**Assignment No.1**

**Adapter & Quality Trimming with TrimGalore**

**Objective**

Learn how to install and run bioinformatics tools in Linux environment, using Google colab. Practice Linux commands, bash shell scripting and file manipulation. Learn to perform adapter and quality trimming on RNA-Seq data using TrimGalore, a wrapper around Cutadapt and FastQC.

**Tasks**

1. **Set Up Environment**
   * Mount Google Drive in Colab.
   * Install TrimGalore and its dependencies (cutadapt, perl).
2. **Process Paired-End RNA-Seq Data**
   * Trim adapters and low-quality bases using TrimGalore for 6 samples:  
     YapPool\_S01, YapPool\_S02, YapPool\_S03, YapPool\_S09, YapPool\_S10, YapPool\_S11.
   * Use the following parameters:
     + Quality cutoff: 20 (Phred score).
     + Minimum read length: 20 bp.
     + Trim 2 bp from both 5’ and 3’ ends (--clip\_R1 2 --clip\_R2 2).
     + Compress outputs (--gzip).
3. **Automate with a Shell Script**
   * Write a script (run\_trimgalore.sh) to process all samples in a loop.
   * Save outputs to a directory: clean.trimgalore.RNA-Seq/.
   * Log trimming statistics for each sample.
4. **Generate QC Reports**
   * Run MultiQC to aggregate TrimGalore and FastQC reports.
   * Interpret the HTML report (check adapter content, read quality, etc.).

**Input Data**

* Paired-end FASTQ files in:  
  drive/MyDrive/Lab\_share/Lab\_data/class\_466/raw.small.RNA-Seq/.

**Deliverables**

1. **Colab Notebook** with:
   * Installation commands.
   * TrimGalore execution for one sample (manual run).
   * Shell script for batch processing.
   * MultiQC output.
2. **Short Report** (PDF or Markdown):
   * Compare input vs. output read counts (from logs).
   * Screenshots of MultiQC plots (e.g., per-base quality).
   * Explain key parameters (e.g., --clip\_R1, --quality).

**Tips**

* Refer to [TrimGalore documentation](https://github.com/FelixKrueger/TrimGalore" \t "_blank).
* Debug errors by checking file paths and dependencies.
* Use !head to inspect FASTQ files before/after trimming.

**Grading**

* **Correctness** (50%): Script runs without errors; outputs match expected formats.
* **Completeness** (30%): All tasks documented in the notebook/report.
* **Critical Thinking** (20%): Interpretation of QC metrics and parameter choices.

1. Use grep, awk and sed to finish sequence file and GFF3 file manipulation