins. The first task to sort out the data was carried out mainly in python. The second task to analyze the data was more complex and I used a panel of softwares that AI recommended to me. Firstly, I used fastqc and then multiqc built-in the Virtual Studio Code. To process the data, I have used the R Studio and relied on dada2 package.

TASK 1: Automated Data Inspection and Sorting

First of all, I have reviewed the names of the samples and checked whether they have a specific tag which could be used for sorting the data. The data were named in the same format, and therefore I have proceeded to review the files. I have listed the first line („header“) from all the files.

The example of a header no.1:

FILE: SOWX62_1.fastq.gz

HEAD: @SRR36802809.1.1 M03692:32:000000000-CYK55:1:1101:13191:1741 length=251

The example of a header no.2:

FILE: SZPB7N_1.fastq.gz

HEAD: @SRR36739997.32201.1 m84288_250212_110628_s4/236717910/ccs/12037_13529 length=1492

The dataset clearly differed in the identification code (letters and numbers) and by the length. I have observed that the data could be (even manually) diversified based on the lentgh, however I used the AI to explore other options. When I reviewed the headers by AI, I got the answer that the data was from Ilumina and PacBio platforms based on two parameters: the „/ccs/“ tag (PacBio) and „M03“ tag and short read length (Illumina). All of the files were sorted to two folders and can be found in \raw_data\ Illumina_Data and PacBio_Data.

CODE:

All was performed using python3. The commands can be found in the following file: sort_files.py

TASK 2: Platform-Specific Bioinformatics Pipeline

Overview:

I have decided to proceed with a further analysis only with the Illumina dataset.

Before I performed the analysis, I have installed the fastqc and multiqc using Terminal, the commands are as follows (performed on Ubuntu):

sudo apt install fastqc

sudo apt install multiqc

Then I have checked the QC data reports on my web browser.

The sequence processing was performed using the R Studio (integrated in VS Code).

QUALITY CONTROL

First, I have performed the quality control analysis for the individual files using fastqc and then created a summarized quality control report using multiqc. Fastqc analysis generated the „.zip“ and „.html“ files for all samples. Multiqc allowed me to analyze the whole dataset and I obtained the „.html“ file. I opened the file on my browser (see the screenshot below) and reviewed the main paramaters, especially *duplicate percentage, GC content, M Seqs* and *phred score*.

The dataset showed high levels of **duplicates** (all above 90 %), which might mean that the mostly only one sequence was amplified and the library shows low-complexity. Therefore, this might imply a bad quality data or some mistake during the process. However, low duplicate levels could be also due to the fact that the amplification was targeted, e.g. only 16S rRNA gene was amplified.

The dataset did not show high variability in the **GC content** which suggests that the samples are similar and do not vary much (potentially a stable microbial community or only one microbial source).

Phred score is good and the samples are located in the green line.

The **adapter content** was less than 1%, which is also a positive outcome.

On the other hand, there are parameters that did not pass the QC: "*Per Base Sequence Content*", "*Per Sequence GC Content*", "*Per Sequence GC Content*", "*Sequence Duplication Levels*" and the "*M Seq"* (total cont of read pairs) parameter which is zero. However, the "QC" failures are systematic, which suggests that these failures are due to the library design or the biologic nature of the samples rather than the sequencing failure. Also, I have discovered that the data have high per-base quality (meaning the sequencing was performed well), the dataset is not contaminated with adapter sequences and the samples do not show high variability and they are consistent.

Therefore, I concluded that it is meaningful to proceed with the further analysis.

CODE:

I have performed all the tasks in VisualStudio Code using python and the scripts can be found in the following files:

qc_analysis.py → generating the QC data

multi_qc_analysis.py →relocating the QC data to the different folder

command (terminal): *multiqc fastqc/ -o multiqc_data/* →generating summarized QC data

BASIC DATA STATISTICS

The read counts are in the range of 6,544 to 14,814 reads per sample, while the average is around 10-15 thousand reads. The read legth is around 248-251 bp which is typical for Illumina. The GC content across samples does not show any outliers and is consistent.

SEQUENCE PROCESSING AND FEATURE INFERENCE

First, I have checked whether the data is paired-end or single-end. I have concluded that the latter is true because the samples' names do not suggest they are paired-end (x_1, x_2). For this reason, it is not necessary to perform merging of the samples' sequences.

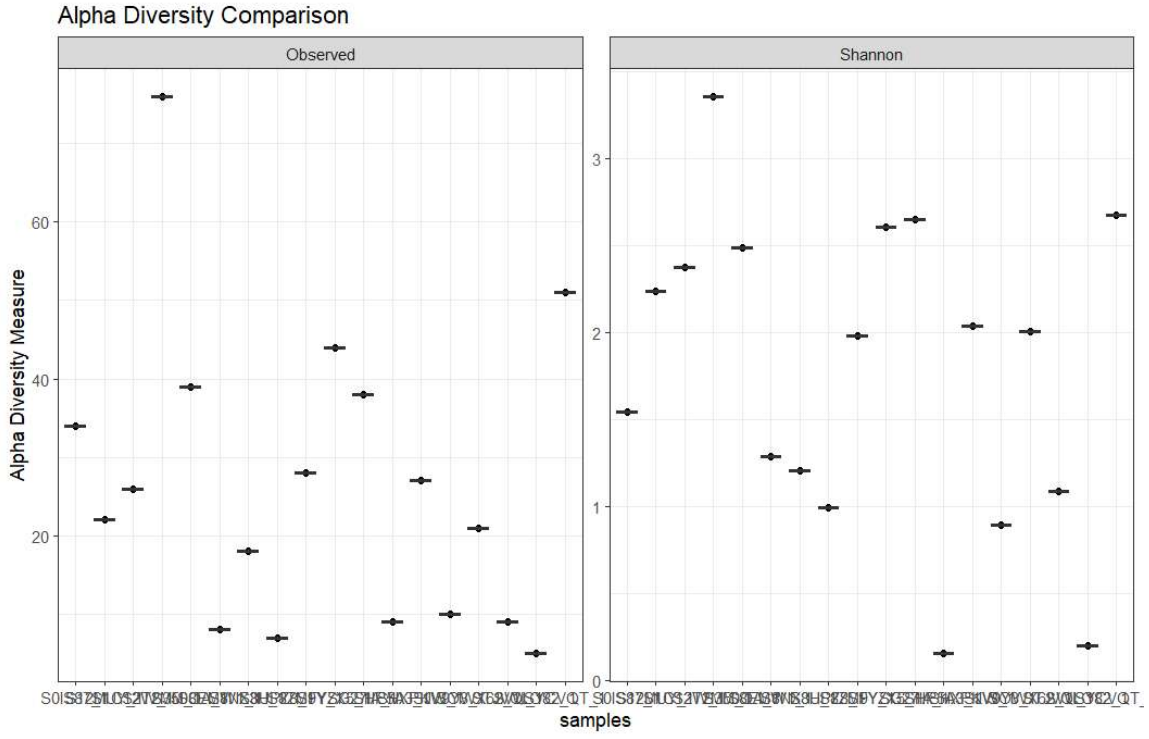
Also, I have omitted trimming of adapters as it does not seem to be necessary according to QC results.

I have created a new folder to concentrate the sequence processing tasks. Then I have performed filtering and trimming of the sequences. Then I dereplicated the sequences and did a "clean up" of the data to eliminate the noise. After I created a ASV table, I have removed chimera sequences.

Then I have downloaded the SILVA database, which is suitable for the 16S sequences. I created a TAXONOMY table – first I have assigned the genus and then species to my sequences.

I have changed the “labelling” of the sequences so the table matches the format of the provided example (those files are tagged “_labeled”).

For calculating alpha-diversity I have decided to choose Shannon Index which considers both richness and evenness and reflects overall community structure. Also, I based this decision on my previous work on thesis where we used this index. Below, you may find the visualization using R Studio. The plot shows two parameters: “observed” means the count of the ASV in the sample, while “Shannon” accounts to the Shannon index. The “observed” refers to the number of different species, while Shannon index shows the diversity of the community.



CODE:

This task was performed using R Studio, specifically the dada2 and ggplot2 package, and with the use of SILVA database for creating a taxonomy table.

The commands can be found in the /data_processing/scripts/script.R

The outputs are in the /data_processing: asv_table_labeled.csv, taxonomy_table_labeled.csv, alpha_diversity_results.csv