Bioinformatics Technical Assignment

Petra Polakovicova
Institute for Clinical and Experimental Medicine,
Center for Experimental Medicine, Prague
petra.polakovicova@ikem.cz

January 16, 2026

1 Overview

Welcome to our challenge! This challenge focuses on your general bioinformatics skills.

You may write the code in the programming language of your choice; however, it must be clearly readable and fully reproducible. You are free to use resources such as online tutorials, AI tools, and other external materials. However, please ensure that all external sources are properly referenced.

If anything about the following questions is unclear, or if you have any feedback, please let us know! Thank you for your time and interest in joining our lab, and good luck!

The goal of this assignment is to evaluate your ability to:

- Inspect and reason about sequencing data
- Automate data processing using the command line or scripting
- Design and execute an appropriate bioinformatics pipeline
- Assess data quality and summarize key statistics
- Communicate your work clearly and reproducibly

There is no single "correct" solution. We are primarily interested in **your reasoning, methodology, and clarity**, rather than only the final results.

2 Provided Data

You will receive a directory containing **multiple fastq.gz files**, representing single-end amplicon sequencing reads of the 16S rRNA gene from microbial samples.

The 16S rRNA gene is widely used in microbial ecology and taxonomy because it contains both highly conserved and variable regions. Amplicon sequencing of the 16S gene allows researchers to profile microbial communities by amplifying and sequencing specific hypervariable regions, enabling identification and comparison of bacterial taxa across samples.

You can download the data here: https://owncloud.cesnet.cz/index.php/s/1VkFDZbz9AiMaNU
Password: BIO2026

Important notes:

- The files originate from **two different sequencing platforms**.
- The sequencing platform information is not explicitly stated in the file names.
- All files are valid FASTQ files.

3 Task 1: Automated Data Inspection and Sorting

3.1 Objective

Identify the two types of sequencing data and **automatically split the FASTQ files into two separate directories**, one per sequencing platform.

Important notes:

- Do **not** manually inspect files one by one (you can do it for testing, but ultimately, you need to submit code that automates the process).
- Use **command-line tools** to analyze the FASTQ files.
- Base your classification on **data-derived properties** to distinguish the platforms.

3.2 Deliverables

A clear explanation of:

- What properties you inspected
- Why those properties distinguish the two platforms

The code or commands used to:

- Extract the relevant information
- Automatically assign files to one of two directories

4 Task 2: Platform-Specific Bioinformatics Pipeline

4.1 Objective

Choose **one** of the two sequencing platforms and process **only the data from that platform** through a complete bioinformatics pipeline.

You are **not given a predefined pipeline**. Part of this task is to demonstrate that you can:

- Identify the appropriate analysis steps for the chosen data type.
- Find and justify suitable tools.

4.2 Expected Components

While the exact implementation is up to you, your final report should include:

1. **Quality Control**
    - Describe the data quality

- Include relevant QC outputs and interpretations

2. **Basic Data Statistics**
   - Summarize key characteristics of the reads (based on what you consider relevant).
   - Clearly state what you consider “basic statistics” and why
3. **Sequence Processing and Feature Inference**
   Perform the necessary steps to obtain:
   - An ASV (Amplicon Sequence Variant) table
   - A taxonomy (TAXA) table

4.3 Output Format

Below are the output formats recommended for the ASV and TAXA tables. Examples can be found in the provided link along with the input data.

SeqID	<SampleName1>	<SampleName2>	...
ACGTAA..	0	30	...
AGGGCT...	20	120	...
...	...	...	...
...	...	...	...

Table 1: ASV table structure

SeqID	Domain	Phylum	Class	Order	..	..
ACGTAA..	Bacteria	Pseudomonadota	Gammaproteobacteria	Pseudomonadales	..	..
AGGGCT...	Bacteria	Bacillota	Clostridia	Oscillospirales	..	..
...	...	...	...	..	..	..
...	...	...	...	..	..	..

Table 2: TAXA table structure

5 Task 3: Alpha Diversity Analysis

5.1 Objective

Using the ASV table you generated, compute and visualize **alpha diversity** across samples.

5.2 Requirements

You may use any programming language (e.g., R, Python). The choice of alpha diversity metric(s) and visualization type is up to you. However, you should be able to explain the details behind the calculations and visualizations.

5.3 Expectations

- Clearly state which diversity metric(s) you selected and why
- Produce at least one meaningful visualization
- Ensure that the code is clear and fully reproducible

6 Code, Reporting, and Submission

6.1 Code

All code must be provided and should be:

- Readable
- Commented where appropriate
- Structured logically
- Reproducible

6.2 Report

Prepare a clear and concise report (in any form you prefer, e.g., Markdown, PDF, HTML, notebook) describing:

1. Your overall approach
2. Key decisions and assumptions
3. Interpretation of results

6.3 Submission

Upload **all code and the report to a GitHub repository**. If you wish, you may demonstrate your Git skills by committing your work incrementally so that the commit history is visible. If you are new to Git or prefer not to do it, you may upload the final files directly using the web interface.

The repository link will be provided to you. However, you must have a valid GitHub account; if you do not already have one, please create it in advance.

Good luck, and we look forward to reviewing your work!