**Dante Freeman**

***Single-Cell RNA-seq Analysis of Non Small Cell Lung Cancer Shows Diverse Cell Populations in Tumor Microenvironment***

**Background:**

Non-small cell lung cancer (NSCLC) is 85% of lung cancer cases and causes the majority of cancer-related deaths worldwide [1]. Single-cell RNA sequencing allows analysis of individual cells within a population, making it a powerful tool for studying cancer cells, immune cells, and healthy tissue cells.

This project analyzes publicly available scRNA-seq data from an NSCLC tumor sample to identify key marker genes and cluster distinct cell populations. The Seurat package in R is used to perform quality control, normalization, principal component analysis, clustering, and marker gene identification.

**Methods**

1. **Data acquisition and loading:**

The scRNA-seq data set is 20k\_NSCLC\_DTC\_3p\_nextgem\_Multiplex\_count\_raw\_feature\_bc\_matrix, and it was loaded in R and processed with Seurat package. The data was imported as a Seurat object for downstream analysis.

1. **Quality Control and Normalization:**

Quality control filters were applied and cells that had too few genes (<200) or too many genes (>2500) were removed. This was done so that cells with poor sequencing would be excluded, and potentially clumped or closely packed cells would be excluded.

The raw expression counts were normalized with LogNormalize. This accounts for differences in sequencing depth across the cells.

A group of black dots

AI-generated content may be incorrect.

*Figure 1: Violin plots of QC Metrics. This figure shows the distribution of the QC metrics.*

* nFeature\_RNA: number of genes detected per cell
* nCount\_RNA: total RNA molecules per cell
* percent.mt: proportion of reads which are mitochondrial genes

Cells with low gene counts are dead or poor quality. High mitochondrial content is stressed or dying cells.

A red and blue dotted line

AI-generated content may be incorrect.

*Figure 2: Feature Scatter Plot of nCount\_RNA vs nFeature\_RNA*

This scatter plot is comparing number of RNA molecules to number of expressed genes. High quality cell data will typically have a linear relationship between these features. This is because more RNA means more genes detected, typically. Low quality cells will be outside this.

1. **Identification of Variable Features**

A red dot on a white background

AI-generated content may be incorrect.

*Figure 3: Highly variable Genes*

This figure shows the top 2000 most variable genes. This is used for dimensionality reduction. The top 10 mostvariable genes are labeled. These genes tell the most meaningful differences in gene expression across cells.

Highly variable genes were identified to capture important variability in gene expression. These genes were used to scale the data, centering and standardizing expression values across cells.

1. **Scaling and PCA**

Dimensionality reduction was performed using principal component analysis (PCA), focusing on the top 15 principal components to summarize the major sources of variation.

1. **Clustering and UMAP Visualization**

Clustering was done using a graph-based algorithm built on a nearest-neighbor graph of cells. Essentially, this is clustering cells based on how similar their gene expression profiles are. To visualize the clustering results, Uniform Manifold Approximation and Projection (UMAP) was used to reduce the high-dimensional data into a two-dimensional space.

**A map of different colors

AI-generated content may be incorrect.**

*Figure 4: UMAP Plot of Cell Clusters*

Uniform Manifold Approximation and Projection visualization of clustered single cells from NSCLC tumor tissue. Each point is a cell. Colors represent clusters based on similar gene expression.

1. **Marker-Gene Annotation and Cell Type Annotation**

Finally, marker genes were identified by comparing gene expression across clusters, using thresholds for minimum expression percentage and log fold change. This was accomplished using Seurats FindMarkers command. Known canonical markers for epithelial, immune, and proliferative cells were visualized using feature plots to aid in biological interpretation.

**A screenshot of a computer

AI-generated content may be incorrect.**

*Figure 5: Feature Plots of Canonical Marker Genes*

This figure shows expression of common marker genes in UMAP cluster space (from previous figure) to aid in cell identification. Higher expression is darker shades of blue.

EPCAM – epithelial tumor cells, Cluster 0

CD3D – T cell marker.

CD14—monocytes/macrophages

MS4A1 – B cell marker

MKI67 – proliferating cells

**Discussion**

**A group of colored dots

AI-generated content may be incorrect.**

*Figure 6. UMAP Plot of NSCLC Clusters Annotated by Cell Type.*  
scRNA-seq data clusters are labeled based on canonical marker gene expression (e.g., *EPCAM*, *CD3D*, *CD14*, *MS4A1*, *MKI67*). Cell types include epithelial (tumor) cells, T cells, B cells, monocytes, proliferating cells, and others. Annotation allows biological interpretation of the tumor microenvironment composition.

In this study, sequencing data from scRNA-seq for NSCLC tumor specimen was prepared, normalized, dimensionality reduced, clustered, then analyzed for markers. Six major cell populations were identified: epithelial cells (EPCAM), T cells (CD3D), B cells (MS4A1), monocytes (CD14), proliferating cells (MKI67), and other stromal/immune type cells.

The epithelial cluster (cluster 0 in figure 4) has a variable expression and proliferation and stress markers. This underscores intratumoral heterogeneity and potential tumor subclones.

Immune infiltration is indicated by presence of adaptive immune cells (B, T cells), and innate cells (monocytes). This indicates an active immune response. This could guide immunotherapy.

MKI67 suggests a highly proliferative tumor. So, anti-proliferative drugs would be potent in this case.

**Future direction:**Multi-omic integration by combining scRNA-seq with scATAC-seq or spatial transcriptomics could help map regulation of genes.

**References**

1. **https://pmc.ncbi.nlm.nih.gov/articles/PMC2718421/#:~:text=Non%E2%80%93small%20cell%20lung%20cancer%20accounts%20for%2085%25%20of%20all,crucial%20for%20determining%20appropriate%20therapy.**