Jonathan Lyda
Dr. Lareau
BIOENG 249
16 December 2024

Lung and Trachea Differential Gene Expression: Insights into Disease Risk

Why is lung tissue more susceptible to diseases like COVID-19, ARDS, and pulmonary fibrosis when compared to other parts of the respiratory system, such as the trachea? The dataset I chose gave me enough high quality lung and tracheal cell-level data to show key differentially expressed genes. Having cell-level data, in addition to mouse-level data, was crucial to directly comparing lung and tracheal cells, leading to the identification of highly expressed, tissue-specific genes that may explain susceptibility to certain diseases.

This project reveals genes that are highly expressed in lung tissue but not in tracheal tissue, which coincide with their biological roles in certain lung diseases. The differences in gene expression between lung and tracheal tissues, particularly the expression of Egfl7, Cldn5, and Esam, create lung-specific susceptibility to disease. Identifying lung-specific genes can direct development of targeted therapeutics and biomarkers for lung-related diseases such as COVID-19, ARDS, and pulmonary fibrosis.

Dataset and Methods

The source of data was from a Single-cell RNA-seq experiment that included gene count matrices for cells grouped by their tissue of origin, metadata, and annotation data. The data size for lung and tracheal data was 3,314 cells and 23,433 genes, with 1,391 lung cells and 1,923 tracheal cells. However, after filtering (filtered library size and lowly expressed genes) there were 3,141 cells (library size) and 17,243 genes. After loading the lung, trachea, annotations, and metadata into multiple data frames, I filtered cells with abnormally low library sizes, removing them using the 4.7th and 99.5th percentiles. I followed this by filtering out rare genes (those that were expressed in fewer than 8 cells). I normalized the library size to correct for differences in sequencing depth, followed by square root transformation, which stabilized the data variance.

With all of the data processed, I applied dimensionality reduction with Principal Component Analysis (PCA), reducing the data to 50 principal components to help with clustering and visualization. I additionally used UMAP as another dimensionality reduction technique to further visualize patterns of cell clustering.

I used KMeans clustering to separate the data into 4 clusters. I tried Phenograph and Spectral clustering, however I went with KMeans clustering due to its accuracy and simplicity. With this clustering method, I used a t-test to identify the 20 highest differentially expressed genes in the data. The most significant genes were ranked by the t-statistic, with Egfl7, Cldn5, and Esam as the top 3 differentially expressed genes. With this differential gene expression analysis, I was able to gain some insight into why certain diseases specifically target lung cells as opposed to trachea cells.

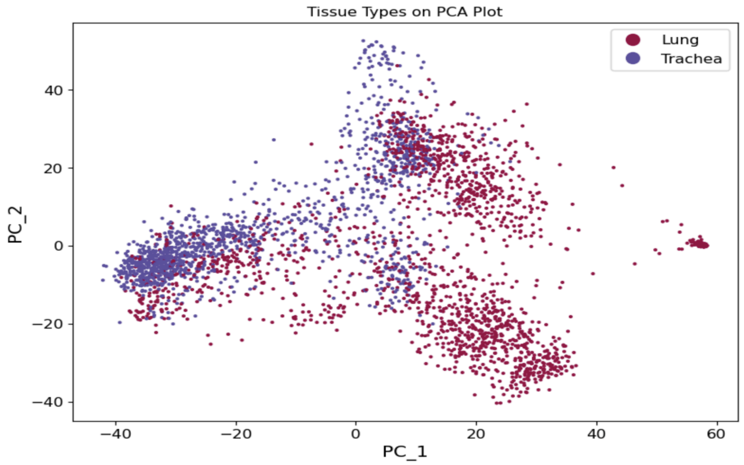

Figure 1: Cells from lung and trachea tissue plotted with the first 2 principle components. The cells cluster with other cells from the same tissue type.

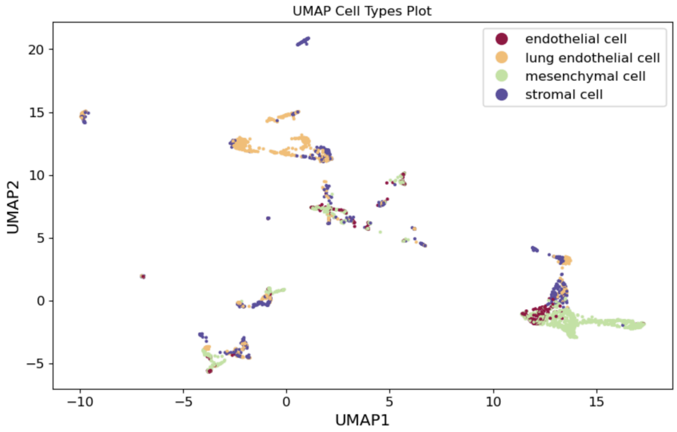

Figure 2: Various cell types from both lung and trachea tissue plotted with UMAP (dimensionality reduction). The cells are clustered with the same cell type.

Results and Key Figures

Egfl7, Cldn5, and Esam were found to be the most differentially expressed genes and they are all located in lung tissue cells (Figure 1 and 3). I also found that the lung and tracheal cell type populations were separated when colored by cell type, indicating differences and distinction

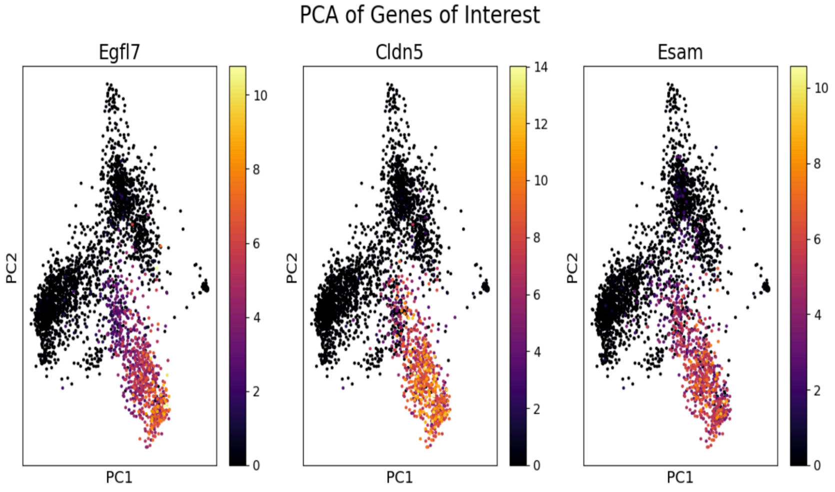

Figure 3: Genes with the highest differential gene expression (Egfl7, Cldn5, Esam). They are all expressed in lung tissue, as the location of expression on the PCA plots above are where lung cells are being plotted.

between cell types (Figure 2). A clear example of this is the separation of endothelial cells from lung tissue and tracheal tissue, despite both being classified as endothelial cells.

Discussion and Interpretation

Lung-specific gene expression can explain susceptibility of the lungs to certain diseases. For example, Egfl7 (epidermal growth factor-like protein) has a direct role in vascular integrity and endothelial migration, which are linked to vascular leakage, making it susceptible to diseases like COVID-19, ARDS, and pulmonary fibrosis. Cldn5 (Claudin-5) creates tight junctions in lung endothelial cells, and its disruption is associated with pulmonary edema in ARDS. Esam (endothelial cell adhesion molecule) keeps endothelial cell junction integrity. Its dysfunction can lead to vascular leakage and inflammation, which can be caused by COVID-19 and ARDS. With this information, researchers can use these genes as biomarkers to monitor changes in their expression for early detection of disease. Additionally, targeting these genes through lung-specific drug delivery could prevent or lessen the damage of diseases like COVID-19, ARDS, and pulmonary fibrosis.

Limitations and Future Directions

The main limitation of my analysis was the size of my dataset, which only had 3,314 total lung and tracheal cells. Additionally, the data from Tabula Muris was only collected from 8 mice total. This leads into future considerations of increasing the size of the dataset, increasing cell count, and the amount of mice sequenced. Expanding the size of the data will lead to more statistically sound and robust results.

I would like to design disease specific studies to determine if Egfl7, Cldn5, and Esam are involved in the progression of ARDS, COVID-19, and pulmonary fibrosis. More specifically, I would like to figure out if Cldn5 and Esam could be targeted to prevent pulmonary edema in ARDS, since both of those genes control vascular permeability.

References

The Tabula Muris Consortium., Overall coordination., Logistical coordination. *et al.* Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* 562, 367–372 (2018). https://doi.org/10.1038/s41586-018-0590-4