Worksheet 8.2 Single Cell RNA-Sequencing

Part 1. Running CellRanger

1. Log on to cluster using terminal. Change working directory to `/scratch/Users/<github_username>`

2. Make a directory called `singlecell` and copy a sbatch template to this folder.

3. In the sbatch file, request 1 node, 1 task, 30 minutes running time, and 16GB memory.

4. Load the cellranger module by `module load cellranger`

5. Execute cellranger with the following command

```
cellranger count --id=<job_name> \
                 --transcriptome=<path_to_ref_genome> \
                 --fastqs=<path_to_fastq> \
                 --sample=<sample_name> \
                 --localcores=1
```

The arguments are

- <job_name>: Any job name you like.
- <path_to_ref_genome>: `/scratch/Workshop/SR2019/8_GATK/singlecellrun/refdata-cellranger-GRCh38-3.0.0`
- <path_to_fastq>: `/scratch/Workshop/SR2019/8_GATK/singlecellrun/fastqs`
- <sample_name>: `hgmm_100`

6. Submit the job.

Part 2. Understanding the Outputs

1. Obtain output files
   - Log on to cluster using terminal. Change working directory to `/scratch/Workshop/SR2019/8_GATK/singlecelloutput/`
   - This folder contains two datasets from two different experiments (1_Ethan_Control and 3_Eric_Control, analyzed by `cellranger count`) as well as the aggregate of these two datasets (Agg1_EthanEric_ControlsOnly, created by `cellranger aggr`). The csv files are used to run `cellranger aggr` and can be ignored for this worksheet.

- Locate the summary files (web_summary.html) from the datasets and the aggregate and transfer the files to your local computer.

2. Dataset 1_Ethan_Control. Answer the following questions with the information from the summary file.

- Number of cells.
  - What is the estimated number of cells in this dataset?
  - How was this number determined? Hint: look at the UMI counts vs Barcodes plot.
  - Why not simply keep all barcodes?
- Reads, transcripts (UMI), and genes
  - On average, how many reads, transcripts, and genes were detected in one cell?
  - Does the number of reads equal to the number of transcripts? Why?
  - How many genes does a human genome contain? Does this number equal to the number of genes detected in a cell? If a gene is not detected, what are the possible reasons?
- Sequencing depth (mean reads/cell)
  - What is “sequencing saturation” reported by cellranger?
  - How does this quantity change as a function of the sequencing depth? Hint: look at the bottom of the “Analysis” panel.
- Graph-based clustering and differential expression analysis
  - How many clusters were generated? How many cells does each cluster contain? Do you always find a clear boundary between two clusters?
  - Do you find any cluster with significantly more/less transcripts detected compared to other clusters? If yes, what are the possible reasons?
  - For each cluster, list the top 5 differentially up-regulated genes. What are their fold changes and p-values? Does every cluster always have differentially expressed genes? Does a large fold change always lead to a small p-value?
  - Based on the differentially expressed genes, infer the activities of the cells in cluster 2 and 3.
- k-means clustering
  - Switch the clustering type to k-means where k equals to the number of clusters generated by the graph-based method. Do you observe the same clusters?
  - Compare k=2, 5, and 10 and observe how the clusters and top differentially expressed genes change as a function of k.
  - In general, how do you choose the value of k?

3. Dataset 3_Eric_Control. Answer previous questions with the information from the summary file.

4. Aggregate Agg1_EthanEric_ControlsOnly. Answer the following questions with the information from the summary file.

- What is the mean reads per cell after aggregation? Why does this number approximately equal to the mean reads in dataset 3_Eric_Control but not dataset 1_Ethan_Control?
- Compare the t-SNE plot of the aggregate with the t-SNE plots of individual datasets. Are the cells from two datasets well-separated on the t-SNE plot? Why?
- Determine which dataset each cell originates from. Do you observe any pattern on the t-SNE plot colored by UMI counts? What biological insights can you conclude?
- Change the clustering type to k-means with k=2. Are the cells clustered by their datasets? Increase the value of k and observe how the clusters change. Does each cluster contain a similar number of cells from these two datasets? What biological insights can you conclude?