412 Take-home exam

Ziyin Zhao

5/2/2022

**Methods**

**Data preparing**

The raw data used for this take-home exam is the data processed in the term project. The SRR ids for each cell sample were obtained via the command:

esearch -db sra -query PRJNA438394 | efetch –format runinfo | cut -d “,” -f 1| grep SRR

FastQs were downloaded through Sratoolkit, then the data of each sample was processed by the following pipeline:

- TrimGalore and FastQC (on both fastq files if paired-end data) -> Trimming FastQs

- STAR -> alignments from FastQs

- R-SEM -> counts from BAMS

The R-SEM raw count and TPM count data for each gene and sample of each donor are merged into a single matrix. Due to issues with the fastq files, 42 samples from donor-40 and 1 from donor-54 were removed from the analysis. The data received would be assessed according to the quality control thresholds: minimum read mapping percentage in excess of 95% and maximum percentage of reads lost to trimming below 5%.

**Seurat**

The Seurat package is used to analyze the data. The matrices of raw counts for each sample (Donor 37, 38, 39, 40, 53, 54) are loaded into R, then combined to generate the Seurat object for further analysis. The Seurat object for total data contains 1042 cell samples; each sample contains 21221 genes. Since the data quality is not good enough, after filtering out data with nFeature\_RNA less than 200 and greater than 2500 were filtered out, 369 samples were filtered out from the seruat subject, 673 samples are remained in the data, with 21221 genes for each sample.

Chart, radar chart

Description automatically generatedChart

Description automatically generatedFigure 1. VlnPlot for total data before and after filtering

The following pipelines are performed on the data:

* Normalizing (NormalizeData)
* Identification of highly variable features (FindVariableFeatures and VariableFeaturePlot)
* Scaling the data (ScaleData)
* Linear dimensional reduction and PCA (RunPCA)
* Determine the "dimensionality" of the dataset (JackStraw and ScoreJackStraw)

FindNeighbors identify the shared nearest neighbor (SNN) of the cells, and the cells are clustered based on the SNN by FindClusters.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 2. UMAP and tSNE plot of clustered cells.

The cells are clustered into 6 groups based on the SNN. The biomarkers of each cluster are identified. A Heatmap of clustered cells is generated to show the expression of top biomarkers of each group in total data.

Graphical user interface, chart, text

Description automatically generated

Figure 3. Heatmap of clustered cells with biomarkers of each cluster

**Analysis**

**Pathway enrichment in epithelial cells**

The biomarker of epithelial cells and cancers is used to identify the epithelial cells and cancer cells in the total dataset:

* Epithelial cells: CAPS, TMEM190, PIFO, SNTN
* Cancer: EPCAM

Diagram

Description automatically generatedChart, diagram

Description automatically generated

Figure 4. Expression of epithelial cell and cancer cell biomarkers

Since the flow cytometry was performed to isolate the Cytotoxic T cell (CTLs) for single-cell sequencing (smartseq2), not many cells other than CTLs are expected to be found in this data. SCINA is performed on the total data to verify the presence of epithelial cells and cancer cells in the data by expressing biomarkers of those two genes. As a result, only 5 cells are classified as epithelial cells. The sample size is too small to perform GSEA (it didn’t pass the thresholds across all subgroups)

**Former smokers and non-smokers in lung cancer data**

Donors 37, donor38, donor 39, and donor 40 have known information on whether they smoke or not. Donor 37, donor38, and donor 39 are former smokers, and donor 40 never smoked before. Donor 37, donor39, and donor 40 are diagnosed with the same type of cancer (Lung Adenocarcinoma), donor 38 is diagnosed with Lung Squamous. To eliminate bias, a subset of the Seurat object is generated, which only contains information for those three donors with the same kind of cancer. 158 samples from donors 37 and 39 are classified as “Ex,” and 107 samples from donor 40 as “Never.“

By comparing the umap between the “Ex” group and the “Never” group, cells in cluster 2 show little in the never-smoking group; meanwhile, cluster 3 is relatively denser in the never-smoking group.

Chart, scatter chart

Description automatically generated

Figure 5. umap of data between Ex-smoker group and never smoke group.

The pathway enrichment is performed between the CTLs of the Ex-smoker group and the never smoke group by performGeneSetEnrichmentAnalysis in the cerebroApp package. Using the h.all.v7.5.1.symbols.gmt, the HALLMARK\_ANGIOGENESIS pathway shows a significant difference between the two groups.

**Expression of IL4 and IL17 signaling pathway**

IL4 and IL17 are not expressed in the CTLs data.

Chart, scatter chart

Description automatically generated

Figure 6. expression of IL4 and IL17 genes in total data

Too low expression results in the inability of the crosstalk to generate the signaling pathway of these two genes. The signaling pathways for IL4 and IL17 identified in the thesis are used to create a feature plot.

IL7R, IRF2, and GTF3A are the genes in the IL4 signaling pathway. FADD, CASP3, CASP8, RELA, IL17RA, IL17RB, and TRADD are the genes in the IL17 signaling pathway. The result shows that those genes are distributed in each cluster.

**Conclusion**

Since the flow cytometry was performed to isolate the Cytotoxic T cell (CTLs) for single-cell sequencing. The types of cells are limited, so studies on the other cells, including lung epithelial cells, cannot be reproduced with this data. The symptoms of COPD in the donors are not introduced in the supplement table. A data set more suitable for research goals should be used to find the link between COPD and lung cancer.

**Reference**

Skander, D. (2019). Integrative'Omics Approach to Investigate Relationship Between COPD and Lung Cancer (Doctoral dissertation, Case Western Reserve University).