Analysis of Squamous Lung Cell Carcinoma Progression Using an Integrative Bioinformatics Approach

Archana Balasubramanian , Shivaramakrishna Srinivasan ,Aaryan Sharma, Aathira Sarath Chandran, Hersh Gupta

I. Abstract

This study integrates single-cell RNA sequencing and spatial transcriptomics to provide a more nuanced understanding of the heterogeneity of squamous lung cell carcinomas. With the implementation of trajectory analysis, we were able to provide a visual representation of cellular populations representing each cell state and be able to distinguish specific points of physiological versus pathological conditions. This study focuses on identifying the key biomarkers responsible for the onset of squamous lung cell carcinoma and pinpointing areas of divergence.

II. Introduction

Lung cancer is an ongoing health concern that persists throughout the world, with its high number of occurrences and mortality rate posing a challenge. More particularly, squamous cell carcinoma of the lungs is the most common form of lung cancer that arises from the squamous epithelial cells lining the bronchial tubes. Despite further research being done for the ongoing development of therapeutics and diagnostics, squamous lung cell carcinoma continues to present challenges that warrant further understanding of its intricacies.

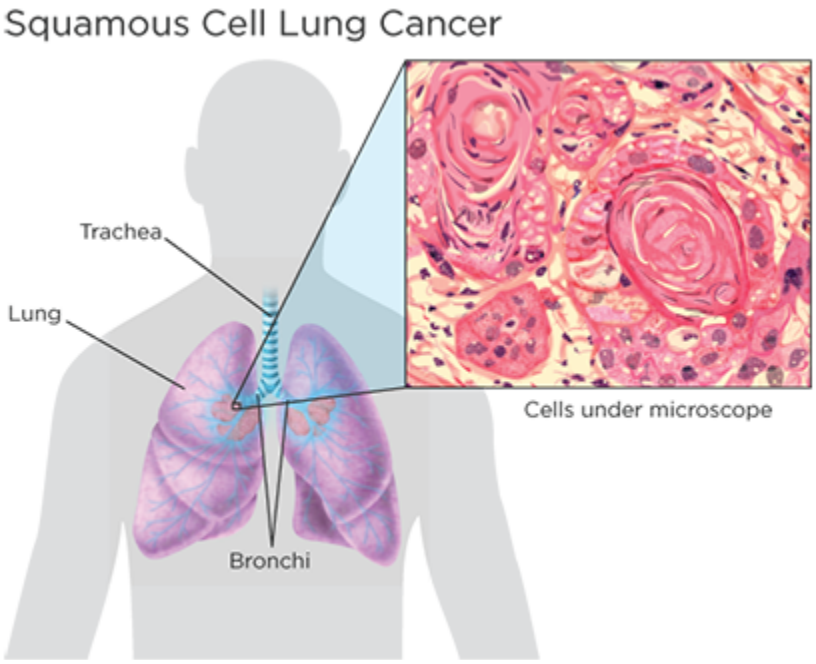

Figure 1. Visual of Squamous lung cell carcinoma; note that the surrounding purple represents keratin and has more of an abnormal shape in comparison to normal lung cell morphology.

Squamous lung cell carcinoma, as mentioned above, is a type of non-small cell cancer that originates from the epithelial lining of the bronchial tubes. Those are the main cells at the lining of the airways within the lungs. It is a widespread form of lung cancer, of which 1 in 3 people diagnosed with lung cancer would be categorized under squamous lung cell carcinoma. A main cause of SLCC is cigarette smoking and/or second-hand smoking, where common symptoms include persistent coughing, coughing up blood, chest pain, and shortness of breath. To get further confirmation in terms of diagnosis, epithelial cells can be taken for analysis via a CT scan. If there are indications of keratinization, then the diagnosis can be confirmed. (**Figure 1**)

Tumor heterogeneity, or the coexistence of multiple different cellular populations within a single tumor, represents a big hurdle to overcome when it comes to cancer research and treatment. That is mainly because there are too many nuances to capture such as distinct molecular profiles and expression patterns. To find one gene that may express SCLC and attempt to silence the expression only to find that due to other genes in the background, this attempt eventually being futile brings concern and a need to understand those nuances.

To capture the tumor environment and express its heterogeneous nature, we aim to integrate RNA sequencing analysis along with spatial transcriptomics to shed light on the nuances. With single-cell RNA sequencing, we aim to classify cells diagnosed with squamous lung cell carcinoma identify what specific marker genes they possess through clustering, and apply trajectory analysis to provide an evolution of the lineage of the given cells to compare and contrast cell states (physiological and pathological). Lastly, spatial transcriptomics will be integrated to provide a visualization of the trajectory analysis, therefore providing an understanding of these nuances from which further identification of marker genes can be applied for squamous lung cell carcinoma diagnosis.

III. Methods

A. Data Acquisition and Processing - Single-cell RNA Sequencing

The single-cell RNA sequencing dataset was obtained from the Human Cell Atlas database [2], from which the lung cancer data was collected from the following sources:

1. **Barbry_unpubl:** Participants recruited from the Pneumology Unit of Nice University Hospital provided nasal and tracheobronchial samples, collected between 1 and 15 December 2020. Details of patient inclusion criteria and procedures are available at ClinicalTrials.gov under reference NCT04529993. Libraries were prepared using the v3.1 protocol for 3′ chemistry from 10x Genomics, followed by sequencing on a NextSeq 500/550 sequencer. Data processing involved Cell Ranger 6.0.0 pipeline with GRCh38 reference genome. Filtering, normalization, and clustering were performed using the Scanpy83 pipeline. [2]

2. **Schiller_2021:** Tumor-free lung samples were obtained during tumor resections at Asklepios Fachkliniken München-Gauting. The tissues were processed for scRNA-seq using 10x Genomics Chromium Single Cell 3′ v3.1 gel beads and reagents. Sequencing was performed on a NovaSeq 6000 instrument. Cell Ranger computational pipeline (v3.1.0) was used for data processing with the GRCh38 human reference genome. Quality control involved filtering barcodes and cells based on transcript counts and mitochondrial content. [2]

3. **Duong_lungMAP_unpubl:** Postmortem human lung samples were obtained from BRINDL, supported by the NHLBI LungMAP Human Tissue Core at the University of Rochester. Single nuclei were isolated and processed for scRNA-seq using 10x Chromium Single Cell 3' Reagent Kits v3. Cell Ranger v3 pipeline was utilized for data processing with GRCh38 reference genome. Downstream analysis included filtering, batch correction, and clustering using Pagoda2 and Seurat. [2]

4. **Jain_Misharin_2021:** Nasal epithelial samples were collected from healthy volunteers at Northwestern Medicine. The scRNA-seq libraries were prepared using 10x Genomics Chromium Single Cell 5′ V1 or V2 chemistry. Sequencing was performed on a NovaSeq 6000 instrument. Data processing involved Cell Ranger 3.1.0 pipeline with GRCh38 reference genome. Cell annotations were assigned based on expert interpretation of marker genes using Seurat workflow. [2]

5. **Schultze_unpubl:** Fresh adjacent normal tumor-free lung tissues from patients with non-small cell lung cancer tumors were obtained at Hannover Medical School. The tissues were processed for scRNA-seq using the Seq-Well protocol. Data processing included quality control, filtering, normalization, and clustering using Seurat v3.2 pipeline. [2]

B. Data Acquisition and Processing - Spatial Transcriptomics

The spatial transcriptomics dataset discussed here originates from human lung cancer tissue samples, specifically diagnosed as squamous cell carcinoma, acquired from Avaden Biosciences by 10x Genomics. These tissue samples were carefully preserved using the FFPE method, a standard technique for long-term storage of tissue specimens while maintaining cellular morphology and molecular integrity.

The Visium CytAssist Spatial Gene Expression for FFPE protocol (CG000518) was followed to prepare the tissue sections for analysis. This involved sectioning the tissue into thin slices (5

μm), mounting them onto glass slides, and then subjecting them to H&E staining after deparaffinization. The stained sections were subsequently imaged using an Olympus VS200 Slide Scanning Microscope equipped with specific settings to capture detailed images for downstream analysis. Following imaging, the tissue sections underwent a series of processing steps, including coverslipping with 85% glycerol, decoverslipping, dehydration, and decrosslinking, to prepare them for spatial gene expression analysis. This analysis was facilitated by the Visium CytAssist instrument, which transferred analytes to a Visium CytAssist Spatial Gene Expression slide.

Sequencing libraries were prepared according to the Visium CytAssist Spatial Gene Expression Reagent Kits for FFPE User Guide (CG000495). Subsequently, the libraries were sequenced using an Illumina NovaSeq platform, with specific sequencing configurations and depths optimized for spatial transcriptomics analysis.

The resulting dataset provides valuable insights into the spatial distribution of gene expression within the lung cancer tissue, shedding light on the molecular mechanisms underlying squamous cell carcinoma progression and potentially informing therapeutic strategies. Key metrics, such as the number of spots detected under tissue, median genes per spot, and median UMI counts per spot, underscore the richness of the spatial transcriptomics data generated from this study.

C. Single-cell RNA Sequencing

Initial Filtering

The raw data of the single-cell RNA sequencing via the Lung Cell Molecular Atlas was subsetted using the Python libraries Squidpy and Scanpy. Initially, the data is loaded from the h5ad file into an AnnData object using Scanpy's 'read_h5ad' function. The function 'metadata_filter' is then defined to facilitate quick filtering and subsetting of the dataset based on metadata. The 'metadata_filter' function was implemented to allow filtering of the cells based on their associated disease status or other metadata attributes.

From the metadata_filter function, two subsets of the original dataset were created: one subset consisting of cells in the baseline physiological state and another with all the cells classified as squamous lung cell carcinoma. The normal cell dataset itself was filtered to include those above the age of 55 via 'normalCellsabv55' as according to the Thienpont 2018 paper where the squamous lung cell carcinoma was collected, the samples were collected from three people aged 86, 70, and 55 respectively. The gene names in the subsets were also indexed by their ENSEMBL IDs for better identification.

The filtered subsets were then written to separate h5ad files for further downstream analysis. The filtered normal cells were stored in the h5ad file ‘normalCellsabv55.h5ad’ while the squamous cells were stored in the h5ad file ‘sclcCellsNew.h5ad.

Preprocessing and Analysis

For preprocessing, the datasets generated from the filtration and subsetting were loaded using the Scanpy library, and a custom function called ‘quality control’ is defined to compute quality control metrics. This function entails filtering out any low-quality cells and genes based on specific thresholds along with any mitochondrial or ribosomal genes in the dataset. Following quality control, normalization was performed to make sure uniform counts were established across the cells, and highly variable genes were identified to focus the analysis on our areas of interest. Data visualization through violin plots helped in analyzing the spread of quality control metrics across cells in both normal (control) and squamous cell datasets. In addition, outlier cells were further filtered out based on those metrics, and high-variable genes were subsetted including only the top genes with the highest variability.

Following preprocessing to get the dataset ready for downstream analysis, we progressed into dimensionality reduction and clustering analysis using Scanpy. To start off, highly variable genes were selected for each dataset, and unwanted sources of variation such as the mitochondrial/ribosomal genes were regressed out. The data was then scaled for comparability for principal component analysis (PCA) to be performed to reduce the dimensionality of the datasets. Visualization of the PCA variance ratios helps assess the variance proportion. Afterward, neighborhood graphs were constructed based on the PCA-reduced space with uniform manifold approximation and projection (UMAP) applied to capture the structure of the datasets. Clustering was performed using the Leiden algorithm, with the resolution parameter set at 0.5 to control granularity. UMAP plots were created with colors generated from cluster identity and cell type to provide insights into the distinct cell population.

Once clustering has been done, further refinement of the squamous cell lung carcinoma (SCLC) dataset is conducted by removing specific cell types such as myofibroblast cells and multi-ciliated epithelial cells. The process makes sure that the subsequent analyses are performed on relevant cell populations. Additionally, differential gene expression analysis is performed using the ‘rank_genes_group’ function to identify the targeted marker genes associated with different cell types for both normal and squamous cell populations. Visualizations through ‘sc.pl.rank_genes_groups’ provide insight into the top differentially expressed genes across cell types. Based on this, the identified marker genes are pulled and saved into CSV files for further analysis. Next, Partition-based Graph Abstraction (PAGA) and draw graph functions were applied for trajectory analysis to infer cell-cell communication and visualize cellular

relationships within the datasets via ‘sc.pl.paga’ and ‘sc.pl.draw_graph’. This helped provide insights into the overall structure and connectivity patterns of cell types. This overall workflow helped enable a comprehensive exploration of cellular heterogeneity to understand the pathways of evolution for the cluster of cells present.

Cell-Cell Network Validation - LIANA

To validate the cell-cell communication among the cells, we implemented the Linking Interaction Analysis for Network-based Algorithms (LIANA) framework. Initially, the scRNA-seq data for squamous cell lung carcinoma (SCLC) cells is loaded and processed using Scanpy. Subsequently, LIANA's **rank_aggregate()** method is employed to analyze cell-cell communication patterns within the dataset. This method allows for the identification of ligand-receptor interactions between different cell types, providing insights into potential signaling mechanisms underlying cellular interactions. The resulting analysis is visualized using LIANA's dotplot function, which visualizes ligand-receptor interactions between specified source and target cell types.

Additionally, the rank_aggregate function from LIANA is applied to perform rank aggregation, a method used to combine rankings from multiple sources into a single consensus ranking. In this context, the rank_aggregate function is utilized to aggregate rankings based on specific cell types or other categorical variables, facilitating the integration of heterogeneous data sources and improving the robustness of downstream analyses.

D. Spatial Transcriptomics

Data Acquisition and Preprocessing

Spatial transcriptomics data was acquired from a publicly acquired dataset from 10X Genomics called "CytAssist_FFPE_Human_Lung_Squamous_Cell_Carcinoma_spatial". The dataset consisted of FFPE Human Lung Cancer tissue from Avaden Biosciences which was prepared and processed according to Visium protocols for spatial gene expression analysis, including H&E staining, imaging, and analyte transfer using the Visium CytAssist instrument for probe extension and library construction. The tissue was diagnosed as Lung Squamous Cell Carcinoma. The data was initially processed using the datasets.visium_sge() function to load the dataset and return as an AnnData object, which included necessary attributes for downstream analysis (e.g., counts, spatial coordinates).

Quality Control and Filtering

Initial quality control was performed to remove cells and genes based on total counts, expressed genes, and mitochondrial content thresholds. Specifically, cells with extreme total counts or a

high percentage of mitochondrial reads were filtered out. Genes detected in a minimal number of cells were also excluded. This step ensured the removal of outliers and improved the homogeneity of the dataset for subsequent analysis.

Normalization and Variable Gene Selection

Normalized gene expression data were obtained through total count normalization, followed by logarithmic transformation. Highly variable genes were identified using the Seurat method, facilitating the focus on genes most informative for identifying cell types or states.

Dimensionality Reduction and Clustering

Principal Component Analysis (PCA) was performed to reduce the dimensionality of the data, capturing the main axes of variation. The PCA representation served as the basis for constructing a neighborhood graph, which, in turn, facilitated the clustering of cells using the Leiden algorithm. This step categorized cells into clusters based on their gene expression profiles, highlighting the transcriptomic diversity within the tissue.

Spatial Mapping and Visualization

The spatial distribution of gene expression was visualized by projecting the data onto the original tissue coordinates. This approach enabled the exploration of the spatial organization of cell clusters and the identification of spatially variable genes, crucial for understanding tissue architecture and cellular interactions.

Identification of Spatially Variable Genes

Spatially variable genes were identified using a statistical framework designed for spatial transcriptomics data, facilitating the discovery of genes whose expression patterns showed significant spatial variation. This analysis provided insights into the spatial heterogeneity of gene expression within the tissue, revealing potential spatially restricted biological processes or cell-to-cell communication pathways.

E. Data Integration

Data Acquisition and Preprocessing

Both the single-cell RNA sequencing (scRNA-seq) data and spatial transcriptomics data were processed using Scanpy's `sc.read_h5ad` and `sc.read_visium` functions, respectively.

Quality Control and Metric Calculation

Quality control metrics were calculated for the spatial transcriptomics dataset to assess data integrity and prepare it for downstream analyses. Metrics including gene counts by cells and total counts were computed with `sc.pp.calculate_qc_metrics`, and the data were then visualized using violin plots to examine distributions of gene and transcript counts across cells, facilitating the identification and removal of potential outliers.

Data Integration

The Scanorama' correct_scanpy' method was employed to integrate the single-cell and spatial transcriptomics datasets. This integration aimed to correct batch effects between the datasets while retaining their unique biological signals. The integrated data were stored in a unified AnnData object, enabling comprehensive analysis across datasets.

Label Transfer and Cell Type Identification

Following integration, a label transfer approach was utilized to assign cell type identities from the scRNA-seq dataset to the spatial transcriptomics data based on cosine distance calculations. This method facilitated the projection of cell type annotations from single-cell data onto spatially resolved data, enhancing the interpretation of spatial distributions of cell types. The transfer was performed using a custom label transfer function, which assigned probabilities to each cell type based on their distances in the integrated space, resulting in a DataFrame that contains the probability of each cell belonging to the scRNA-seq-derived cell types.

Spatial Visualization

The cells' spatial distribution and inferred cell types were visualized on the tissue's high-resolution image. This visualization used Scanpy's `sc.pl.spatial` function, highlighting the locations and densities of different cell types within the tissue context. The `img_key="hires"` parameter ensured that the highest resolution of the tissue image was used for overlaying spatial gene expression data.

Cell Type Consolidation

A categorical column `cell_type` was created to consolidate cell type annotations by selecting the maximum probability cell type for each cell from the label transfer output. This consolidated cell type information was used to generate a final spatial plot, illustrating the distribution and localization of various cell types within the tissue, and providing insights into the tissue architecture and cellular microenvironment.

IV. Results

A. Single-Cell RNA Sequencing

The comprehensive preprocessing applied to the single-cell RNA sequencing data successfully ensured high data quality for downstream analysis. Initial quality control metrics were computed to remove any sort of noise along with factoring out any mitochondrial and ribosomal genes in the dataset. The violin plots provided visual insights into the distribution of quality control metrics across cells in both normal and squamous cell lung carcinoma conditions. (**Figure 2**). The data was grouped into highly variable genes and others to get an insight to the nature of the data before further downstream analysis. (**Figure 3**)

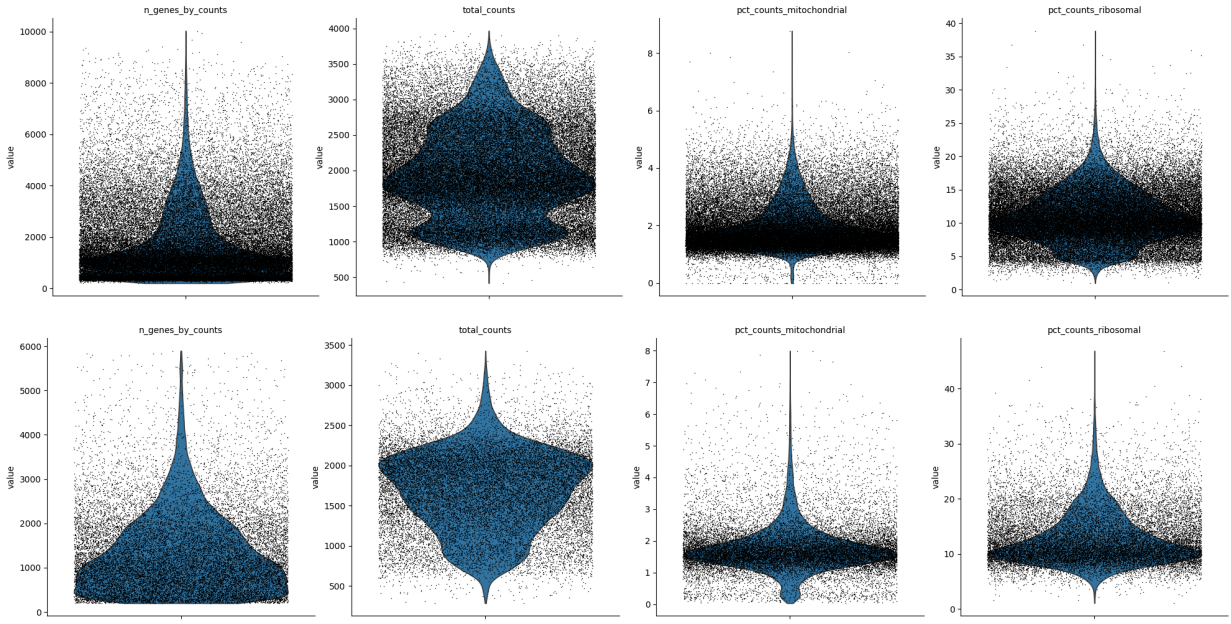

Figure 2: Violin plots illustrating the distribution of gene expression metrics for **normal (no disease condition) cells (top) and squamous cell lung carcinoma cells (bottom)** in the dataset including the number of genes by counts, total counts, percentage of counts attributed to mitochondrial genes, and percentage of counts attributed to ribosomal genes across single-cell samples, with jittering applied for improved visualization. Panels are organized into multi-panel format."

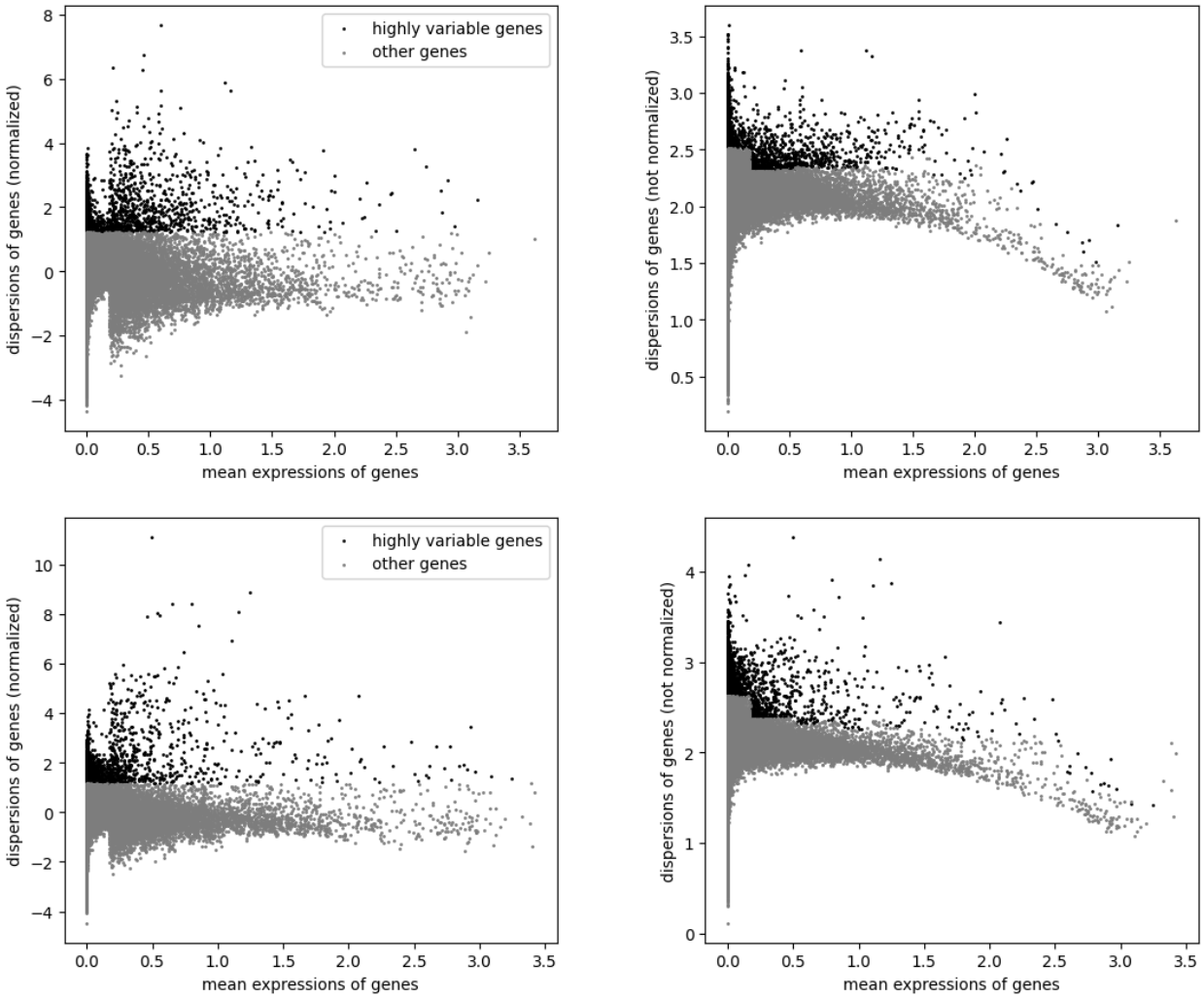

Figure 3: Identification of highly variable genes in the normal cells (no disease condition) (top panel) and cells with squamous cell lung carcinoma (bottom panel), with the top 2000 genes selected. The plot displays the gene expression variance versus mean expression, highlighting genes deemed highly variable for downstream analysis.

Dimensionality reduction via principal component analysis (PCA) and PCA variance ratio visualization effectively reduced the dimensionality of the datasets while focusing on the majority of the variance. (**Figure 4**) From the PCA variance ratio, we set the UMAP graph to include PCA ratios until PCA30 based on visual analysis of the elbow.

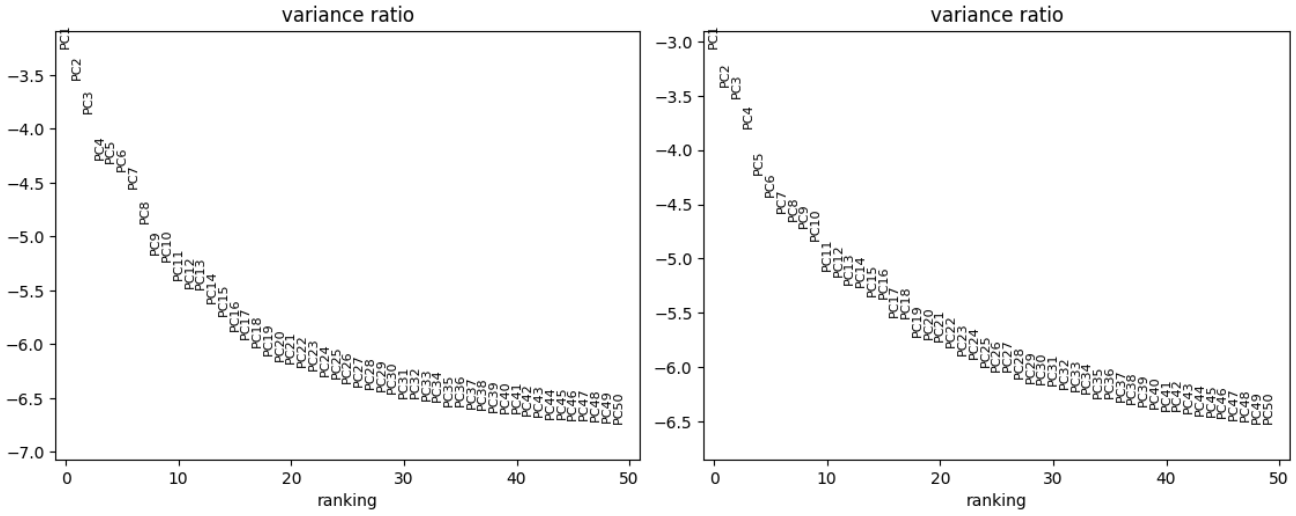

Figure 4: Principal Component Analysis (PCA) applied to normal cells (left) and squamous cell lung carcinoma (right) data after regressing out effects of total counts, percentage of counts from mitochondrial genes, and percentage of counts from ribosomal genes, followed by scaling. The plot displays the variance ratio explained by each principal component, with the logarithmic scale applied, considering the top 50 principal components.

Subsequent clustering analysis utilizing the Leiden algorithm revealed distinct cell populations within both the squamous cell lung carcinoma (SCLC) and control datasets, as depicted in **Figure 5**. The visualization of the UMAP plots colored by cluster identity and cell type highlighted the presence of identifiable cell clusters in both datasets. Notably, there appeared to be fewer clusters in the SCLC dataset compared to the control dataset, suggesting potential differences in cellular heterogeneity between the two conditions. Additionally, a prominent cluster aggregation was observed around the central area of the UMAP plot, indicative of potential inter-cluster connections or cellular transitions. This clustering pattern may signify shared biological processes or functional relationships among cell populations, warranting further investigation into the underlying mechanisms driving cellular interactions and dynamics within the analyzed samples.

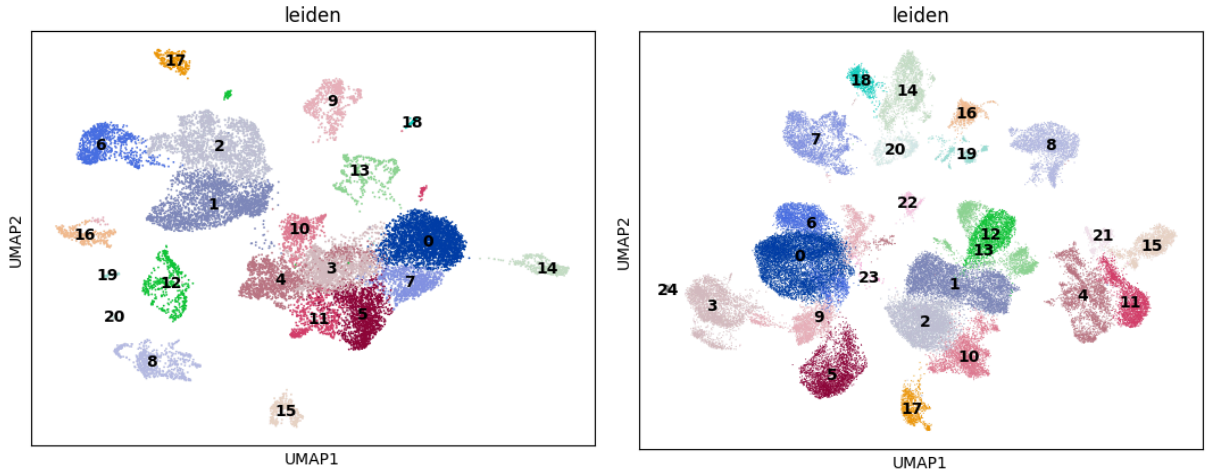

Figure 5: UMAP visualization of the lung cell atlas dataset annotated with Leiden clusters (resolution = 0.5) and cell types for both squamous cell lung carcinoma (left, top) and normal (right, top).

The clustering analysis was reiterated to provide a comprehensive overview of all cell types present in the dataset, resulting in **Figure 6.**

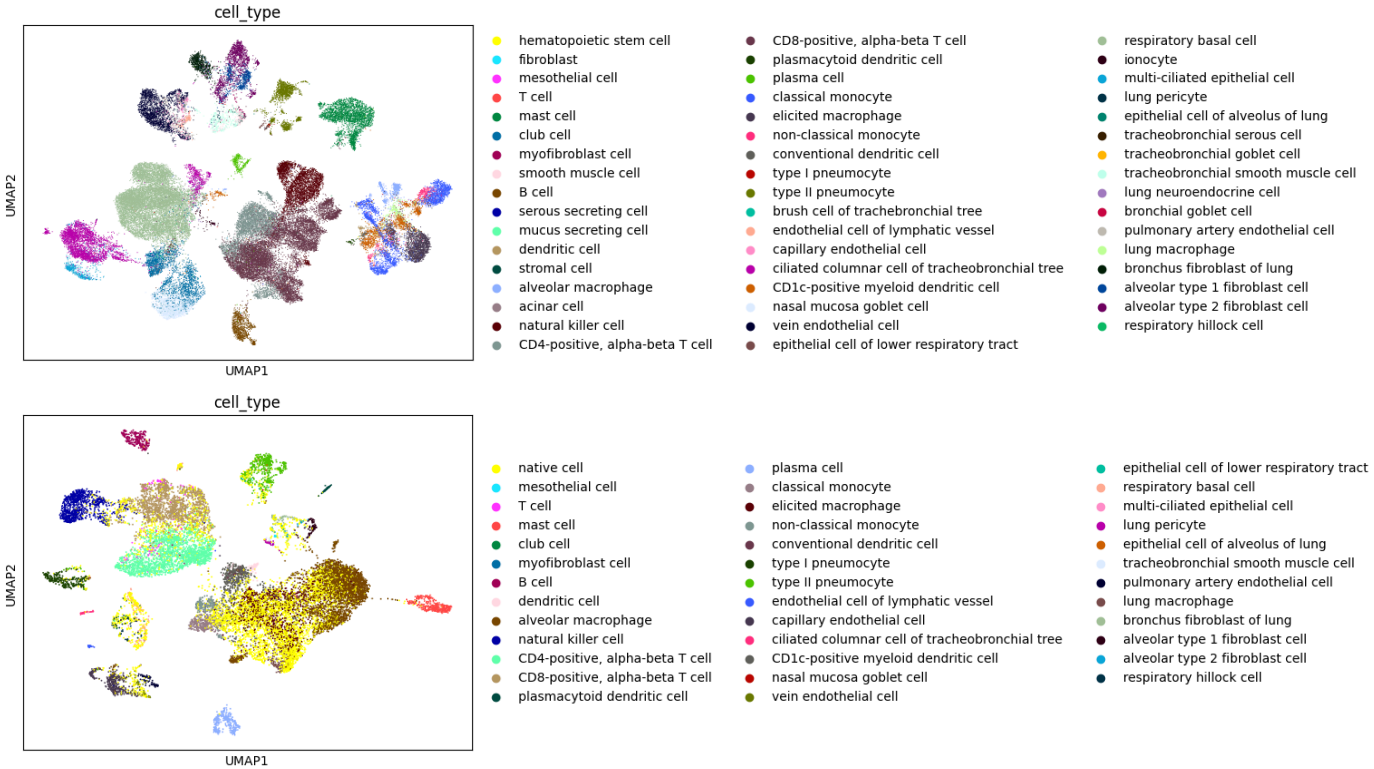

Figure 6: UMAP visualization of the lung cell atlas dataset annotated with original cell type annotations for both control/normal condition (top) and squamous cell lung carcinoma (bottom).

The visualization revealed several distinct clusters, including a predominant population of native cells and unique clusters of prominent categories of cells such as CD4+ T cells, CD8+ T cells. Additionally, the UMAP plot displayed a mixture of clusters in the central area, indicating the presence of diverse cell populations with potentially overlapping or intermediate phenotypes. This expanded clustering analysis offers a more detailed understanding of the cellular composition and heterogeneity within the dataset, highlighting the complexity of the biological system under investigation. The identification of specific cell clusters provides valuable insights into the cellular diversity and potential functional roles of different cell types in the studied biological context.

With further refinement via the removal of the aforementioned specific cell types, differential gene expression analysis identified marker genes associated with different cell types for the squamous and normal datasets respectively using the t-testing statistical method. (**Tables 1 and 2**) Each dataset was then filtered based on specific p-values and log-fold changes to lead to the final datasets for squamous lung cell carcinoma and normal conditions for marker analysis (**Tables 3 and 4**) The trajectory analysis for all cell types of the squamous dataset were graphed (**Figure 7**) which revealed various cell-cell communication patterns that seem to stem from the native cell.

	0	1	2	3	4
hematopoietic stem cell	DUT	SNHG8	COMMD6	BTF3	SRGN
fibroblast	IFITM3	C1R	NDUFAF3	IGFBP4	FBLN1
mesothelial cell	IFITM2	NAMPT	SSU72	MRPS21	TMEM50A
T cell	STMN1	HMGB2	HMGN2	CORO1A	TUBB
mast cell	TPSB2	TPSAB1	CPA3	VIM	GATA2
club cell	WFDC2	KRT19	SLPI	VMO1	AGR2
myofibroblast cell	LGALS1	VIM	COL1A2	CLU	TAGLN
smooth muscle cell	TAGLN	ACTA2	MYL9	TPM2	CALD1
B cell	CD74	MS4A1	CXCR4	RPS27	HLA-DQB1
serous secreting cell	PHLDA1	LPO	EIF2B3	DHRSX	THAP7
mucus secreting cell	BPIFB2	AZGP1	MUC5B	TFF3	AGR2
dendritic cell	GPX4	CD83	PFN1	TXN	BIRC3
stromal cell	TAGLN	ACTA2	MYL9	TPM2	ACTG2
alveolar macrophage	C1QA	FCER1G	C1QB	TYROBP	APOC1
acinar cell	WFDC2	SLPI	SCGB3A1	KRT19	PIGR
natural killer cell	NKG7	GNLY	PRF1	GZMB	KLRD1
CD4-positive, alpha-beta T cell	CXCR4	BTG1	IL7R	IL32	TRAC
CD8-positive, alpha-beta T cell	CCL5	IL32	B2M	CD3D	TRAC
plasmacytoid dendritic cell	CD74	GZMB	RPS11	SEC61B	JCHAIN
plasma cell	JCHAIN	MZB1	SSR4	IGKC	DERL3
classical monocyte	S100A9	TYROBP	LYZ	S100A8	FCER1G
elicited macrophage	TYROBP	CXCL8	SOD2	FCER1G	VIM
non-classical monocyte	FCER1G	TYROBP	AIF1	LST1	SRGN
conventional dendritic cell	HLA-DQB1	CD74	CST3	VIM	SNX3
type I pneumocyte	S100A10	EMP2	AGER	DSTN	NAPSA
type II pneumocyte	SFTPB	NAPSA	SFTPC	SFTPA1	SFTPA2
brush cell of trachebronchial tree	MARCKSL1	KHDRBS1	SKP1	AZGP1	VDAC1
endothelial cell of lymphatic vessel	CCL21	IGFBP7	GNG11	VIM	TFPI
capillary endothelial cell	SPARCL1	GNG11	TMSB10	AQP1	IFITM3
ciliated columnar cell of tracheobronchial tree	CAPS	C20orf85	RSPH1	LRRIQ1	C9orf24
CD1c-positive myeloid dendritic cell	CD74	HLA-DQB1	TYROBP	CST3	AIF1
nasal mucosa goblet cell	WFDC2	LCN2	S100P	LYPD2	VMO1
vein endothelial cell	SPARCL1	IGFBP7	AQP1	ACKR1	VIM
epithelial cell of lower respiratory tract	WFDC2	CYB5A	GSTP1	SCGB3A1	SLPI
respiratory basal cell	KRT17	KRT19	KRT15	S100A2	CD9
ionocyte	RARRES2	APLP2	CD9	CD24	SEC11C
multi-ciliated epithelial cell	CAPS	WFDC2	LRRIQ1	TPPP3	C20orf85
lung pericyte	GPX3	A2M	IGFBP7	COX4I2	CALD1
epithelial cell of alveolus of lung	SCGB3A2	SFTPB	NPC2	CYB5A	NBEAL1
tracheobronchial serous cell	LYZ	LTF	SLPI	WFDC2	PIGR
tracheobronchial goblet cell	MUC5B	C3_ENSG00000125730	TFF3	AGR2	PIGR
tracheobronchial smooth muscle cell	TAGLN	IGFBP7	CALD1	MYL9	ACTA2
lung neuroendocrine cell	BEX2	GRP	PCSK1N	AZGP1	TMED10
bronchial goblet cell	WFDC2	CXCL17	BPIFB1	PIGR	SCGB1A1
pulmonary artery endothelial cell	TM4SF1	IFITM3	CAV1	CLDN5	RAMP2
lung macrophage	CD74	TYROBP	C1QA	C1QC	C1QB
bronchus fibroblast of lung	DCN	MGP	FBLN1	LUM	SPARCL1
alveolar type 1 fibroblast cell	DCN	MGP	LGALS1	IFITM3	C1R
alveolar type 2 fibroblast cell	DCN	MGP	TIMP1	FBLN1	LUM
respiratory hillock cell	S100A2	GSTP1	GAPDH	SERPINB3	LY6D

Table 1: Ranking of genes associated with specific cell types in the lung cell atlas dataset for the squamous cell lung carcinoma condition, determined using the t-test metric. The table displays the top 5 ranked genes (indexed from 0-4, 0 being the highest) for each cell type (top 20 were obtained)

	0	1	2	3	4
native cell	RNASE1	CTSB	CD68	CD14	SPP1
mesothelial cell	PLA2G2A	TM4SF1	HP	SLPI	ITLN1
T cell	STMN1	HMGB2	H2AZ1	IL32	TUBB
mast cell	TPSB2	TPSAB1	CLU	CPA3	RGS1
club cell	FXYD3	AKR1C1	AKR1C3	CCND1	GPX2
B cell	CD74	RPS3	RPL13	CD79A	HLA-DQB1
dendritic cell	BIRC3	GPX4	BTG1	CD74	MARCKSL1
alveolar macrophage	C1QA	APOC1	C1QB	FABP4	MARCO
natural killer cell	NKG7	PRF1	GZMB	KLRD1	CST7
CD4-positive, alpha-beta T cell	IL32	TRAC	CD3D	BTG1	LTB
CD8-positive, alpha-beta T cell	CCL5	IL32	GZMA	TRAC	NKG7
plasmacytoid dendritic cell	CXCR4	PLAC8	PLD4	GZMB	LILRA4
plasma cell	MZB1	FKBP11	XBP1	SSR4	IGKC
classical monocyte	S100A8	LYZ	S100A9	TYROBP	AIF1
elicited macrophage	TYROBP	CSTB	CTSB	CD68	LGALS1
non-classical monocyte	AIF1	LST1	FCER1G	TYROBP	SAT1
conventional dendritic cell	CPVL	C1orf54	HLA-DQB1	SNX3	CST3
type I pneumocyte	AGER	EMP2	CAV1	KRT7	FXYD3
type II pneumocyte	SFTPB	SFTPA2	NAPSA	SFTPA1	SLPI
endothelial cell of lymphatic vessel	CCL21	IGFBP7	GNG11	CAVIN2	RAMP2
capillary endothelial cell	CLDN5	IFITM3	TMEM100	B2M	RAMP2
ciliated columnar cell of tracheobronchial tree	TMEM190	HOATZ	CAPS	C9orf24	TPPP3
CD1c-positive myeloid dendritic cell	HLA-DQB1	CD74	HLA-DRB5	CST3	CTSH
nasal mucosa goblet cell	TPT1	RPL27	RPS12	RPL21	TDRP
vein endothelial cell	IGFBP7	TM4SF1	IFITM3	ACKR1	VWF
epithelial cell of lower respiratory tract	WFDC2	SCGB3A2	SFTPB	SFTPA2	CYB5A
respiratory basal cell	HSPB1	PERP	KRT19	KRT15	MIF
lung pericyte	BGN	IGFBP7	PTN	SPARCL1	A2M
epithelial cell of alveolus of lung	FOLR1	SFTPB	HSBP1L1	ST3GAL5	CYB5A
tracheobronchial smooth muscle cell	ACTA2	PPP1R14A	TPM2	TAGLN	SPARC
pulmonary artery endothelial cell	CLDN5	DEPP1	IFITM3	MGP	MT2A
lung macrophage	CST3	TYROBP	MS4A6A	C1QA	CD14
bronchus fibroblast of lung	BGN	COL3A1	COL1A1	COL1A2	COL6A2
alveolar type 1 fibroblast cell	RARRES2	DCN	A2M	LUM	MFAP4
alveolar type 2 fibroblast cell	C1R	TIMP1	DCN	MT2A	COL1A2
respiratory hillock cell	KRT6A	HSPB1	S100A11	KRT17	S100A2

Table 2: Ranking of genes associated with specific cell types in the lung cell atlas dataset for the **normal (no disease)** condition, determined using the t-test metric. The table displays the top 5 ranked genes (indexed from 0-4, 0 being the highest) for each cell type (top 20 were obtained)

	group	names	scores	logfoldchanges	pvals	pvals_adj
0	native cell	RNASE1	38.676357	1.932592	4.225046e-302	8.956675e-298
1	native cell	CTSB	36.607067	1.275711	7.102813e-277	7.528627e-273
2	native cell	CD68	34.104832	1.314612	4.680133e-242	3.307138e-238
3	native cell	CD14	33.905663	1.617134	5.814181e-237	3.081370e-233
4	native cell	SPP1	33.428078	1.870765	2.347294e-229	9.952057e-226
...	...	...	...	...	...	...
742432	respiratory hillock cell	VPS25	2.306676	1.333766	2.485431e-02	4.828064e-02
742433	respiratory hillock cell	SENP5	2.304311	1.957749	2.500051e-02	4.853794e-02
742434	respiratory hillock cell	CDK4	2.303652	1.208497	2.503381e-02	4.859814e-02
742437	respiratory hillock cell	TIMM8A	2.297578	2.600601	2.541156e-02	4.925989e-02
742438	respiratory hillock cell	TNS4	2.294225	5.311518	2.561954e-02	4.963977e-02

30203 rows × 6 columns

Table 3: Marker genes identified using the rank_genes_groups() function from single-cell data analysis for **squamous cell lung carcinoma**. Marker genes were filtered based on adjusted p-values (<0.05) and log-fold changes (>1), highlighting genes significantly associated with specific cell types or conditions.

	group	names	scores	logfoldchanges	pvals	pvals_adj
0	hematopoietic stem cell	DUT	123.528229	3.296336	0.000012	0.000019
1	hematopoietic stem cell	SNHG8	54.172615	2.547828	0.000215	0.000323
2	hematopoietic stem cell	COMMD6	46.221798	1.633149	0.000216	0.000324
3	hematopoietic stem cell	BTF3	43.502277	1.382027	0.000255	0.000381
4	hematopoietic stem cell	SRGN	43.147533	2.126544	0.000303	0.000451
...	...	...	...	...	...	...
1251795	respiratory hillock cell	TMEM258	2.385739	1.435117	0.031716	0.045686
1251796	respiratory hillock cell	FGFBP1	2.380333	3.928392	0.032052	0.046141
1251797	respiratory hillock cell	CHCHD2	2.373320	1.106016	0.032476	0.046711
1251799	respiratory hillock cell	PHLDA3	2.339743	2.774838	0.034631	0.049622
1251800	respiratory hillock cell	UNC5B-AS1	2.336436	3.742538	0.034851	0.049917

73314 rows × 6 columns

Table 4: Marker genes identified using the rank_genes_groups() function from single-cell data analysis for **normal cells**. Marker genes were filtered based on adjusted p-values (<0.05) and log-fold changes (>1), highlighting genes significantly associated with specific cell types or conditions.

The thickness of the connections seems to represent how strong the connections among the specific cells are, as some connections such as the lung and alveolar macrophages plus the T cell and natural killer cells show that relationship. There are a lot of other connections as clearly seen in the trajectory analysis but considering that these different cell types are all contributing to

squamous lung cell carcinoma, it gives reason to prove the connections. With the normal dataset trajectory analysis (**Figure 8**), it is a much more cluttered web of connections that show some form of interlinking among the cell types. The randomness of this compared to a more specific approach can help place emphasis on the differences in trajectory analysis between diseased and control states.

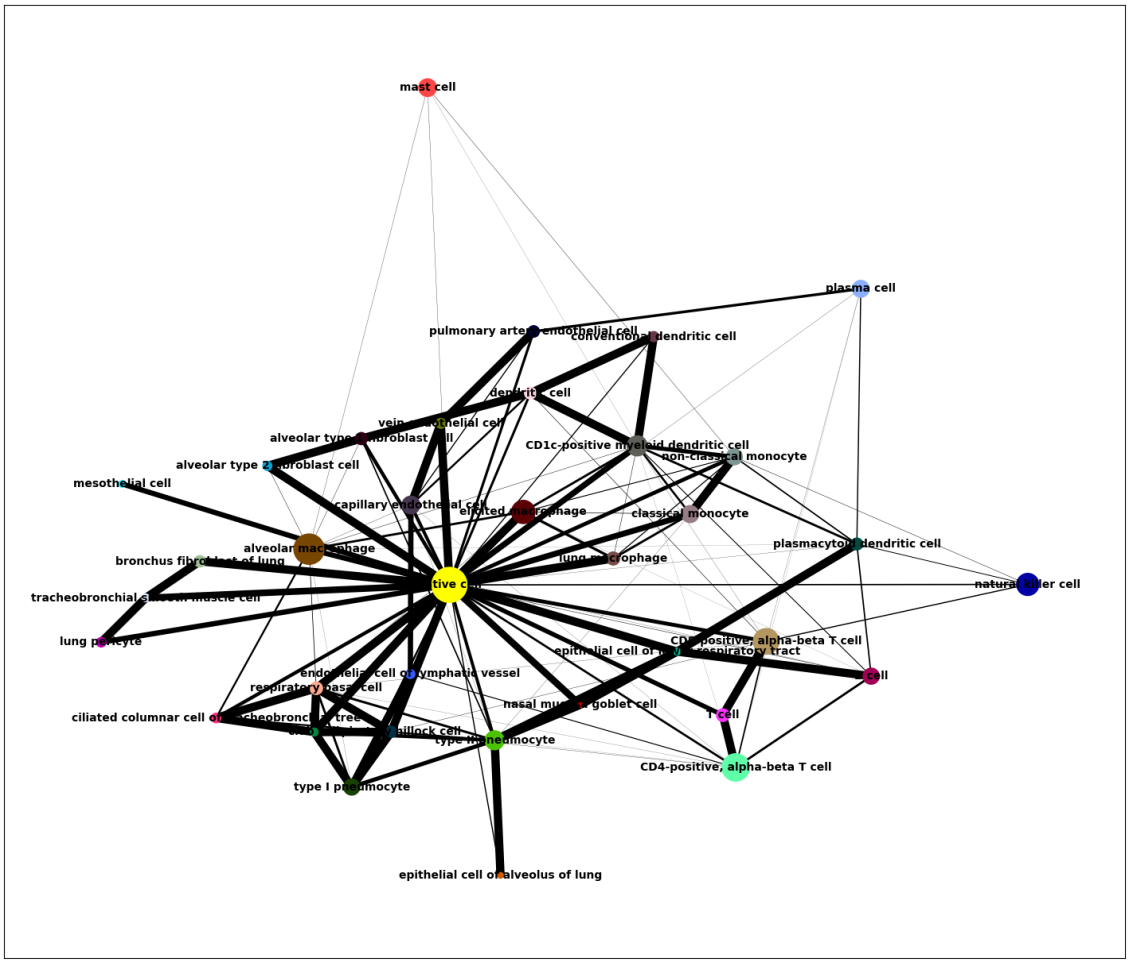

Figure 7: Partition-based Graph Abstraction (PAGA) analysis performed on the single-cell lung cancer (SCLC) dataset, with cells grouped according to their cell types, for the squamous cell lung carcinoma data. The resulting PAGA graph is visualized with nodes colored by cell type, utilizing a larger figure size for enhanced visualization.

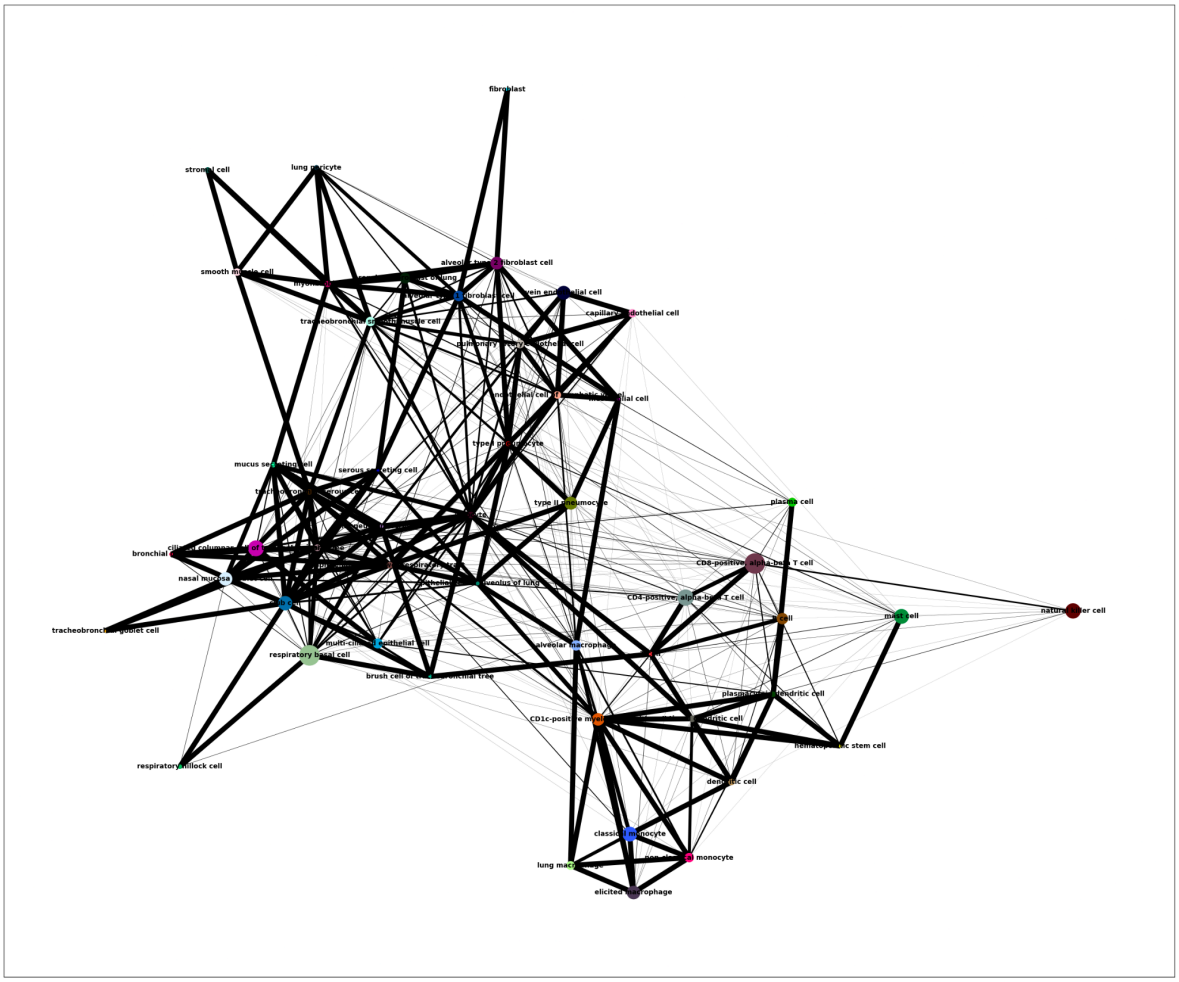

Figure 8: Partition-based Graph Abstraction (PAGA) analysis performed on the single-cell lung cancer (SCLC) dataset, with cells grouped according to their cell types, for the **normal cells'** data. The resulting PAGA graph is visualized with nodes colored by cell type, utilizing a larger figure size for enhanced visualization.

B. Spatial Transcriptomics and Integration

Quality control measures applied to the spatial transcriptomics data led to the refinement of the dataset, as illustrated in **Figure 9**. Violin plots representing the distribution of total gene counts and expression levels across the dataset were generated. Cells with extreme total counts, which could indicate doublets or other aberrations, and those with high mitochondrial content, often a sign of cellular stress or death, were systematically excluded. Additionally, genes that were expressed in a minimal number of cells were removed to eliminate background noise and focus on biologically relevant signals. These filtering steps resulted in a more homogeneous dataset, thus ensuring a more reliable basis for subsequent analyses of spatial gene expression patterns.

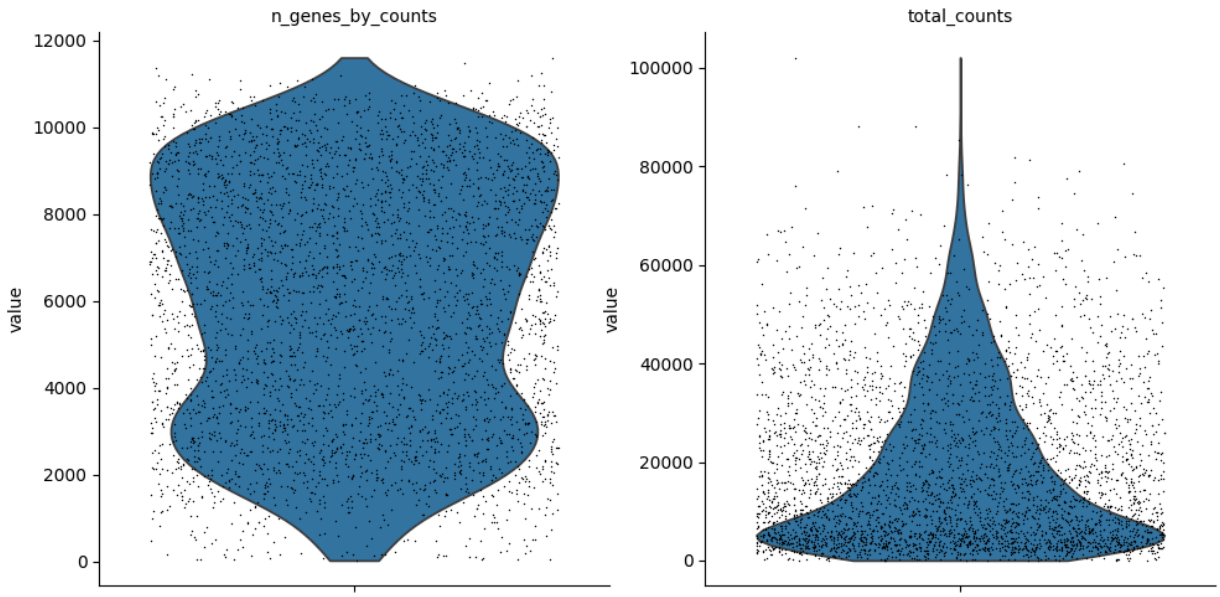

Figure 9: Quality Control Metrics for Spatial Transcriptomics Data. Violin plots depict the distribution of (A) the number of genes detected by counts (n_genes_by_counts) and (B) the total expression counts (total_counts) across spatially resolved transcriptomic spots. The shape of the distributions provides insights into data quality, with the breadth of each plot indicating the variability within the dataset. Such plots are instrumental in identifying outliers and determining thresholds for data filtering to ensure robust downstream analysis.

Spatial transcriptomics analysis revealed distinct spatial distributions for various quality control metrics within the tissue sample. The visualization of 'in_tissue' indicates areas of the tissue where transcriptomic data was captured. The 'n_genes_by_counts' highlights regions with varying gene detection levels, reflecting potential cellular complexity or density. The 'log1p_n_genes_by_counts' and 'total_counts' plots elucidate the expression intensity and the extent of transcriptomic activity, respectively. These spatial maps allow for the observation of transcriptomic heterogeneity across different tissue sections, which is fundamental for understanding the molecular architecture of the sample as seen in **Figure 10.**

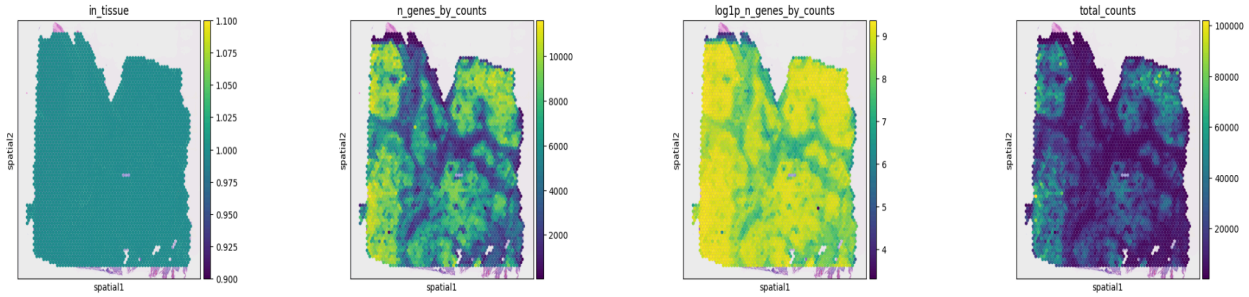

Figure 10: Spatial distribution of transcriptomic data quality control metrics. (A) 'in_tissue' displays areas where transcripts were detected, (B) 'n_genes_by_counts' shows the number of genes detected per spot, (C) 'log1p_n_genes_by_counts' presents the log-transformed values to normalize gene count distribution, and (D) 'total_counts' indicates the total number of transcripts detected per spot. These spatial heat maps provide insights into the tissue's gene expression landscape, facilitating the identification of regions for detailed analysis.

Integration of single-cell RNA sequencing (scRNA-seq) with spatial transcriptomics data elucidated the distribution of specific cell types across the sampled tissue. Analysis results for four cell types are presented, showing distinct spatial patterns. For instance, CD8-positive, alpha-beta T cells exhibited localized clusters, potentially indicating areas of immune activation or interaction with other cell types. Type I pneumocytes, commonly associated with gas exchange surfaces, showed a more diffuse distribution, aligning with expected pulmonary structures. Epithelial cells of the lower respiratory tract and plasmacytoid dendritic cells, each with their unique functions in respiratory health and immune response, displayed varying degrees of spatial heterogeneity. These results, representing a subset of all analyzed cell types, highlight the intricate cellular architecture and microenvironments within the tissue as seen in **Figure 11**.

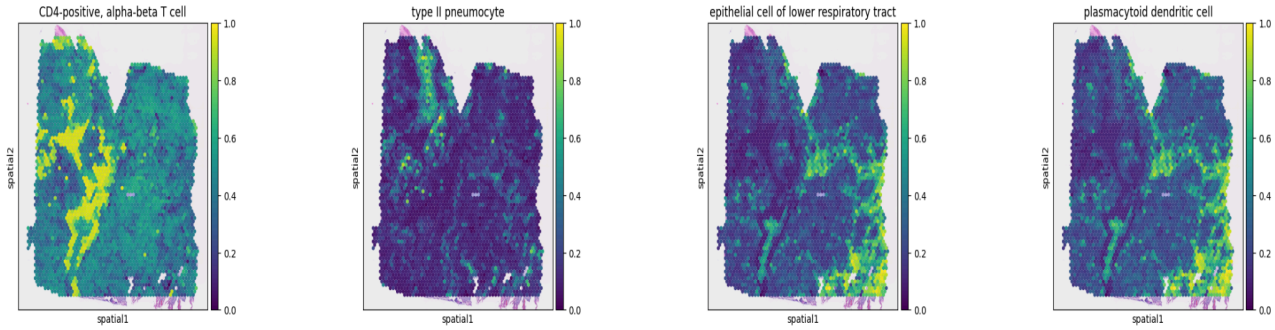

Figure 11: Integrated scRNA-seq and Spatial Transcriptomics Cell Type Distribution. Spatial maps illustrate the distribution and probability density of (A) CD8-positive, alpha-beta T cells, (B) type I pneumocytes, (C) epithelial cells of the lower respiratory tract, and (D) plasmacytoid dendritic cells within the sampled tissue. Each panel represents the inferred abundance of the respective cell type, providing insights into the cellular composition and potential functional zones within the tissue context.

Following the consolidation of cell type annotations, spatial distribution maps were generated to visualize the landscape of various cell types within the tissue. The map revealed a complex mosaic of cellular environments, with distinct cell types occupying specific niches in the tissue architecture. By assigning the most probable cell type to each spatial location, we observed a diverse cellular composition, reflecting the intricate interplay between different cell populations within the tissue microenvironment as seen in **Figure 12**.

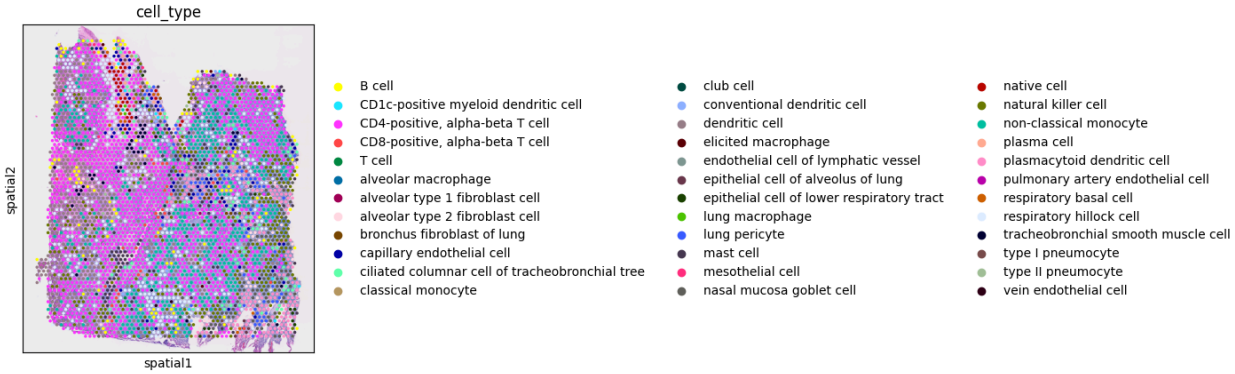

Figure 12: Comprehensive Cell Type Distribution in Spatial Transcriptomics. This spatial plot showcases the distribution of multiple cell types within the tissue sample, with each color representing a different cell type as determined by the highest probability from the label transfer output. The plot provides a detailed visualization of the tissue's cellular heterogeneity, contributing valuable insights into the tissue's structural organization and the spatial interconnectivity of the cellular components.

V. Discussion

Through integrated analysis methods, certain differentially expressed genes showed promise as markers for squamous cell carcinoma, specifically in the lung. Type 1 pneumocytes (T1P) are a critical component of pulmonary function. Over 70 percent of each alveolar surface is composed of T1P, making them a major component of gas exchange. As such, change in the cell state of these cells is of particular concern. The analysis revealed two genes that are dysregulated and differentially expressed for type 1 pneumocytes: CAV1 and AGER.

AGER is a multi-ligand cell surface receptor that is relevant to homeostasis, development, and inflammation. In T1P, **AGER** (the gene encoding the receptor for advanced glycation end-products, RAGE) plays critical roles in lung physiology and pathology. AGER, through its interaction with various ligands, is involved in cellular signaling pathways influencing inflammation, cell proliferation, and survival. It is implicated in the pathogenesis of several lung conditions, including pulmonary fibrosis, acute lung injury, and chronic obstructive pulmonary disease (COPD), and importantly squamous cell lung cancer, through its roles in inflammation and fibrosis. Within T1Ps, AGER was upregulated. Based on the UMAP of cancerous cells, T1P has a clear region towards the left side and can be identified as a present cell type. Further, the

next UMAP for differential gene expression indicates upregulation for cells in that same area. Therefore, we suggest that this is an upregulation of AGER in T1Ps.

Caveolin-1, found in type 1 pneumocytes, plays multiple essential roles in pulmonary function. It is integral to maintaining the structural integrity of the alveolar epithelium, regulating barrier function, modulating signaling pathways, and orchestrating responses to injury and inflammation. When caveolin-1 function is disrupted in type 1 pneumocytes, it has been associated with various lung diseases, including acute respiratory distress syndrome (ARDS), pulmonary fibrosis, and lung cancer. Understanding the nuances of caveolin-1's involvement in type 1 pneumocytes is crucial for understanding the mechanisms underlying these diseases and identifying potential therapeutic targets to address them.

From the analysis, it can be seen that CAV1 is also highly upregulated in the cluster of cells correlated to T1Ps in the cancerous cells. However, this upregulation is not seen in normal cells, marking the distinction between cancer and non-cancerous cells.

LIANA was used to validate the relationship of spatially close cell types. This method employs a receptor-ligand analysis to show that cells clustered together are related. That is, there exist definitive receptor interactions between those cell clusters.

VI. Conclusion:

In conclusion, our study successfully harnesses the complementary strengths of single-cell RNA sequencing and spatial transcriptomics to illuminate the complex landscape of squamous cell lung carcinoma. This integrative approach transcends the limitations of individual sequencing methods, offering a multi-dimensional view that captures not only the heterogeneity of cellular states but also their precise spatial contexts within the tumor microenvironment. By addressing this critical gap, we provide a more refined understanding of cellular dynamics and gene expression patterns, uncovering novel biomarkers and offering insights into the transition from physiological to pathological states. Such depth of analysis sets the stage for targeted therapeutic strategies, potentially revolutionizing personalized medicine for lung cancer patients.

VII. References:

[1] Yosef, N., et al. "An integrated cell atlas of the lung in health and disease." Nature, vol. 625, no. 7895, 2022. Available: https://www.nature.com/articles/s41591-023-02327-2.

[2] Human Cell Atlas. "HCA Lung." Available: https://data.humancellatlas.org/hca-bio-networks/lung.

[3] Lung Cell Atlas. "New repo lung cell atlas." Available: https://github.com/LungCellAtlas.

[4] Treutlein, B., et al. "Single cell RNA analysis identifies cellular heterogeneity and adaptive responses of the lung at birth." PubMed Central (PMC), vol. 9, no. 1, 2019. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6318311/.

[5] Chakraborty, S., et al. "A comparative analysis of single cell small RNA sequencing data reveals heterogeneous isomiR expression and regulation." PubMed Central (PMC), vol. 118, no. 8, 2021. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8857176/.

[6] Brüffer, C., et al. "miRMaster 2.0: multi-species non-coding RNA sequencing analyses at scale." Nucleic Acids Research, vol. 49, no. W1, 2021. Available: https://academic.oup.com/nar/article/49/W1/W397/6238409.

[7] Picelli, S., et al. "Full-length RNA-seq from single cells using Smart-seq2." Nature Protocols, vol. 9, no. 1, 2014. Available: https://www.nature.com/articles/nprot.2014.006.

[8] Stuart, T., et al. "Review scRNA seq." Nature Methods, vol. 18, no. 3, 2021. Available: https://www.nature.com/articles/s41592-021-01171-x.

[9] 10x Genomics, "Human Lung Cancer (FFPE) Spatial Gene Expression Dataset," 10x Genomics, [Online]. Available: https://www.10xgenomics.com/datasets/human-lung-cancer-ffpe-2-standard

[10] Liana-Py. "Basic Usage." Accessed: [3/10/2024]. [Online]. Available: https://liana-py.readthedocs.io/en/latest/notebooks/basic_usage.html

[11] J. M. Taube et al., "Association of PD-1, PD-1 Ligands, and Other Features of the Tumor Immune Microenvironment with Response to Anti-PD-1 Therapy," Nature Medicine, vol. 24, no. 10, pp. 1686–1691, Oct. 2018. DOI: 10.1038/s41591-018-0096-5

[12] M. Kukacka et al., "Inference of Variable Neuronal Responses from Frequent Calcium Imaging Data," Nature Communications, vol. 12, no. 1, pp. 1–14, Jun. 2021. DOI: 10.1038/s41467-021-22801-0

[13] LUNGevity Foundation, "Squamous Cell Lung Cancer." Accessed: [3/7/2024]. [Online]. Available: https://www.lungevity.org/for-patients-caregivers/lung-cancer-101/types-of-lung-cancer/squamous-cell-lung-cancer

Supplementary Figures

In our exploration of squamous cell lung carcinoma (SqCLC), we identified Type I pneumocytes as a potent cell marker, exhibiting pronounced differences between diseased and control states. The UMAP visualizations in our study distinctly showcased Type I pneumocytes, known for their role in gas exchange and lung homeostasis, were markedly altered in the disease state as compared to the control. Notably, in the SqCLC samples, there was a conspicuous reduction in the presence and distribution of these cells, which aligns with the disruption of alveolar structures often observed in lung carcinoma. Conversely, in the control samples, Type I pneumocytes maintained a homogeneous and widespread distribution, indicative of healthy lung tissue architecture.

The disease state showed a fragmented presence of Type I pneumocytes, suggesting a compromised alveolar function and potential keratinization, a hallmark of SqCLC. This aberrant cell population topology hints at the underlying pathological processes disrupting lung parenchyma integrity. The clear contrast between the healthy and diseased tissue samples not only underscores the utility of Type I pneumocytes as a biomarker for SqCLC but also opens avenues for further investigation into the mechanisms driving the transition from a normal to a pathological state.

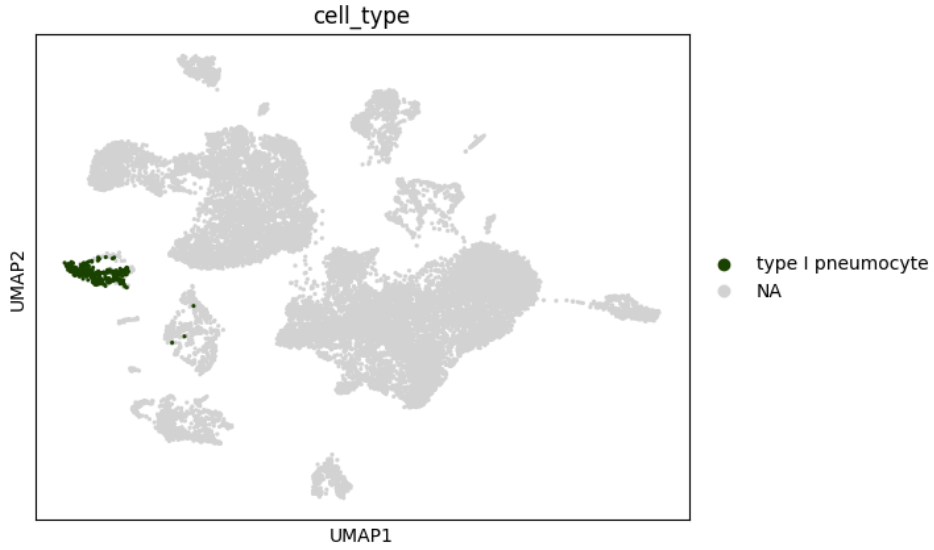

Figure S1 The image depicts a UMAP plot identifying type I pneumocytes (colored green) in the squamous cell lung carcinoma dataset

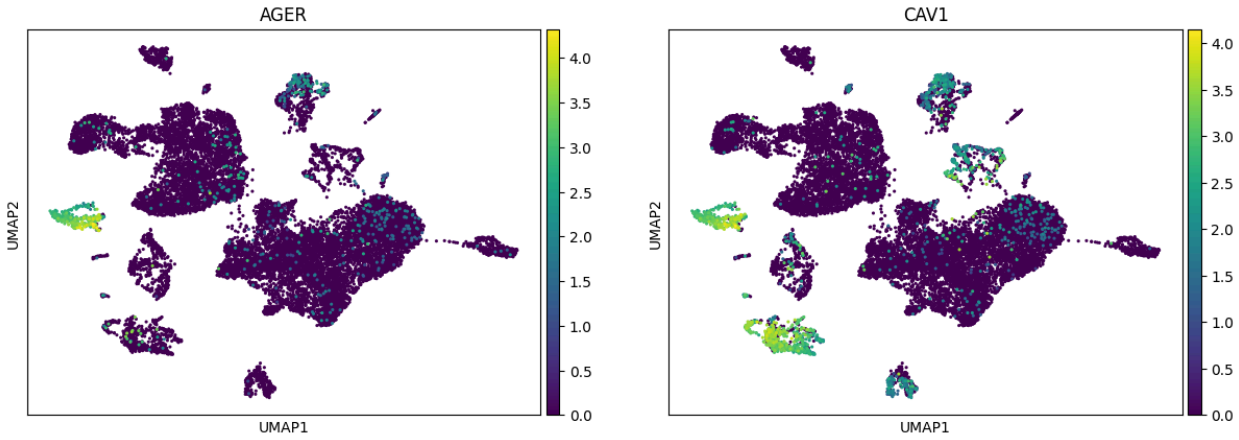

Figure S2 Depicts two UMAP plots representing the expression levels of the AGER gene and the CAV1 gene across the squamous cell lung carcinoma dataset.

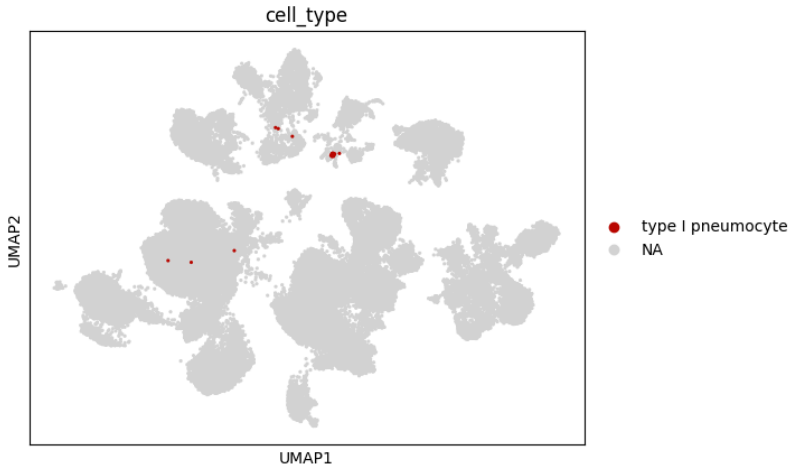

Figure S3 The image depicts a UMAP plot identifying type I pneumocytes (colored green) in the control dataset

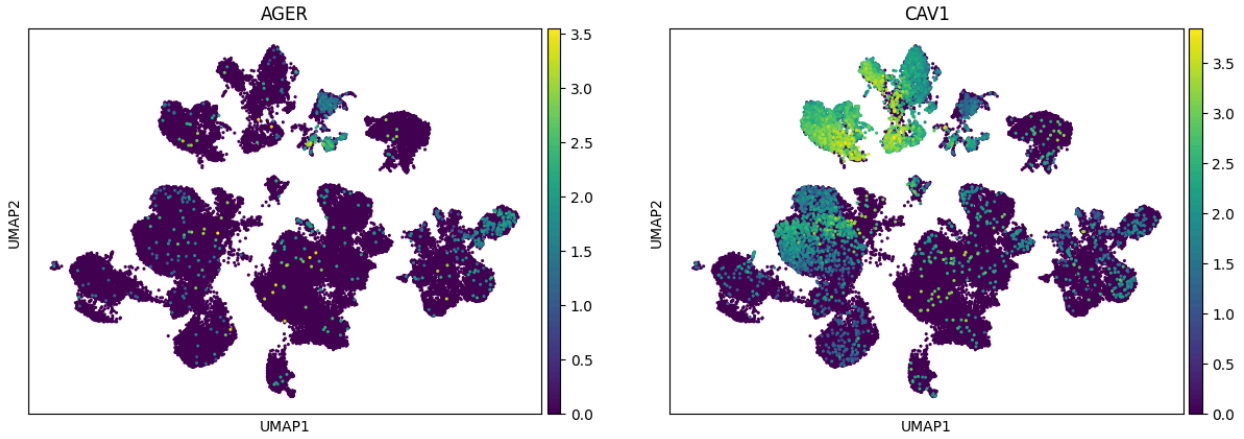

Figure S4 The image depicts two UMAP plots representing the expression levels of the AGER gene and the CAV1 gene across the control dataset

The observed patterns of ligand-receptor interactions, derived from our integrated single-cell RNA and spatial transcriptomics dataset, underpin the complex intercellular communication mechanisms disrupted in squamous cell lung carcinoma (SqCLC). The dot plot accentuates a highly interactive milieu, contrasting with the less complex control state, underscoring the perturbed signaling in SqCLC that may facilitate tumor proliferation and evasion of immune surveillance.

These findings underscore the relevance of Type I pneumocytes beyond their traditional role in gas exchange. In the diseased state, these cells exhibit a distinctive signaling signature, potentially contributing to the pathogenic landscape of SqCLC. This aberration from normalcy is reflective of the tumor's ability to co-opt normal physiological processes for malignant progression. The detailed signaling interactions identified present a fertile ground for therapeutic exploitation, with the potential to dismantle the pathological communication networks without disrupting the essential functions of Type I pneumocytes.

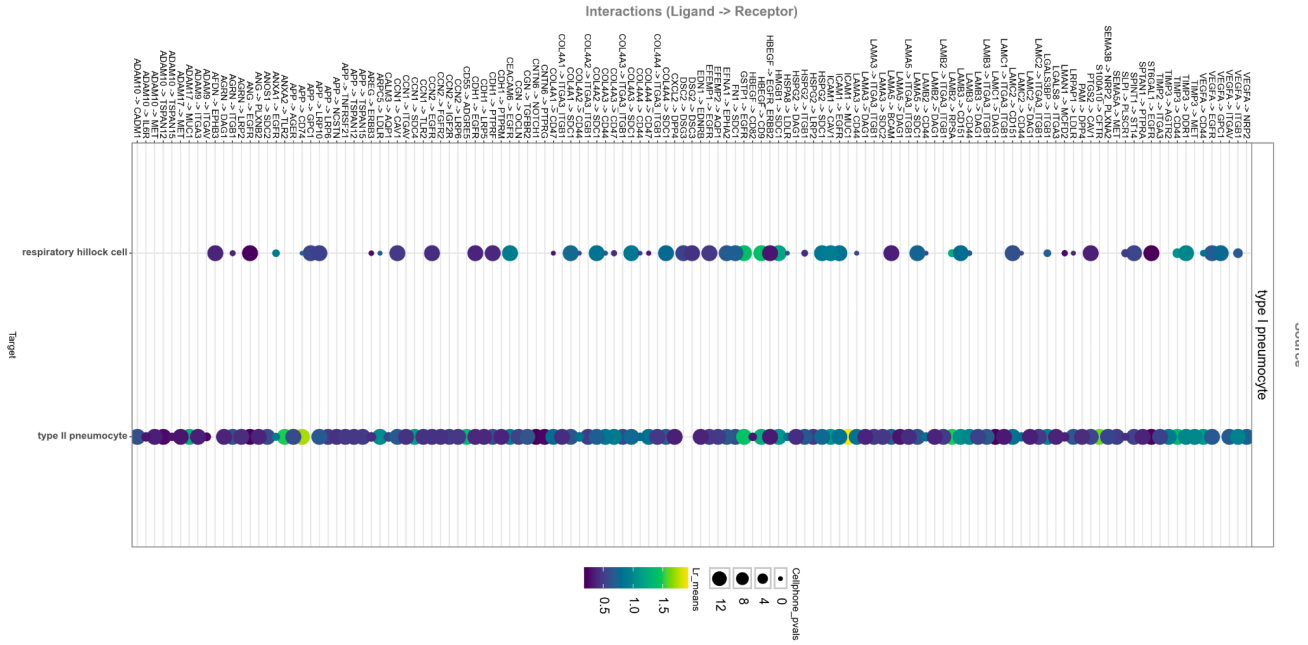

Figure S5: LIANA analysis for each receptor interaction from type I pneumocyte. The targets are the respiratory hillock cell and type II pneumocyte from which the ligand-ligand interactions have been plotted. As seen, you can see the range in means highlight the strength in interactions and size of p-values to further supplement whether the source and target ligand interactions are of significance.