Navigating the Tumor Microenvironment: Identifying Novel Biomarkers in Non-Small Cell Lung Cancer Using Single-Cell Transcriptomics

Abstract: The high degree of cellular heterogeneity in Non-Small Cell Lung Cancer (NSCLC) tumor formation calls for expansive exploration into tumor microenviroments (TME), indicating that TME research is underlooked in the status quo, consequently making the identification of noval genetic biomarkers a flagship for future cancer drug innovations. This study aimed to discover novel NSCLC by i) conducting standard procedure single cell RNA analysis towards both epithelial cell subtypes and immune cell subtypes, ii) utilizing clinical data and implementing statistical models to recognize potential genetic risk factor genes, iii) associating stastically significant NSCLC risk factor genes to potential cell subtypes, and iv) identifying novel biomarkers with protein binding pockets to allow for future drug design innovations. Using single cell sequencing statistics of the 44 tumor samples obtained from 5 NSCLC patients, 60,760 cells were processed, clustered, and annotated. From the 26 cell-type and 8 Epithelial cell-subtype clusters, the basal and brush cell subtype indicated the highest proportion and number of copy number variation. Further clinical prognostic analysis identified 14 statistically significant risk factor genes. In addition, correlation analysis between previously identified genetic biomarkers ($p<0.05$) and 30610 preprocessed T cells demonstrates strong correlation between *RNASET2* and T-regulatory cell-subtype ($r=0.698$) as well as CD4+ cell-subtype ($r=0.663$). Finally, the protein structure visualization of all identified genetic biomarkers ($p<0.05$) revealed two genes (*RNASET2* and *LAMB1*) with protein binding sites, encouraging drug designs with ligand sites that is complementary with the structure of these protein binding pockets. This project hypothesizes that there exists previously undiscovered or untapped biomarkers within the NSCLC tumor micro-environment, and thus, will utilize bioinformatics techniques to discover novel biomarkers, in attempt to provide new directions towards drug design and innovation. Ultimately, it contributes to the medical field by identifying thirteen immunological cancer risk factors, two biomarkers with significant drug development benefits, and one stromal biomarker.

Keywords: NSCLC, Biomarker, Tumor Micro-environment, Bioinformatics

1. Introduction

Despite recent developments, lung cancer, with non-small cell lung cancer (NSCLC) being its most regular subtype (80%-85%), remains a leading cause of death– not only does lung cancer hold the highest mortality rate within the 36 cancer types, but it is also the most commonly diagnosed cancer type. Previous research attributed its high mortality rate (2%-20% of NSCLC patients survive five years from diagnosis) due to the huge amounts of late diagnosis (Rodak et al., 2021). In the past, Scientists have already identified the leading oncogene mutations that are heavily correlated to lung cancer, including *EGFR, TP53, KRAS, MET. ALK, ROS1, and BRAF*. It is worthy to note that in order for proto-oncogene to be activated into oncogenes, there must be an occurrence of gain-of-function mutations that most commonly appears as point mutations in proteins, an overexpression of proteins caused by localized reduplication, and false gene expression caused by chromosomal reduplication. Therefore, this research project is designed in a way to identify important genes or carcinogenic factors that are beneficial to the treatment of NSCLC through single-cell sequencing technology. Ultimately, this study hopes to discover cancer risk biomarkers that may potentially reduce the number of late diagnosis, which is a staggering number considering that in 2020 alone, approximately 2,206,771 individuals were diagnosed with lung cancer drawn from a sample of 185 countries (Rodak et al., 2021).

To achieve the aim, this study analyzed datasets drawn from single cell RNA sequencing. In this paper, the scRNA-seq approach was selected because it could not only identify health cells from tumor cells at early developmental stages, but also, with the aid of other recent technologies like CyTOF (cytometry by time-of-flight), scRNA-seq could confirm the association between *TNFRSF9* and higher levels of tumor necrosis (Rodak et al., 2021). Before the single cell RNA sequencing approach was popularized, researchers could only examine cell features with large quantities of raw material, which makes the costs of conducting studies, especially manual labors, with a cell lens very difficult and costly. In essence, the single-cell RNA sequencing approach analyzes cells individually, which helps researchers understand the role specific cells play in a complex biological system through providing in-depth features on a singular cell. Through analyzing a population of cells with the same functions, the single-cell RNA sequencing approach offers insights into population heterogeneity within cells within the same tissue. The single cell RNA sequencing technique, established by Tang et al. in 2009, allowed more in-depth analysis and decreased costs significantly in cell-based research. Currently, researchers can analyze transcriptome at "single-cell level for over millions of cells" in a single study (Jovic et al., 2022).

Furthermore, this study focuses on cellular compositions of the tumor microenvironment (TME). Recent studies found that the dynamic between cancer cells and TME may be an "active promoter of cancer progression" because the TME may "support cancer cell survival, local invasion, and metastatic dissemination" depending on its components (Melo et al., 2021). Immune cells and stromal cells are the major cellular components of the TME. Specifically, immune cells are grouped into adaptive immune cells, which use immunological memory to strengthen immune responses, and innate immune cells, which produce defense responses immediately. Both groups of immune cells can either suppress or promote tumor genesis. Stromal cells, including vascular endothelial cells, fibroblasts, adipocytes, and stellate cells, vary hugely between different TME. They contribute to tumorigenesis by secreting chemicals that lead to “angiogenesis, proliferation, invasion, and metastasis” (Melo et al., 2021). This project aims to identify the composition of NSCLC cells, further understanding of the relationship between epithelial cell subtypes and NSCLC, employ clinical prediction models to identify potential carcinogens, and research the connection between immune T-cells and tumor formation. Consequently, this study hypothesizes of the existence of undiscovered or untapped biomarkers within the NSCLC tumor micro-environment.

2. Method

2.1 Single Cell RNA sequencing

To generate the dataset, researchers previously employed the technique of single cell RNA sequencing in order to use cell composition as a lens to predict lung cancer. This involves three major steps including first, isolating tissues into singular cells, a step which allows researchers to analyze cells independently; the second step involves capturing the singular cells and extracting the RNA through adding poly-A and poly dt; the last step sequences the RNA. Specifically, this study utilized the technology 10xGenomics, a microfluidics-based method, to allow for single-cell RNA sequencing.

2.2 Dataset Sourcing

This study utilized an open-sourced data in the GEO database, explicitly, it attained single cell sequencing information from 44 tumor samples from 5 NSCLC patients. The statistical distribution of the sample is shown in the table below:

Table 1 Sample Information Table

	Tumor edge tissues	Tumor middle tissues	Tumor center tissues	Healthy tissues (control)
Patient 1	3 samples	3 samples	2 samples	absent
Patient 2	3 samples	3 samples	3 samples	3 samples
Patient 3	2 samples	2 samples	2 samples	2 samples
Patient 4	2 samples	2 samples	2 samples	2 samples
Patient 5	2 samples	2 samples	2 samples	2 samples

2.3 Raw Data Preprocessing

After downloading the dataset, cellranger 7.1.0 processed the dataset: comparing reference genomes, data filtering, data correction, and other procedures to conduct process analysis on the human reference genome, GRch38. The analysis results in three files, namely the barcodes.tsv, features.tsv, and matrix.mtx files for subsequent data analysis.

2.4 Quality Control and Data Integration

This study utilized Seurat packet 4.3.0 to perform analysis. In order to identify cell subjects to include in this study, cells with less than 300 genes and cells where the mitochondrial gene represents over 10% of the entire genetic expression was filtered out. FindIntegrationAnchors was then used to remove the batches and integrate the dataset, which serves the consequent normalizing and standardizing processes.

2.5 Cluster Analysis and Annotation

After the filtering, the remaining 60,760 cells were clustered and annotated. Next, this study employed dimensionality reduction techniques like principal component analysis (PCA) with a coefficient for the total of PCs to compute the data (npcs) of 80. Finally, the UMAP was mapped with the default parameters and carried out cluster analysis on the dataset with a 0.5 resolution for the Louvain algorithm. Ultimately, the UMAP allowed me to annotate the graph form marker-genes cited in multiple PubMed papers and the CellMarker dataset.

2.5.1 NSCLC Cell Composition

NSCLC cells are composed of epithelial cells, endothelial cells, NK cells, cancer cells, myeloid cells, T cells, B cells, fibroblasts cells, and neuroendocrine cells, for which the marker genes *SFN*, *COL4A1*, *NKG7*, *EPCAM*, *CD163*, *CXCL13*, *CD79A*, *COL1A1*, *C1R*, *CENPF* were selected respectively.

2.5.2 T-cell Specific Cluster Analysis and Annotation

30610 T cells from the dataset processed through Seurat were first selected. Following which, dimensionality reduction techniques like PCA with a coefficient for the total of PCs were used to compute the data (npcs) of 50. Finally, the UMAP was mapped with the default parameters and carried out cluster analysis on the dataset with a 0.4 resolution for the Louvain algorithm. The UMAP allowed me to annotate the graph with marker-genes cited from multiple references, primarily, the Human Protein Atlas.

2.6 Using Copy Number Inference to Identify Tumor Cells

The packet infercnv (copy number inference) was downloaded, applied with the cut-off parameter as 0.1, and referenced with the copy number variation results for epithelial cell subtypes to that of T-cells.

2.7 GO, KEGG, GSEA functional enrichment analysis

$Padjust$ and log_2FC scores of the marker genes of both cancer cells and epithelial cells were calculated by selecting genes where $Padjust < 0.05$ and $log_2FC > 1.5$ to conduct functional enrichment analysis. Specifically, clusterProfiler 4.8.3 was used to perform enrichment analysis, namely Gene Ontology (GO) and Kyto Encyclopedia of Gene and Genomes (KEGG) on their differential expression gene. Fgsea 1.26.0 was then downloaded to perform Gene Set Enrichment Analysis (GSEA). This R-package allows for quick and accurate calculations of low GSEA p-values. The results of the top 10 genes filtered through GO and KEGG and the top 4 genes found through GSEA with ggplot2 (3.4.3 version) were then visualized.

2.8 Hub Genes – PPI

This data was imported into STRING to visualize protein-protein interaction network (PPI). Cytoscape_v3.10.1, which allowed for the use of cytoHubba(v0.1) plug-in to filter hub genes, was then downloaded. Although there are 11 topical analysis methods in CytoHubba (v0.1) including Degree, Edge Percolated Component, Maximum Neighborhood Component, Density of Maximum Neighborhood Component, and Maximal Clique Centrality, this paper adopted the MCC (Maximal Clique Centrality) algorithm.

2.9 Clinical Prognostic Analysis

Before graphing the Kaplan Meier (K-M) Survival curve, the data on lung adenocarcinoma (LUAD) that was sourced from the TCGA dataset was preprocessed. Then, the packet, survival, in Rstudio was used to visualize the K-M Survival curve. Further Cox Proportional analysis and the LASSO regression model were implemented to filter out significant risk variables. This hypothesis was then confirmed through graphing a reciver operating character curve (ROC) and calculating the area under its curve (AUC). Lastly, correlation analysis was used to identify the possible cell subtypes that the genetic biomarkers may resemble.

2.9.1 Graphing K-M Survival Curve

Matrix datasets and clinical sample information for both LUAD and LUSD TPM, the two subtypes of NSCLC, from TGCA were downloaded to perform survival analysis. Using the survival packet 3.5.7 and ggplot2 (3.4.3 version), this study graphed the change of expression levels of hub genes through time under two survival states (alive and dead).

2.9.2 Cox Proportional Hazard model

A open-sourced clinical dataset from the TCGA dataset was preprocessed to conduct regression analysis. Specifically, the survival packet 3.5.7 on R was downloaded to make calculations on the model, identifying p value < 0.05 and HR $\neq 1$ as factors to a significant variable. Further, the median score of the risk score was used to identify high and low risk factors. Then, this study also graphed the results that it generated from the random forest algorithm and risk factors analysis. It is also important to understand that the primary construction of the regression analysis relies on this equation:

$$h(t) \ = \ h_0(t) \times exp(b_1x_1 + \ b_2x_2 + \ldots + b_px_p)$$

Where:

t represents the survival time;

$h(t)$ is the hazard function determined by a set of p covariates $(x_1, x_2, \ldots, x_p)$;

The coefficients $(b_1, b_2, \ldots, b_3)$ measure the impact (i.e, the effect size) of covariates;

The term is called the baseline hazard, which is the value of hazard when x_i is equal to zero (Abd ElHafeez et al., 2021).

2.9.3 LASSO Regression Model – filtering significant risk variables

LASSO regression analysis was performed based on overfitting prevention techniques, using the glmnet packet 4.18 to construct a regression model. Lambda.min was selected as the coefficient for filtering significant risk variables.

2.9.4 Graphing ROC Curve

SurvivalROC 1.0.3.1 was used to calculate the ROC for the one-year, three-years, and 5-years survival rate, based on the K-M Survival Curve that was previously predicted. This study further used the percentage of true positive and false negative events to predict the AUC, assuming that when AUC > 0.6, the result is significant.

2.10 Pseudotime Analysis

Monocle 2.28.0 was then downloaded to use the cluster biomarker differential gene approach to retrieve 7974 genes. Then, the data-set was filtered out by p-value, which narrowed it down to 2000 genes; constructed subjects for CellDataSet; estimated the size factor and dispersion; and defined the trace genes to construct the tracing model. From these steps, a heatmap, containing the top 100 genes, a dispersion map with all 2000 genes, and three pseudo time graphs, was created The pseudotime graphs are displayed by T-cell subtype, color, and state respectively.

2.11 Cell and Gene Correlation Analysis

Correlation analysis was conducted between the 7 predicted carcinogens and the 6 marker genes of the t cell subtypes that were annotated. This procedure was performed through the pearson approach, using the cor function to calculate r (the correlation coefficient). Lastly, the dataset was visualized with the ggplot2 package in Rstudio.

2.12 Protein Binding Pocket Detection

This study utilized Pymol, an open-source proprietary molecular visualization system, to detect protein binding pockets in the thirten statistically significant risk factors filtered out by the Cox Proportional Analysis and LASSO regression model and PPI.

3. Results

3.1 Identifying Cell types

After performing dimensionality reduction techniques, the first round of cluster analysis was confucted. The UMAP algorithm classified the dataset into 26 clusters. This study annotated the independent clusters by first grouping them into possible pairs (clusters that shares the same cell-type) based on gene-expression details derived from the wilcoxon heat graph and dot graph were mapped in R (figure 1c); Then identifying marker genes by references and CellMarker database. This resulted in nine total cell types which are epithelial cells, endothelial cells, NK cells, cancer cells, myeloid cells, T cells, B cells, fibroblasts cells, and neuroendocrine cells. UMAPs for each specific marker gene (*SFN, COL4A1, NKG7, EPCAM, CD163, CXCL13, CD79A, COL1A1 C1R, CENPF)* were then created to confirm that each marker gene is distinctly associated with the specific cell types selected. One anomaly was found: the gene expression of NKG7(the marker gene for NK cells) is equally representative for T cells; however, their similarities are understandable as bioinformatic analyses of transcriptional profiles in 2011 showed that NK cells and T cells are highly similar. The violin graph confirmed this pattern.

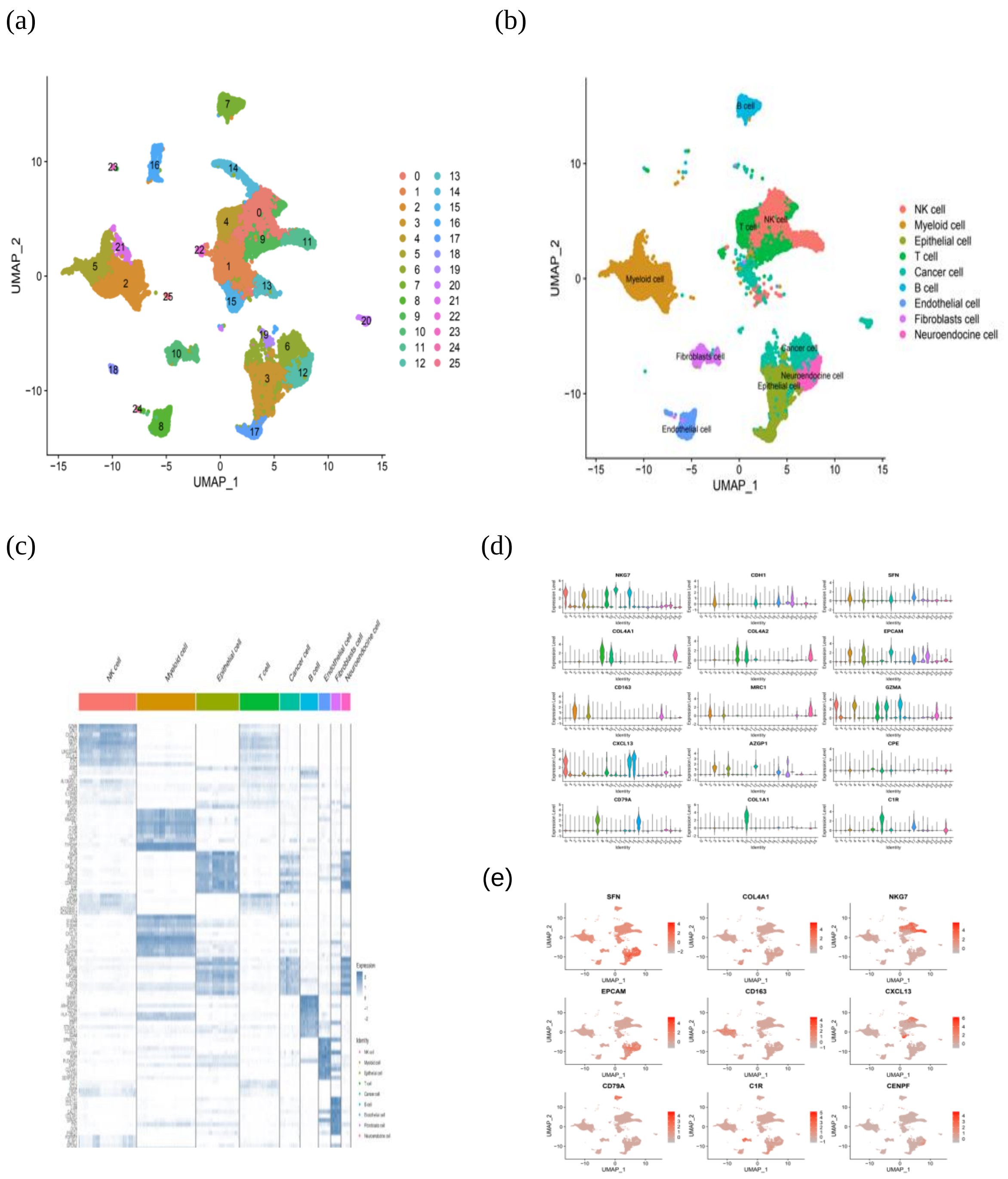

Figure 1 Visualization of cell type and marker genes

(a) UMAP of cell type unannotated
(b) UMAP of cell type annotated
(c) Wilcoxon Heat Map
(d) Violin Graph for cell type marker genes
(e) UMAP of cell type marker genes

3.2 Identifying Epithelial cell sub-types

From literature reviews and additional references, which indicates that most carcinogenic cells are derived from epithelial cells, the paper narrowed down to perform the same process of cluster analysis for the subtypes of epithelial and cancer cells. The results of copy number variation analysis further supports this claim. Specifically, epithelial cells were grouped into the subcategories of cilia cells, secretory club cells, cancer cells, AT1 cells, AT2 cells, basal cells, and brush cells. Again, these cell types were matched with marker genes that was found in CellMarker (and references to confirm their correlation by mapping UMAPs and violin graphs), namely, *TMEM190*, *HLA-DRB1*, *KRT19*, *TM4SF1*, *SFTPA2*, *TUBB*, and *KAZN*.

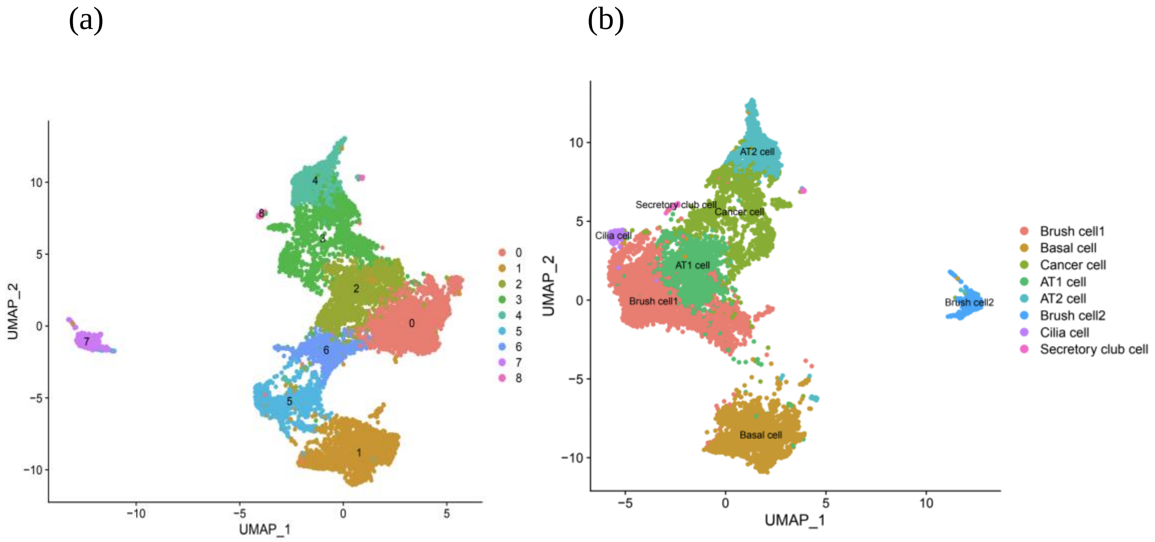

Figure 2 UMAPs of epithelial cell sub-types
(a) Epithelial cell sub-type UMAP prior annotating
(b) Epithelial cell sub-type UMAP post annotating

3.3 Inference Copy Number Variation

The previous decision is supported by the results of the analysis on copy number variation. The decision of analyzing the subtypes of epithelial cells was further validated by the infercnv. Figure 3a depicted that unlike the control, where modified expressions remained close to 1, which meant that there are limited copy number variations—all of the subtypes of epithelial cells showed both areas with concentrated higher and lower modified expressions in its genomic region in the observation cells relative to the reference cells. Large levels of genetic mutation, which can be inferred from dramatic concentrations of modified expressions, indicates changes to cell division, which leads to canceration. Figure 3b showed, on the other hand, the total number of copy number variation in each subtype of epithelial cells. It demonstrated that basal cells have the largest range of copy number variation. The significance of this dataset was later explored through analyzing the proportion of tumor cells to that of paranormal cells for each epithelial cell subtype. This dataset reveals that the proportion of tumor cells is highest in brush cells and the lowest cilia cells.

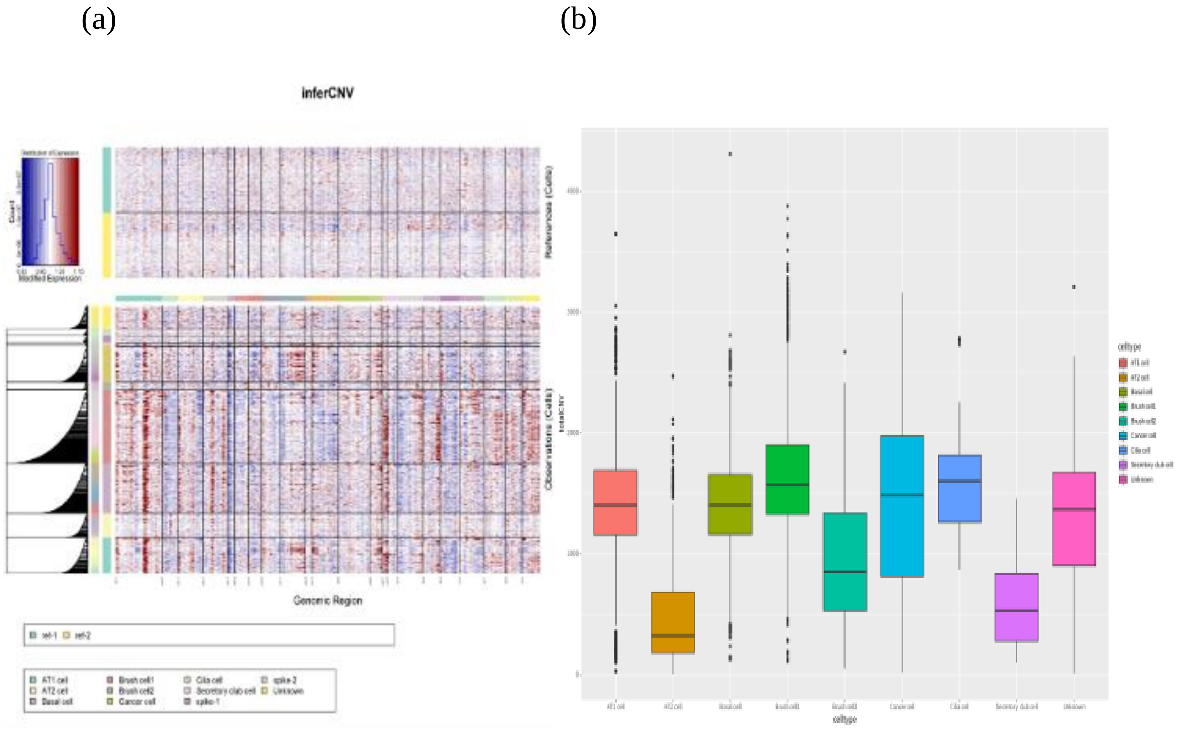

Figure 3. Visualization of Copy Number Variations
(a) InferCNV
(b) Epithelial cell sub-type copy number result boxplot

3.4 Identifying preferred tumor subtype

Two bar plots were createdto reflect the distribution of sample origins, namely, LUAD, lung squamous cell carcinoma (LUSC) and NSCLC. Figure 4a graphs the distribution by cell numbers, showing the number of tumor cells in each cell-type for each subtype of lung cancer respectively; Figure 4b graphs the distribution by percentage, showing the percentage of tumor cells that can be classified under the three subtypes for each cell type. Ultimately, lung adenocarcinoma and lung squamous cell carcinoma are both subtypes of NSCLC, therefore, the percentage of sample originated from LUAD and LUSC need to be compared; and the graph shows that the majority of the participants were LUAD patients in the sample dataset that this research sourced from.

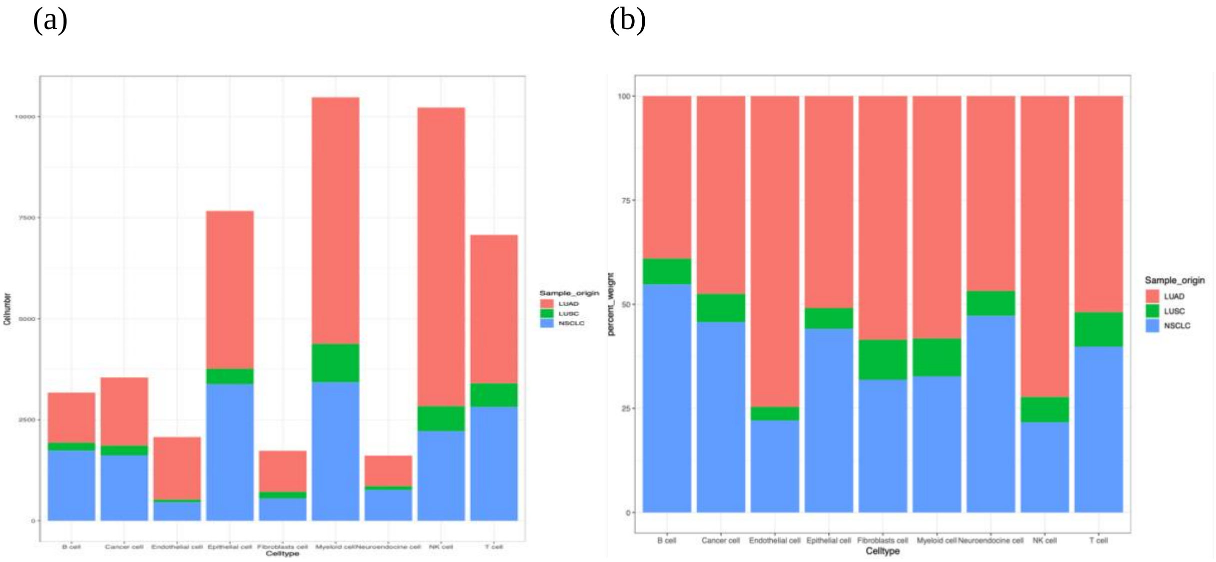

Figure 4. Visualization composition of NSCLC subtype
(a) Distribution of NSCLC Subtype by Cell Number
(b) Distribution of NSCLC Subtype by proportion

3.5 GO and KEGG Enrichment Analysis

The first enrichment analysis of marker genes was performed with Gene Ontology (GO), as it supports organisms with usable OrgDb objects. This approach enables annotating genetic datasets through three categories: molecular function, biological processes, and cell components. The results of the GO graph recognized binding activities to be the most common biological process for the marker gene set; Ribosome and cell adhesion materials are the most common components; The molecular functions varied but most were associated with transportation and movements.

Similarly, another enrichment analysis was done using KEGG, and the results came from differential expression analysis. The KEGG method uses a path lens, which allows me to view the results more dynamically. Note that the size of the dots represents the gene count enriched in the pathway while the bluer the color, the more significant is the pathway enrichment. Here, the analysis identified ribosome and COVID-19 to be the most significantly enriched pathways.

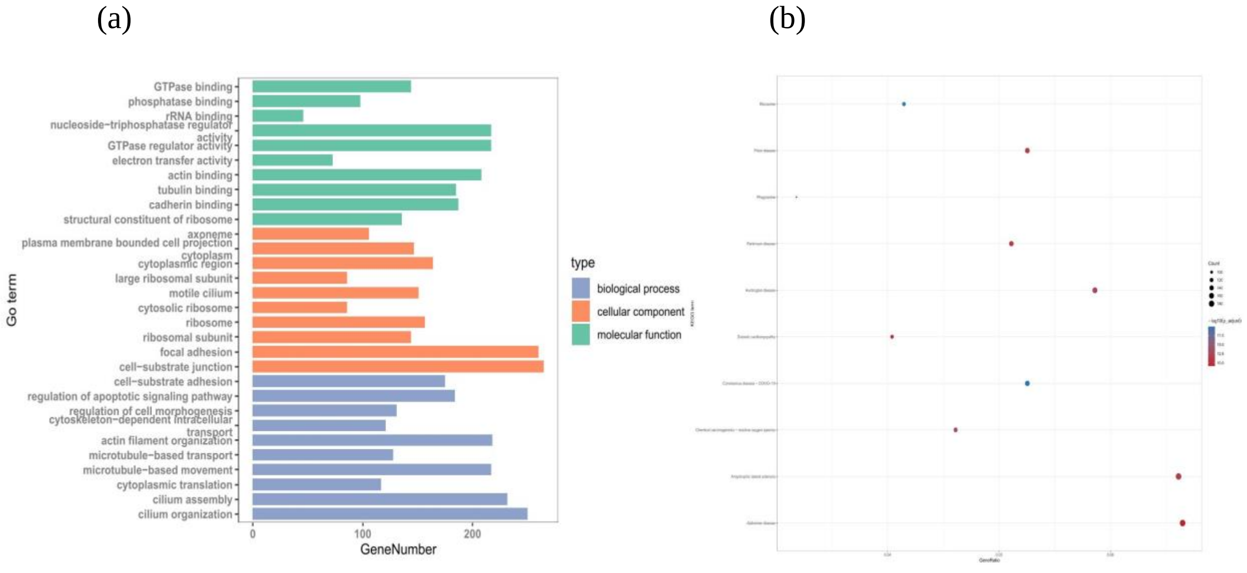

Figure 5. Enrichment Analysis Results
(a) GO results
(b) KEGG results

3.6 Expression Value Difference Analysis and Correlation Analysis

The GSEA method was chosen to conduct expression value difference and correlation analysis for the dataset, msigdb, in the Fgsea packet. This provides the richness score (ES) and the P-value of the genelist, allowing me to standardize the dataset and find the normalized richness score (NES) and the adjusted P-value (p-adjust). Two graphs were mapped out accordingly and four pathways with the highest NES were identified. Furthermore, one cancer-related pathway, namely, chemical carcinogenesis, was found, with an NES of 1.58 and p-value <0.05.

Table 2 GSEA result table

ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust
hsa05310	Asthma	12	0.79678681	2.11757417	0.00110865	0.02912719
hsa04672	Intestinal immune network for IgA production	19	0.71177536	2.07966955	0.00105597	0.02912719
hsa05322	Systemic lupus erythematosus	29	0.69370757	2.19392531	0.00102459	0.02912719
hsa05320	Autoimmune thyroid disease	19	0.64722322	1.89106074	0.00211193	0.0406899
hsa05330	Allograft rejection	19	0.64722322	1.89106074	0.00211193	0.0406899
hsa05204	Chemical carcinogenesis - DNA adducts	16	0.55788199	1.58066254	0.03870968	0.41368451

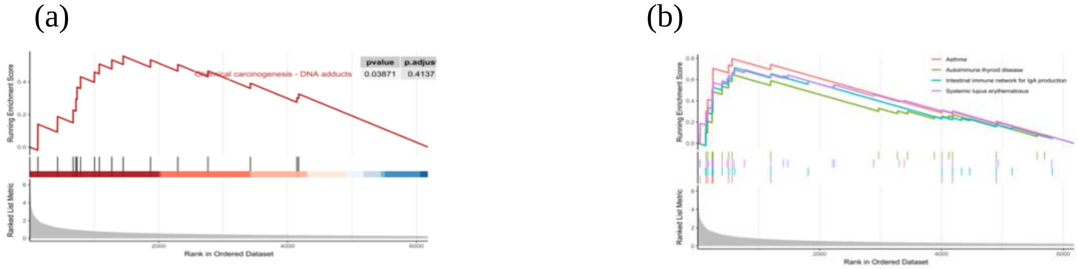

Figure 6. GSEA results
(a) GSEA result cancer related ID
(b) GSEA result top 4 IDs

3.7 Hub genes

First, a PPI graph was generated by the website STRING. Then, to identify the top six hub gene for further analysis, the plug-in Cytohubba in Cystoscope was utilized to filter out the six most important hub genes, which are namely, *NDUFS7*, *ATP5PD*, *NDUFB8*, *NDUFS3*, *COX6A1*, and *NDUFB4*. Specifically, the MCC algorithm was used to select the hub genes because past studies found it to have the best precision in predicting essential proteins in PPI networks of other model organisms, including yeast. This step allowed us to use these differentially expressed genes to draw the Kaplan Meier survival curves.

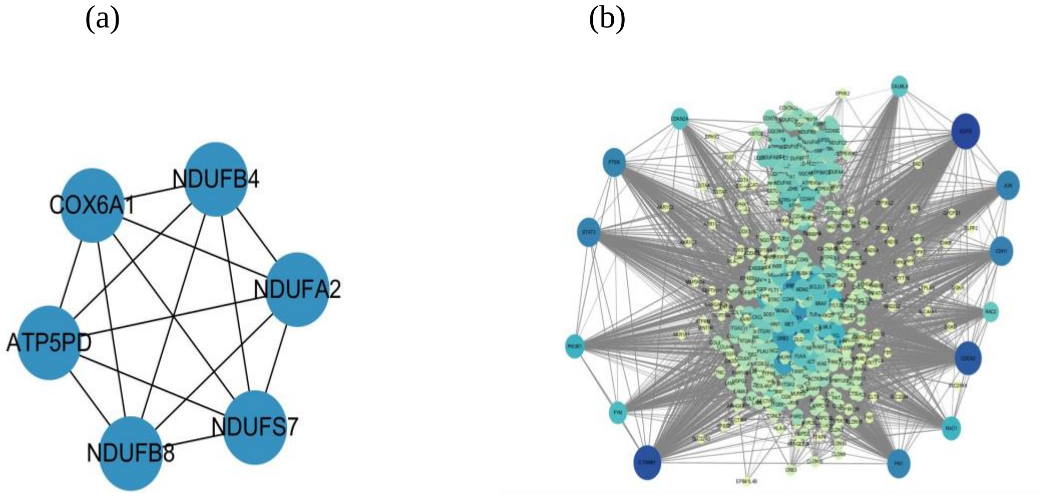

Figure 7. PPI & Cytohubba visualization

(a) Final hub gene selection

(b) Cytoscape rough results

3.8 Kaplan Meier survival curves

Before drawing the Kaplan Meier survival curves (KM), the dataset, retrieved from the TCGA database, was preprocessed. TPM expression matrix for the LUAD, a subdivision of NSCLC, the TCGA clinical samples from this database, and the differential expressed genes matrix that found through the previous process were all utilized. With the aid of the Kaplan Meier survival curves, this study identified the significance each hub gene must contribute to individuals' survival rates and time. One gene produced a statistically significant result: In the graph below, the blue line represents the overall survival rate over years in those who exhibited high expression levels of *NDUFB8* while the red line represents that for those who exhibited low gene expression levels for the same gene. With a p value of 0.038 (p <0.05), this study hypothesize that NDUFB8 has a high potential research value in the field of LUAD tumor type NSCLC.

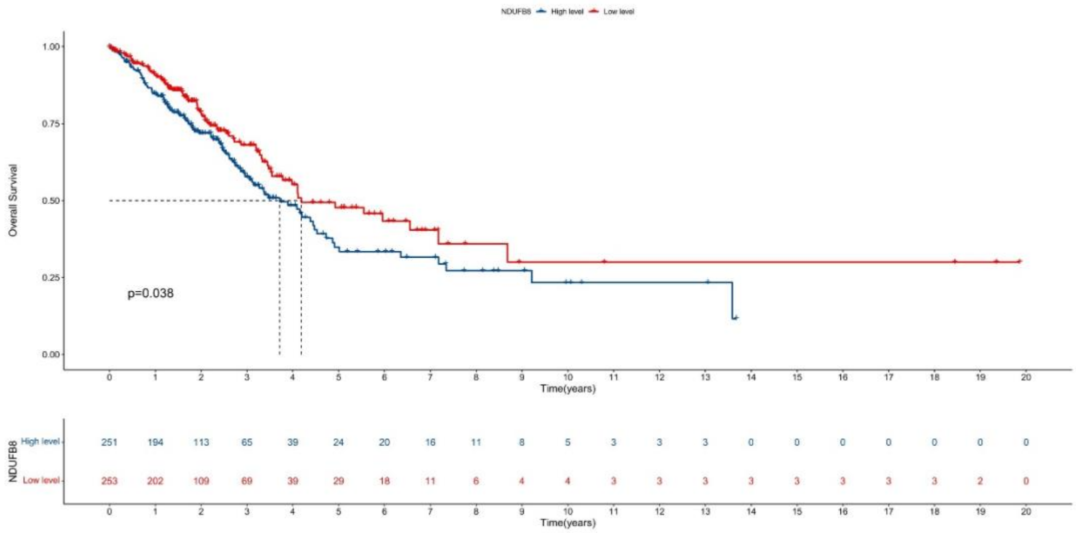

Figure 8. K-M survival curve for *NDUFB8* gene

3.9 Simplification techniques

The single factor cox regression analysis was conducted to screen for important differentially expressed genes that act as independent variables in relationship to survival rates for NSCLC patients, which limits the scope of later analysis procedures. To the same end, a least absolute shrinkage and selection operator regression analysis (LASSO), which simplified the original linear regression model and provided a solution to overfitting by limiting the coefficient for independent variables, was also conducted. In the end, this process used regularization techniques to narrow the number of independent variables from 393 to 93. Figure 9b plots the LASSO model. The graph, with binomial deviance and log(lambda) as its dependent and independent variables, shows two dotted lines which represent the lambda min and lambda 1SE respectively. Note that they are important because the range of values in between the two lines minimizes the binomial deviance. Finally, the lambda min was used in this study because it filters out less genes which broadens the scope of this study.

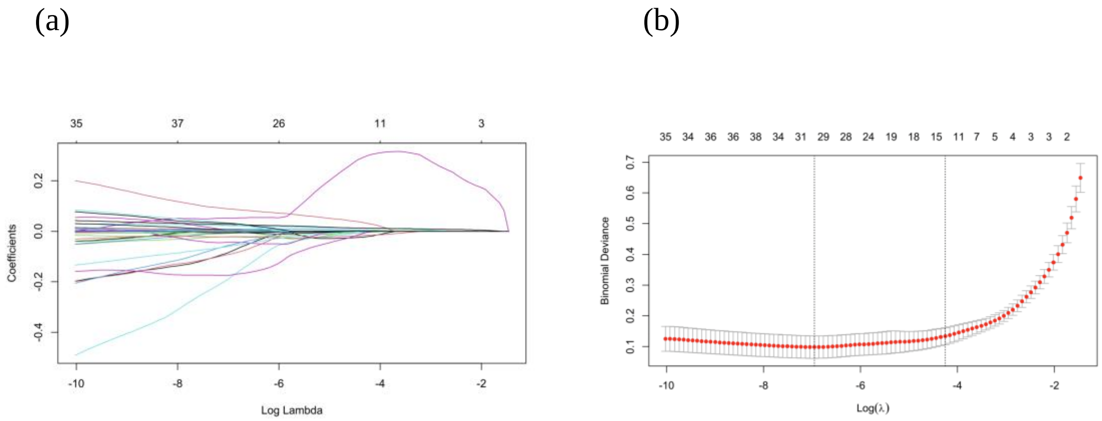

Figure 9. Simplification technique Visualization
(a) LASSO regression analysis coefficient distribution
(b) LASSO analysis result

3.10 Identifying statistically significant change factors

Statistically significant risk factors were identified through multivariate cox regression analysis and the random forest algorithm. The multivariable cox analysis and the random forest algorithm both recognized *RNASET2, HPSE2, LAMB1, RXFP1, KLK6,* and *CAV2* as statistically significant risk variable genes ($p<0.05$). Further they showed that while *HPSE2* and *RXFP1* are associated with decreased risk of lung cancer and increased survival times, *RNASET2, LAMB1*, and *CAV2* are associated with increased risk of lung cancer and decreased survival times. However, the random forest algorithm filtered out *ADAM15* as statistically insignificant ($p>0.05$) while the multivariate cox analysis didn't.

Figure 10c showed how the risk scores of genes are identified in the random forest algorithm graphs; while Figure 10a and 10b visualizes risk with the cox hazard model: On the left side of the graphs the risk scores were lower and more patient samples were alive while on the right side of the two graphs, the risk scores were higher, and the number of dead patient samples was larger than that of alive samples.

Table 3 Multivariable Cox Analysis Results

gene	coef	exp(coef)	se(coef)	z	p
RNASET2	0.033947	1.03453	0.0109	3.114	0.001843
HPSE2	-0.2223	0.800678	0.101267	-2.195	0.028152
LAMB1	0.009352	1.009396	0.003629	2.577	0.009963
GOLM1	0.000693	1.000694	0.000444	1.562	0.118173
RXFP1	-0.22614	0.797604	0.114406	-1.977	0.048079
KLK6	-0.00912	0.990918	0.00392	-2.327	0.019954
KLK8	0.024677	1.024984	0.014092	1.751	0.079928
SPATS2	0.016496	1.016633	0.008426	1.958	0.050258
PLEKHA6	0.007752	1.007782	0.004734	1.637	0.101527
SRGN	0.000288	1.000288	0.000159	1.811	0.070206
ADAM15	0.002233	1.002236	0.001123	1.989	0.046695
CAV2	0.005874	1.005891	0.001564	3.755	0.000173
CAVIN3	0.014055	1.014154	0.008837	1.59	0.111727

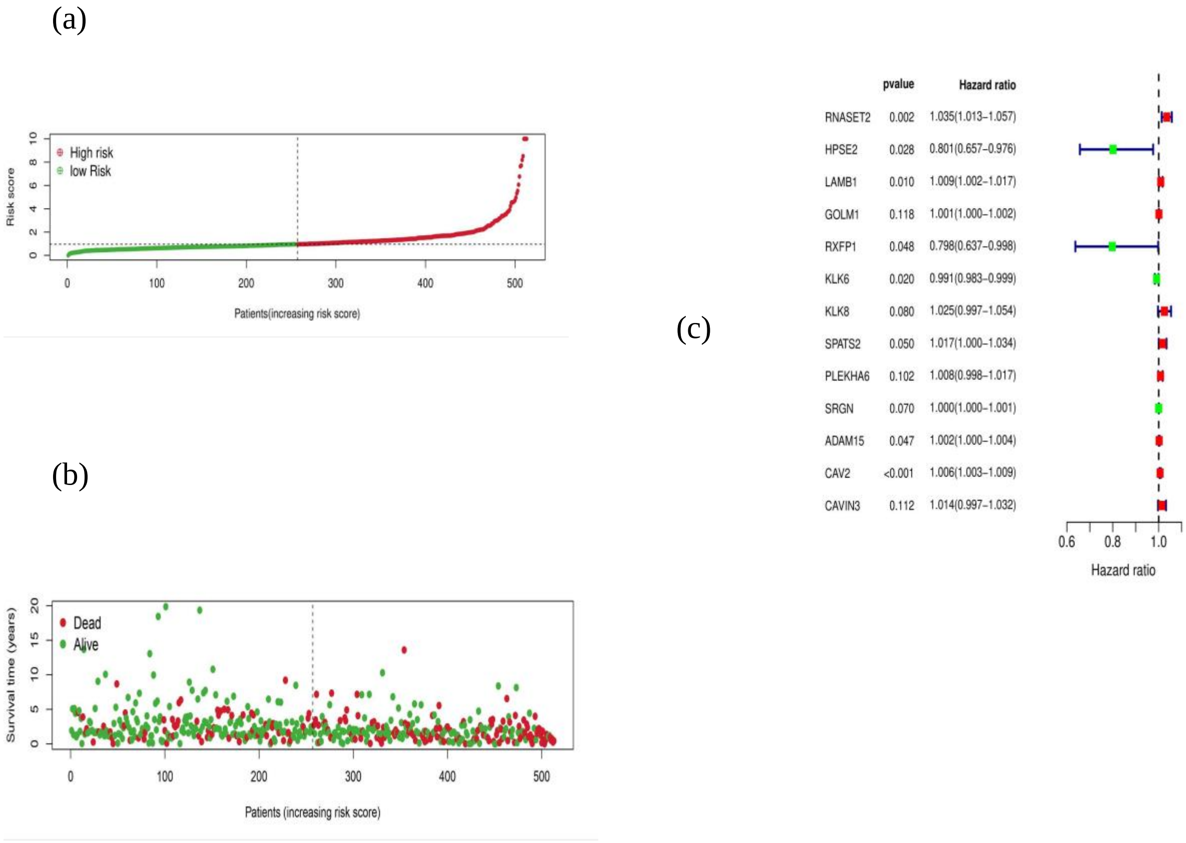

Figure 10. Hazard Ratio Visualized

(a) risk score visualization (based on high/low risk)

(b) risk score visualization (based on living status)

(c) Random Forest Algorithm Visualization

3.11 Model Validation

The hypothesis on the final risk factors was validated by the combined ROC (receiver operating characteristic curve) curve with an AUC (area under curve) around 0.6. In general, an AUC of 0.6 - 0.8 is regarded as acceptable, meaning that the risk factors found in this study affects the ability to diagnose patients of the disease NSCLC.

This research further proceeded to analyze the tumor microenvironments in depth, examining how the interactions between tumor cells and components of their respective microenvironment contributes to the final diagnosis, survival outcomes, and clinical treatments. Since a tumor microenvironment is often divided into one dominated by immune cells and one that is majority composed of fibroblasts, this study decided to target specific relationships between immune cells and tumor diagnosis. Ultimately, immune cells play two significant roles in tumor regulation, namely, the prevention of tumor progression or the promotion of tumor progression, depending on the phenotypes of the immune cells, which makes them a necessary component to examine in depth. Specifically, T-cells were selected due to its long history with NSCLC research and the large varieties of T-cell subtypes.

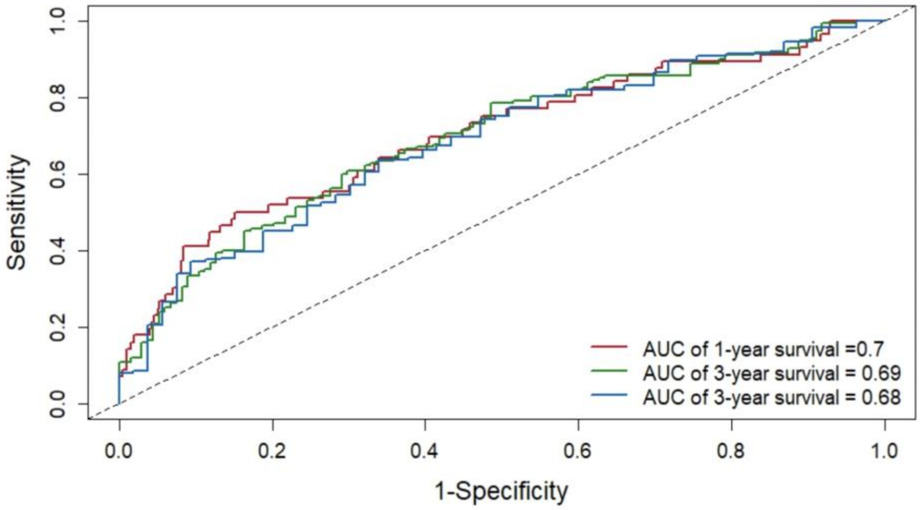

Figure 11. ROC graph

3.12 Identifying T Cell Subtypes

The open-sourced dataset that have undergone the Seurat analysis process and collected data on T-cells was retrieved to be processed with dimensionality reduction techniques to filter out unnecessary information and conducted cluster analysis. The UMAP algorithm classified the dataset into 8 clusters, which were then annotated after being grouped into larger sections, and associated each group with unique genes with high expression levels in the wilcoxon heat map that were graphed in R. Lastly, possible T-cell subtypes that each cluster may represent were identified after the genes with information from other references including the human protein atlas were matched accordingly. Ultimately, 6 T cell subtypes and 6 respective marker genes, which are CD8+ T cells, Vd2 gdTCR cells, Tregs cells, cytotoxic T cells, CD4+ T cells, MAIT cells, and the genes, *CCL4L2*, *KLRCL*, *TNFRSF4*, *HSPA1B*, *SERPINE2*, and *RYR2* were identified. To make sure that the marker genes that were identified associates to the T cell, relatively uniquely, a violin plot and UMAPs designated for each gene were plotted. Note that because the patterns for Treg cells and CD4+ T cells are relatively similar because T-reg cells may also be recognized as a branch of CD4+ T cells.

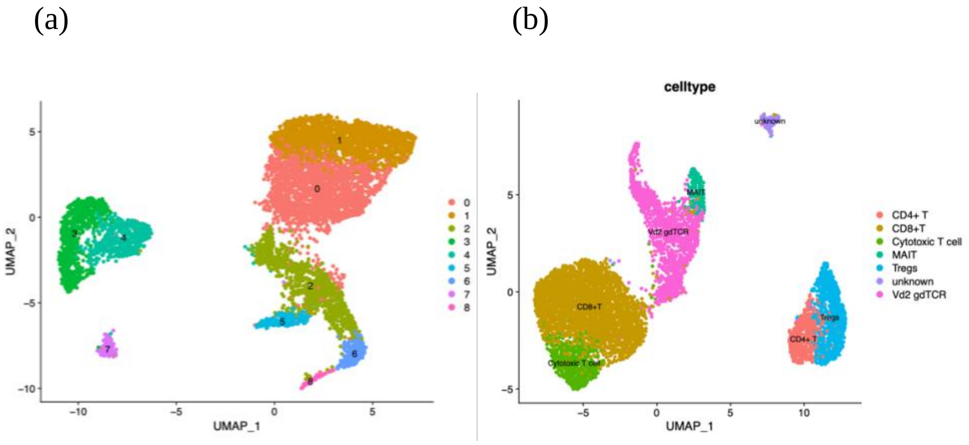

Figure 12. T cell subtype UMAP
(a) T cell subtype UMAP unannotated
(b) T cell subtype UMAP annotated

3.13 Pseudotime Analysis

The cluster differential gene bio-markers were used to obtain 7973 genes, and filtered it to 2000 genes by order of p-value from the smallest to the largest. Then the order genes were employed to find how dispersed the 2000 genes are, shown in Figure 13b. After these preparations, pseudotime analysis was condusted. Left 2 Figure 13a orders the pseudotime of each marker gene by the shade of the color, with the darker color meaning that a gene has a lesser pseudotime. Moreover, since, from right to left, a singular pivot point separates the pseudotime graph into three branches, a pseudotime graph by the three states, also seen as the three branches was then mapped out. After which, a pseudotime graph by cell type (left 1 and 3 Figure 13a) was drawn out. Tregs and CD4+ T cells, Vd2 gdTCR ($\gamma\delta T$cells), and CD8+ T cells were found to be concentrated around state 1, 2, 3, respectivel. The top 100 genes were also selected by order of p-value, to graph a heatmap (Figure 13c), which was grouped into four clusters to visualize the change in mean expression as the pseudotime moves on x-axis of the graph is pseudotime. Finally, the information obtained from the previous graphs were utilized to map out Figure 13d. Figure 13d (left 1) is a jitter plot that visualizes the mean expression levels of each marker gene that were found during T-cell subtype identification under each pseudotime state; Figure 13dleft 2 is a violin plot that expresses the amount of the six marker genes under the three states respectively; Figure 13d right 1 shows the trends of each marker gene as the pseudotime progresses. Specifically, the three visualizations suggested that *CCL4L2*, the marker gene for CD4+T cells, *HSPA1B*, the marker gene for cytotoxic T cells, and *TNFRSF4*, the marker gene for T-regulatory cells have visible trends.

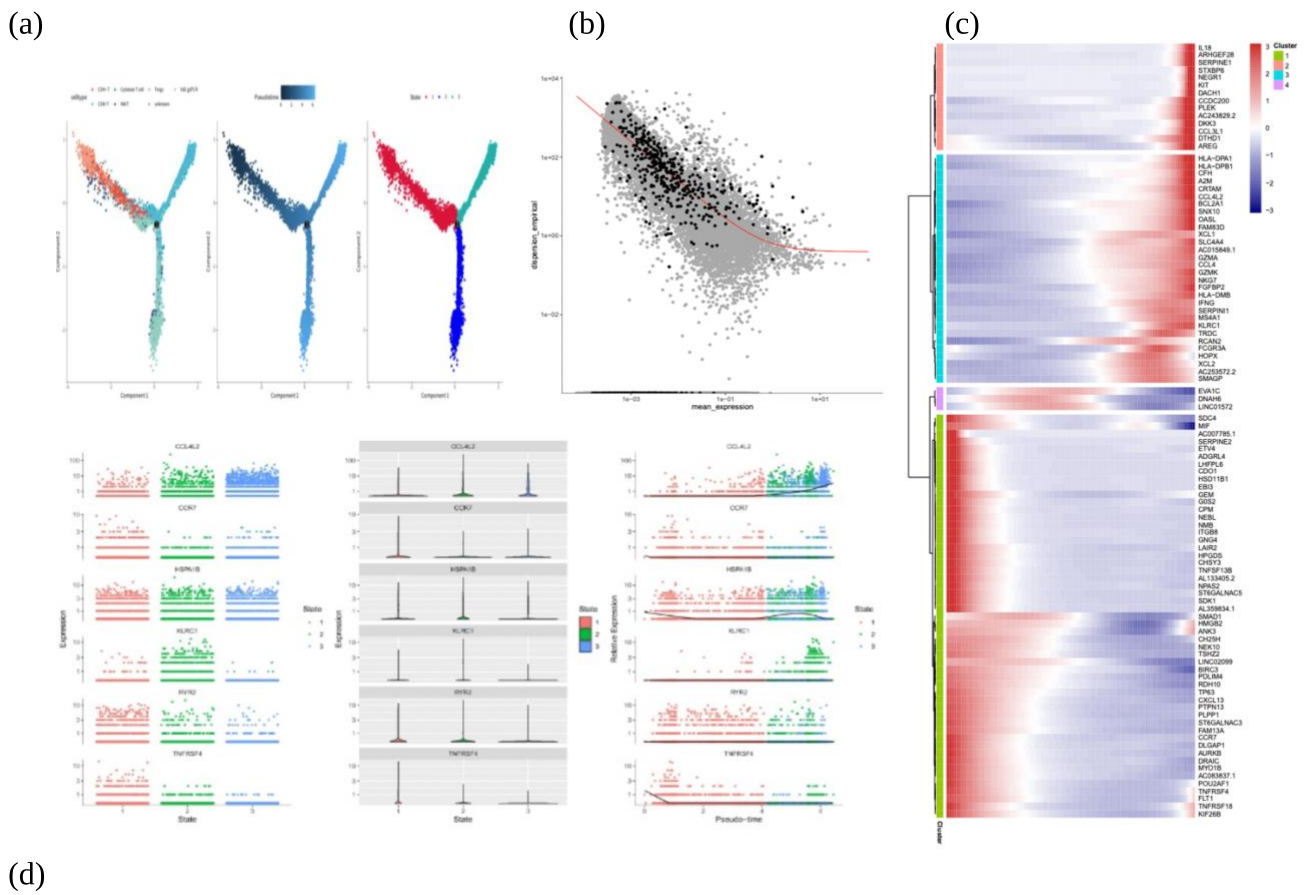

Figure 13. Pseudotime Visualization

(a) Pseudotime graph

(b) Dispersion map

(c) Marker gene heat map

(d) Pseudotime state progression

3.14 Correlation Analysis

A significant correlation (r>0.5 and p<0.01) was found between *RNASET2* and both treg cells and CD4+T cells. This was explainable because treg cells, also known as regulatory T cells, are a subtype of CD4+T cells, and accounts for 5~10% of peripheral blood (PB) CD4+T cells.

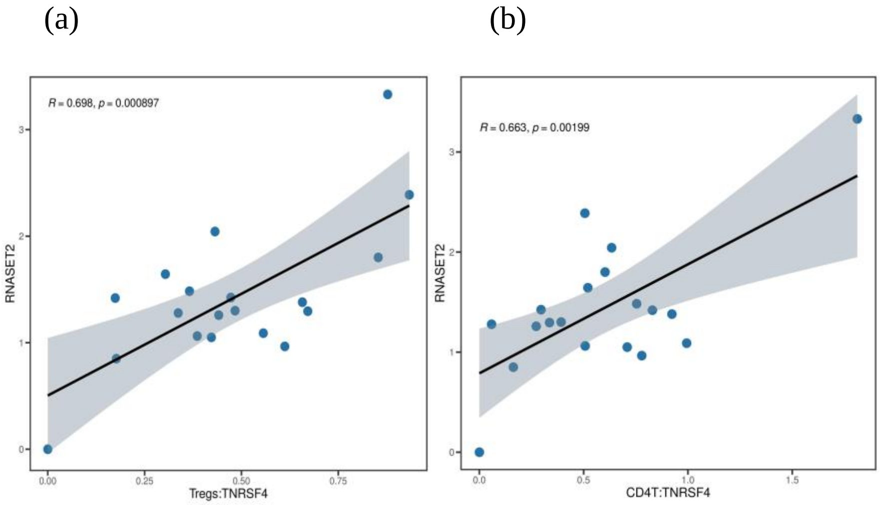

Figure 14. Correlation Analysis Visualization

(a) Correlation Graph between *RNASET2* and Treg cells

(b) Correlation Graph between *RNASET2* and CD4+ cells

3.15 Protein Binding Pocket Visualization

Pymol, an open-sourced molecular visualization application, was utilized to visualize the protein structure of statistically significant risk variable genes and identify whether they contain protein binding pockets. Since to either suppress or express a gene, an activator or an inhibitor protein must bind to the gene's protein binding pocket, potential drug discovery is linked hand-in-hand to the shape and location of potential protein binding pockets. From the thirteen risk variable genes found through multivariable cox analysis, two genes contained potential protein bind pockets: specifically, LAMB1 and RNASET2, both of which are positively with increased risks of NSCLC and a decreased survival rate.

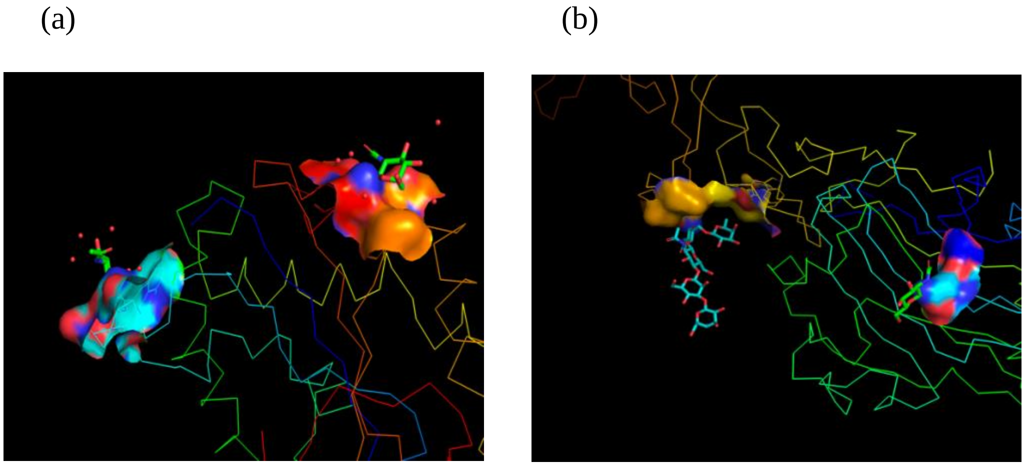

Figure 15. Protein Binding Pocket Visualization

(a) Visuation for RNASET2

(b) Visuation for LAMB1

4. Discussions

In the study, the relationship between epithelial cells and lung cancer were confirmed, while finding specific epithelial cell subtypes that may induce tumors. While previous research has established that a large proportion of lung cancer developed from epithelial cells which line human airways, this paper identified that in particular, the basal and brush cell subtype of epithelial cells are significant for future research (Orr & Hynds, 2021). This was concluded because from the copy number variation inference, brush cells were identified as the epithelial cell sub-type with the highest proportion of tumor cells and basal cells to have the highest copy number variation. This conclusion was reached because copy number variations are commonly caused by genome rearrangements, which leads to errors in the cell division process. The drastic changes in cell division eventually promotes tumorigenesis.

Through conducting multivariate cox analysis and using the random forest algorithm, *RNASET2* was recognized a statistically significant risk variable gene ($p < 0.05$ HR = 1.035) associated with increased risk of lung cancer and decreased survival times. Furthermore, there was a significant correlation (r>0.5 and p<0.01) between *RNASET2* and CD4+T cells (especially in its T regulatory subtype). While previous studies have already detected this enzyme in particularly embryonic tissues and immune cells, particularly in the monocyte/macrophage's lineage, research has not been done on the relationship between the human *RNASET2* gene and NSCLC. Before this study, research on the *RNASET2* gene had been focused on its connection to ovarian cancer. RNASET2 has been identified as a tumor suppressor gene in preventing ovarian tumorigenesis (Ji et al., 2021). This conflicted with the results obtained from this study: *RNASET2* associates with increased rates of lung cancer. Ultimately, this distinction makes possible interpretations of overexpression of RNASET2 in cancer patients a significant and interesting field for further research.

Later, through the PPI map that was generated on STRING and the Kaplan Meier survival curves, a unique hub gene that has high potential research value for further studies in NSCLC research identfied. Previous studies targeting the relationship between NSCLC and oxidative phosphorylation (OXPHOS), discussed notable changes in the expression of *NDUFB8* in LUAD tumor cells compared to healthy cells. *NDUFB8* (Ubiquinone Oxidoreductase Subunit B8) which is involved in the assembly of the mitochondrial respiratory chain complex I, is also a biomarker of Alzheimer's disease and Parkinson's disease (Stelzer G et al., 2023). The gene, along with *NDUFV1, NDUFV2, NDUFS1, NDUFS2, NDUFS3, NDUFS7, NDUFS8,* forms the core subunits of complex I, and catalyzes the oxidation of NADH and electron transfer, which is consequential for the OXPHOS machinery. Previous studies have established that other components of the OXPHOS machinery are highly predictive in detecting early-stage NSCLC, in particular, "the expression of OXPHOS genes is negatively correlated with the prognosis of lung adenocarcinoma"

(Kalainayakan et al., 2018). However, while it is conclusive that high levels of OXPHOS genes helps detect early-stage cancer because the process energizes cancer cells, there is very limited research on *NDUFS8* on the area of NSCLC making further studies necessary.

Finally, this paper recommend researchers to focus future clinical drug discovery for NSCLC on inhibiting the expression of *LAMB1* and *RNASET2*. Although previous researchers have suggested that there was a down-regulation of *RNASET2* in cancerous gastric tissues in comparison to adjacent non-cancerous tissues, the same article describes decreasing levels of *RNASET2* in cancerous stomach tissues as a "consequence of GAC (Gastric Adenocarcinoma Cancer), instead of the driver." Furthermore, a CRISPR/Cas9 analysis found that "RNASET2 knock out had no significant effect on angiogenesis and tumor cell growth but increased cell differentiation," suggesting that the correlation between low levels of *RNASET2* and tumor growth is not causational (Hosseini et al., 2022). However, it is worthy to note the potential harms of *RNASET2* deficiency. Specifically, *RNASET2* deficiency may impair "the recycling of ribosomal RNA," which is an essential element of protein synthesis (Cox, 2023). Similarly, RNaseL's inactivation of *RNASET2* "could lead to a failure to degrade either extracellular and/or intracellular ssRNAs, which triggers the innate immune response with downstream consequences on development." Ultimately, this study also confirms this hypothesis, finding a significant correlation between *RNASET2* and CD4+T cells, especially T-regulatory cells, cell subtypes that compose the immune system. Additionally, research on Ovarian Cancer "suggested a role for RNASET2 as a class II Tumor Suppressor Genes, whose function is abolished in cancer tissues mainly by downregulation of its expression rather than by mutational events" (Bruno et al., 2022). While this paper targets on providing plausible solutions to NSCLC, the effects of

RNASET2 on other cancer subtypes should also be taken into consideration under the larger umbrella of the effects of minimizing *RNASET2* expression.

On the otherhand, although drug innovation that focuses on inhibiting *LAMB1* is limited, *LAMB1* has been historically associated with the initiation and progression of pneumoconiosis. Studies found that reduced activity of *LAMB1* transcription correlates to reduced levels of Coal worker's pneumoconiosis (CWP). Similarly, an analysis of the effects of *LAMB1* on gastric cancer prognosis states that the gene “elevated cell proliferation, invasion, and migration” by “promotes cell growth and motility via the ERK/c-Jun axis.”

Moreover, future research can focus on creating new, distinct algorithms to calculate the ligand compatibility between the protein binding pockets and potential medical treatments. To specify, studies should focus on both the positive and negative effects that drug innovations and protein binding pockets may influence each other. Note that the relationship between drugs and protein binding pockets models the induced-fit theory, stating that both the protein structure and ligand site modify its shape to maximize binding efficiency and success. Another means of future research is to produce new algorithms to predict the probability of cryptic protein pockets, similar to the PocketMiner graph neural network, and further expand the predicting algorithm from the probability of potential cryptic pockets to their potential shapes (Meller et al., 2023)

There are also limitations to this study that are worth mentioning in this paper. While this study aims to perform a thorough analysis of NSCLC tumor structure and potential carcinogens, most of this study focuses on LUAD, a subtype of NSCLC, rather than LUSC. This is because the open-sourced dataset that this study adopted found a higher proportion of tumor cells under the LUAD subtype. This means that the results of this study are more generalized towards LUAD patients.

5. Reference

Bruno, A., Noonan, D. M., Valli, R., Porta, G., Taramelli, R., Mortara, L., & Acquati, F. (2022). Human rnaset2: A highly pleiotropic and evolutionary conserved tumor suppressor gene involved in the control of ovarian cancer pathogenesis. *International Journal of Molecular Sciences*, *23*(16), 9074. https://doi.org/10.3390/ijms23169074

Chen, G., Ning, B., & Shi, T. (2019). Single-Cell rna-seq technologies and related computational data analysis. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00317

Cox, T. M. (2023). Lysosomal diseases. *Encyclopedia of Cell Biology*, 977-1028. https://doi.org/10.1016/b978-0-12-821618-7.00282-0

The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses (PMID: 27322403 ; Citations: 2,595) Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary, D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5 [PDF]

Hosseini, S. A., Salehifard Jouneghani, A., Ghatrehsamani, M., Yaghoobi, H., Elahian, F., & Mirzaei, S. A. (2022). CRISPR/Cas9 as precision and high-throughput genetic engineering tools in gastrointestinal cancer research and therapy. *International Journal of Biological Macromolecules*, *223*, 732-754. https://doi.org/10.1016/j.ijbiomac.2022.11.018

Ji, M., Zhao, Z., Li, Y., Xu, P., Shi, J., Li, Z., Wang, K., Huang, X., & Liu, B. (2021). FBXO6-mediated rnaset2 ubiquitination and degradation governs the development of ovarian cancer. *Cell Death & Disease*, *12*(4). https://doi.org/10.1038/s41419-021-03580-4

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., & Luo, Y. (2022). Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine, 12*(3). https://doi.org/10.1002/ctm2.694

Kalainayakan, S. P., FitzGerald, K. E., Konduri, P. C., Vidal, C., & Zhang, L. (2018). Essential roles of mitochondrial and heme function in lung cancer bioenergetics and tumorigenesis. *Cell & Bioscience*, *8*(1). https://doi.org/10.1186/s13578-018-0257-8

Meller, A., Ward, M., Borowsky, J., Kshirsagar, M., Lotthammer, J. M., Oviedo, F., Ferres, J. L., & Bowman, G. R. (2023). Predicting locations of cryptic pockets from single protein structures using the pocketminer graph neural network. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-36699-3

Melo, C. M., Vidotto, T., Chaves, L. P., Lautert-Dutra, W., Reis, R. B. D., & Squire, J. A. (2021). The role of somatic mutations on the immune response of the tumor microenvironment in prostate cancer. *International Journal of Molecular Sciences*, *22*(17), 9550. https://doi.org/10.3390/ijms22179550

Mithoowani, H., & Febbraro, M. (2022). Non-Small-Cell lung cancer in 2022: A review for general practitioners in oncology. *Current Oncology*, *29*(3), 1828-1839. https://doi.org/10.3390/curroncol29030150

Orr, J. C., & Hynds, R. E. (2021). Stem cell–derived respiratory epithelial cell cultures as human disease models. *American Journal of Respiratory Cell and Molecular Biology*, *64*(6), 657-668. https://doi.org/10.1165/rcmb.2020-0440tr

Remark, R., Becker, C., Gomez, J. E., Damotte, D., Dieu-Nosjean, M.-C., Sautès-Fridman, C., Fridman, W.-H., Powell, C. A., Altorki, N. K., Merad, M., & Gnjatic, S. (2015). The non–small cell lung cancer immune contexture. A major determinant of tumor characteristics and patient outcome. *American Journal of Respiratory and Critical Care Medicine*, *191*(4), 377-390. https://doi.org/10.1164/rccm.201409-1671pp

Rodak, O., Peris-Díaz, M. D., Olbromski, M., Podhorska-Okołów, M., & Dzięgiel, P. (2021). Current landscape of non-small cell lung cancer: Epidemiology, histological classification, targeted therapies, and immunotherapy. *Cancers*, *13*(18), 4705. https://doi.org/10.3390/cancers13184705

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Tan, Z., Chen, X., Zuo, J., Fu, S., Wang, H., & Wang, J. (2023). Comprehensive analysis of scRNA-Seq and bulk rna-seq reveals dynamic changes in the tumor immune microenvironment of bladder cancer and establishes a prognostic model. *Journal of Translational Medicine*, *21*(1). https://doi.org/10.1186/s12967-023-04056-z

Wang, C., Yu, Q., Song, T., Wang, Z., Song, L., Yang, Y., Shao, J., Li, J., Ni, Y., Chao, N., Zhang, L., & Li, W. (2022). The heterogeneous immune landscape between lung adenocarcinoma and squamous carcinoma revealed by single-cell RNA sequencing. *Signal Transduction and Targeted Therapy*, *7*(1). https://doi.org/10.1038/s41392-022-01130-8

Zhang, Y., Wang, D., Peng, M., Tang, L., Ouyang, J., Xiong, F., Guo, C., Tang, Y., Zhou, Y., Liao, Q., Wu, X., Wang, H., Yu, J., Li, Y., Li, X., Li, G., Zeng, Z., Tan, Y., & Xiong, W. (2021). Single-cell RNA sequencing in cancer research. *Journal of Experimental & Clinical Cancer Research*, *40*(1). https://doi.org/10.1186/s13046-021-01874-1