# Screening Problem

Write a WDL workflow to analyze gene-count matrices of human bone marrow single-cell data using SCANPY.

## 1. Data

You are given 8 10x channels (i.e. 8 donors' data). You can download them either via HTTPS:

| Sample | HTTPS Link |
|---|---|
| MantonBM1_HiSeq_1 | https link |
| MantonBM2_HiSeq_1 | https link |
| MantonBM3_HiSeq_1 | https link |
| MantonBM4_HiSeq_1 | https link |
| MantonBM5_HiSeq_1 | https link |
| MantonBM6_HiSeq_1 | https link |
| MantonBM7_HiSeq_1 | https link |
| MantonBM8_HiSeq_1 | https link |

Or via google-cloud-sdk:

```
gsutil -m cp -r gs://terra-featured-workspaces/Cumulus/cellranger_output
```

Each h5 file contains the unfiltered gene-count matrix of one donor's data. Your analysis will start from them.

## 2. Analysis Steps

You may follow SCANPY's tutorial on how to analyze single-cell data.

The analysis consists of the following steps:

1. Quality control: For cells, keep those with $ 500 \leq$ number of genes expressed $< 6000$, and with percent of mitochondrial genes (name prefix `MT-`) $< 10\%$. For genes, keep those expressed in at least $0.05\%$ of cells.
2. Log-normalize data with $100,000$ reads per cell.
3. Select top 2000 highly variable genes.
4. Get Principal Component Analysis (PCA) embedding with 50 PCs.
5. Get Nearest Neighborhood graph of 100 neighbors from the 50 PCs.
6. Leiden clustering on PCA embedding with resolution 1.3.
7. Calculate UMAP embedding from PCA embedding, and generate UMAP plot with cells colored by their leiden labels.
8. Find marker genes for each leiden cluster using Mann-Whitney-U test, and generate the gene rank plot.

## 3. Develop a WDL workflow for analysis

1. Build a docker containing all the necessary pakcages you'll need for the tasks above.
2. Develop a WDL workflow in version 1.0 specs to wrap the steps above, which runs on top of the docker image you build in Step 1.
3. For each channel, respectively, analyze the data using the steps above.

When finished, the workflow should give the following output:

| Name | Type | Description |
|---|---|---|
| count_matrix_h5ad | `Array[File]` | The count matrix in `h5ad` format containing filtered cells and clustering results. It's a list since each channel should return one such file. |
| umap_png | `Array[File]` | The UMAP plot of each channel. |
| gene_rank_png | `Array[File]` | The gene rank plot of each channel. |

# 4. Test and Run in local machine

1. Install Java environment (e.g. OpenJDK) and Docker on your computer.
2. Download Cromwell which is the workflow engine for WDL.
3. (Optional) You may use `womtool` downloaded from the link above to check the syntax of your WDL worflow:

```
java -jar womtool.jar validate your-workflow.wdl
```

4. Run your workflow with Cromwell:

```
java -Dconfig.file=custom.conf -jar cromwell.jar run your-workflow.wdl -i inputs.json
```

where * You can modify custom.conf that we provide as a template. In specific, change Line 33 to map the correct path into your docker container during execution. * `inputs.json` is a JSON file you'll need to create, which contains the workflow inputs.

# 5. Results

Create a GitHub/GitLab/BitBucket repo, including:

- Source code: `Dockerfile` and WDL file
- README file stating how your workflow works and how you run it
- UMAP and gene rank plots generated by your workflow