## Part 1: Docker Setup

- Create a Docker container using a Dockerfile that starts from a base Ubuntu image.

## Part 2: Package Installation

- Attempt to install the following R packages:
    - Seurat (follow installation guide: https://satijalab.org/seurat/articles/install_v5.html)
    - Signac (GitHub repository: https://github.com/stuart-lab/signac)
- If any installation fails, in the report, provide a clear explanation of the issues encountered and why the installation was unsuccessful.
- Possibly try to find a work around to the installation fails, if you fail at all in installing the package(s) describe the reason why it was impossible to find a work around.

## Part 3: Data Processing and Analysis

Use the dataset: “PBMCs 3k cells from a healthy donor”

- use the material provides as part of the exam.

### *Step 1: Matrix Conversion*

- Use the Matrix R package to convert the sparse matrix into a full matrix.
- Save the result as a data.table object.

### *Step 2: Split Gene Expression and ATAC-seq Data*

- From the data.table object, separate:
    - Gene expression data (rows labelled with Ensembl gene IDs, e.g., ENSG00000243485)
    - ATAC-seq peak data (rows labelled with genomic coordinates, e.g., chrN:NNNN-NNNN)

### *Step 3: Summarize Data*

- For each dataset (expression and peaks), compute the column-wise sum to produce:
    - A single vector of total expression per gene
    - A single vector of total chromatin accessibility per peak region

### *Step 4: Create Genomic Ranges*

- Convert both the summarized gene expression and peak data into GenomicRanges objects.
- Add the summarized data as metadata to their respective GenomicRanges.

### *Step 5: Gene Annotation for ATACseq data*

- Using the annotation file Homo_sapiens.GRCh38.114.gtf.gz:
    - Create a GenomicRanges object only for protein-coding genes and only for gene features.
    - Remap the ATAC-seq GenomicRanges to this object and attach the summarized peak data from step 4.

### *Step 6: Finalize Expression Data*

- Subset the expression GenomicRanges, step 4, to include only protein-coding genes.
- Add gene symbol identifiers to the object.

### *Step 7: Data Normalization and Integration*

- Normalize both expression and ATAC-seq data using CPM:
    - Divide each column by the column sum, multiply by $10^6$, add a pseudo-count of 1, and apply $\log_2$.
- Merge expression and ATAC data based on common genes.
- Provide a summary table of the number of ATAC peaks that could not be merged and a plot of peak intensity distribution chromosome by chromosome. Provide a summary table of the genes which do not show association with ATAC peaks and plot their expression distribution chromosome by chromosome

### *Step 8: Visualization*

- Generate a scatter plot using ggplot2:
    - X-axis: log-transformed expression CPM
    - Y-axis: log-transformed ATAC CPM
- If the plot is too much busy of data divide the plot in the 24 chromosomes

## Part 4: Reporting

- Create an HTML report using RMarkdown that includes:
    - All code used
    - Output results (tables, plots, etc.)
    - A description of any problems or errors encountered
    - Possible workarounds or solutions you tried
    - Describe how to repeat the analysis using the docker provided.
- Provide the docker with the software installed which includes the scripts used to perform the analysis.