Bioinformatics | BIOL3802 + BIOL8802 ONC-U S2 2025

Assignment 1: Merging, Quality Control, and Filtering

- Access Deepthought using Jupyter Notebooks (http://deepteachweb.flinders.edu.au/jupyter) and complete the following exercises.
- Submit answers to the **bolded questions** (Questions 1, 2, 3, 4, 5, 6, 7, 8)

Part A — Merging, Quality Control, and Filtering Exercise

Step 1 — Download and Prepare Data

- Download the dataset (8 FASTQ files in .zip format) using the wget command:

wget https://zenodo.org/record/1236641/files/test_fastq_small.zip

- Unzip the downloaded file to obtain the 8 separate FASTQ files.
- These 8 files represent 4 samples, each with:
    - Forward reads (R1)
    - Reverse reads (R2)
      This is because the samples were generated via paired-end sequencing.

Step 2 — Initial Quality Control

- Run FastQC on the Test01 R1 and R2 FASTQ files.

  Question 1: /4

    - **From the FastQC report, provide screenshots of:**
        - **Basic Statistics for each file**
        - **Per base sequence quality plots for both the Test01 R1 and R2 FASTQ files.**

Step 3 — Merging Paired-End Reads

- Use the following fastp command to merge paired-end reads for Test01:

```
fastp \
 -i Test01_L001_R1_001.fastq \
 -I Test01_L001_R2_001.fastq \
 --merge \
 --merged_out Test01_merged.fastq \
 --disable_adapter_trimming \
 --disable_quality_filtering \
 --html fastp_merge_report.html \
 --json fastp_merge_report.json
```

<u>Question 2</u>: Briefly explain what the above code does to the R1 and R2 FASTQ files. /2

Step 4 — Post-Merging Quality Control

- Run FastQC on the merged Test01 FASTQ file from Step 3.

  <u>Question 3:</u> Are the sequences in the merged file longer or shorter than the original R1 and R2 reads? Explain why this might be the case. /2

Step 5 — Repeat Merging for All Samples

- Repeat the fastp merging step for:
  - Test02
  - Test03
  - Test04
    (Make sure to update the output filenames accordingly.)

Step 6 — Filtering Merged Reads

- Using the fastp filtering command from the GitHub tutorial, filter all four merged FASTQ files using:
  - Quality threshold: -q 30
  - Minimum length: -f 30

  <u>Question 4:</u> Run FastQC on the new filtered Test01 FASTQ file.
  Compare the Basic Statistics report before and after filtering — are there differences? /2

<u>Part B — Alignment Exercise</u>

Step 7 — Setup

- Copy the files from:
  https://github.com/N-falk/Bioinformatics_2025/tree/main/Assignment1
  onto your Deepthought.
- Install Minimap2:

curl -L https://github.com/lh3/minimap2/releases/download/v2.30/minimap2-2.30_x64-linux.tar.bz2 | tar -jxvf -

./minimap2-2.30_x64-linux/minimap2

- Create a Conda environment and install Samtools:

conda install -c bioconda samtools

Step 8 — Index the Reference

- From the directory containing the minimap source files that you just donwloaded, index the reference FASTA file using the following minimap2 command:

minimap2 -d ref.mmi reference.fasta

<u>Question 5:</u> When aligning DNA sequences, what is the purpose of the reference genome? /3

Step 9 — Align Reads to the Reference

- Run the following commands to perform alignment, sorting, indexing, and statistics:

minimap2 -ax sr ref.mmi sample_merged.fastq.gz | samtools sort -o merged.sorted.bam

samtools index merged.sorted.bam

samtools flagstat merged.sorted.bam

<u>Question 6:</u> Briefly describe the difference between SAM and BAM file formats. /4

<u>Question 7:</u>

- **What percentage of reads from Test01 mapped to the reference? /1**
- **What might this indicate? /1**

Step 10 — Align E. coli Reads

- Repeat the alignment process for the dataset:

ecoli_1K_2.fq.00.0_0.cor.fast.gz

<u>Question 8</u>:

- **What percentage of reads from the *E. coli* FASTQ file mapped to the reference? /1**
- **What might this indicate? /1**