Journal of Advanced Zoology
ISSN: 0253-7214
Volume 45 Issue S-1 Year 2024 Page 158:170

Comparative Analysis Of Lung And Breast Cancer Complexity Using Single-Cell RNA Sequencing Data

Sudarshan Gogoi[1]*, Subham Pradhan[2], Dibyajyoti Chintey[3], Naseem Zoha Ansari[4], Chaitanya Saikia[5]

[1],[2,4]Department of Mathematics, Sikkim University, Gangtok-737102, Sikkim, India*
[3]Department of Mathematics, Dibrugarh University, Dibrugarh-786004, Assam, India
[5]Department of Mathematics, Assam Science and Technology University, Guwahati-781013, Assam, India

***Corresponding Author:** Sudarshan Gogoi*
**Department of Mathematics, Sikkim University, Gangtok-737102, Sikkim, India*
Email: sgogoittb@gmail.com

Article History

Received: 15 Sep 2023
Revised: 28 Dec 2023
Accepted: 10 Jan 2024

Abstract

We performed a thorough analysis of six single-cell RNA sequencing (scRNA-seq) datasets from the 10X Genomics Database in this work. We used Principal Component Analysis (PCA) to reduce dimensionality, clustering, quality control, normalization, identification of high variable features, data preprocessing, and Uniform Manifold Approximation and Projection (UMAP) for visualization. To better comprehend cellular heterogeneity, we also identified marker genes for each cluster and looked at gene correlation networks. In comparison to breast cancer datasets, our results showed that lung cancer datasets had more edges and marker genes in their gene correlation networks. This implies that the lung cancer samples have higher levels of molecular complexity and heterogeneity. Furthermore, a detailed depiction of the cellular environment that highlighted the complex interactions between cell groups was made possible by the UMAP visualization. The underlying biology of lung and breast cancers is better understood because to the discovery of marker genes and the examination of gene correlation networks. The found intricacy in datasets related to lung cancer could have consequences for comprehending disease subgroups, signaling pathways, and overall heterogeneity. This work lays the groundwork for future investigations into the molecular details of cancer and the development of tailored treatment plans.

Keywords: Normalization, scaling, PCA, UMAP, Clusters, Marker Genes, Gene Correlation, Network

1. Introduction

Single-cell RNA sequencing (scrRNA-seq), has completely changed the study of cellular heterogeneity and gene expression patterns. The 10X Genomics Chromium platform has become a prominent instrument for

scRNA-seq profiling of individual cells, enabling the deciphering of intricate biological systems. Through the process of barcoding individual cells and creating cDNA libraries, this approach allows thousands to millions of cells to be processed in parallel, hence facilitating high-throughput single-cell transcriptomics. Droplet-based partitioning, which is used by the 10X Genomics platform, collects cellular transcriptomes in nanoliter-sized droplets, making it possible to isolate, barcode, and then sequence individual cells. By using this method, scientists can investigate the diversity of cells within tissues, find uncommon cell populations, and define signatures of gene expression that are particular to a given cell type with a high degree of sensitivity and throughput.

In cancer research, single-cell RNA sequencing (scRNA-seq) has been a game-changer, providing deep insights into the complex dynamics of tumor ecosystems(Imodoye et al., 2024). With the use of this technique, researchers can look at individual tumor cells, giving them a thorough grasp of the variety and complexity of these structures. In traditional methods, it was possible to miss significant changes between a mixture of cells when studying them collectively. But scRNA-seq goes farther, allowing scientists to examine the genetic makeup of each individual cell independently. By doing this, researchers can find different cell kinds inside tumors, comprehend how each type of cell works differently, and pinpoint particular genetic markers that are particular to each type of cell.

Machine learning (ML) plays a pivotal role in the analysis of single-cell RNA sequencing (scRNA-seq) data, especially in the field of cancer research(Vrahatis et al., 2020). Researchers can now interpret enormous volumes of complex biological data from scRNA-seq experiments with more ease because to this potent computational method. Molecular recognition, molecular signatures, cell type classification, and cancer cell behavior prediction are all accomplished through the use of ML algorithms. Significant biological insights are extracted from scRNA-seq data by applying machine learning (ML) models, including clustering algorithms (k-means, DBSCAN, hierarchical clustering), dimensionality reduction techniques (PCA, t-SNE, UMAP), and classification techniques (random forests, support vector machines).

Seurat is a popular R tool made especially for handling, examining, and displaying results from single-cell RNA sequencing(Slovin et al., 2021). It is especially useful in the study of cancer since it offers a set of instruments and algorithms designed specifically for scRNA-seq analysis. Seurat gives researchers the ability to preprocess scRNA-seq data in the context of cancer by carrying out quality control, normalization, and filtering procedures to guarantee high-quality data for further analysis. It makes it easier to identify cell clusters, shows how various cell types relate to one another within the tumor ecosystem, and describes the gene expression profiles linked to particular cell populations. Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE), two of Seurat's dimensionality reduction techniques, enable the display of intricate cellular landscapes in cancer tissues. The heterogeneity and spatial structure of cells within tumors are revealed by Seurat, which helps reduce the dimensionality of the data while maintaining vital biological information. This information is essential for comprehending the course of cancer, immune responses, and treatment outcomes. Seurat also integrates with a variety of ML algorithms and statistical methods, allowing researchers to find regulatory networks or signaling pathways related to cancer biology, perform differential expression analysis, and identify marker genes unique to particular cell types or states.

The complexity of various cancer types can significantly differ. There are several distinct diseases that make up cancer, each with distinct features of its own. Certain cancers may be more complex due to the presence of multiple cell types, similar to the multiple colors in a painting. Different cancer types can also exhibit differences in the way cells interact and behave within a tumor. This variety is influenced by various factors, including the cancer cells' growth rate, ability to disseminate, and response to various treatments. It is important for medical professionals and researchers to comprehend these variations in complexity as it aids in the development of efficient methods for the diagnosis, treatment, and management of each kind of cancer.

In our study, we gathered scRNA-seq datasets representing lung and breast cancer, sourced from publicly available the 10X Genomics Database (https://www.10xgenomics.com/products/single-cell-gene-expression). Using single-cell RNA sequencing data and Seurat analysis, we compared the complexity of lung and breast cancer. To do this, we first clustered similar cells from several datasets. After obtaining these clusters, we examined the genes known as marker genes that were highly expressed in each one. We removed the remaining gene data for each cluster in order to concentrate on these particular genes. Next, we investigated the relationships or connections between these filtered genes within the clusters. We constructed networks as dataframes that represented the clusters for each type of cancer using these linkages. We were able to learn more about the intricacy of both tumors by looking at and comparing these networks. This method is unique in

that it gives us a clearer understanding of the complexity of the cancers by allowing us to directly observe how genes interact within specific groups of cells related to the cancers. Other methods may not go as deeply into these particular relationships and interactions between genes in different types of cancer.

2. Methodology

2.1 Data Preprocessing and Quality Control

We initially gathered six distinct single-cell RNA sequencing (scRNA-seq) filtered datasets from the 10X Genomics Database (https://www.10xgenomics.com/products/single-cell-gene-expression). Of these, three were related to breast cancer and three to lung cancer, considering one as a validation dataset for each cancer and other two as training datasets. In Table1, we presented a thorough overview of the datasets.

Cancer Type	Dataset No.	Slide Serial Number	Chemistry	Mean Reads per Spot	Median Genes per Spot	Number of spots under tissue	Remark
Breast	1	V11J26-008-B1	Spatial 3' v1	32,524	5,244	2,518	Training data
	2	V19L29-095-A1	Spatial 3' v1	72,436	3,671	4,325	Training data
	3	V19B23-014-A1	Spatial 3' v1	47,223	3,654	4,898	Validation data
Lung	1	V42A20-354-D1	Visium V4 Slide - FFPE v2	35,785	6,174	3,858	Training data
	2	V52Y10-286-B1	Visium V5 Slide - FFPE v2	1,36,379	10,087	6,195	Training data
	3	N/A	Single Cell 5' R2-only	26,051	1,459	2,616	Validation data

Table 1. Overall description of the cancer datasets

We began our analytical pipeline by importing each dataset into R studio and converting it into a Seurat Object in order to enable comprehensive analysis. Each dataset was transformed into a Seurat Object independently, meeting the requirements of at least five cells and two hundred features. A Seurat Object, which functions as a structured container to store and handle single-cell RNA sequencing (scRNA-seq) data within the R environment, was created using the CreateSeuratObject() function in Seurat (Ji et al., 2019). By arranging the unstructured datasets, it produces a structured Seurat Object. The elements that comprise the Seurat Object are gene expression counts, cell metadata, gene metadata, and other pertinent data. Thus the metadata of the Seurat object is divided into four columns: the first column lists the cell names; the second column contains the dataset's identification; the third and fourth columns list the number of RNA molecules and genes that are present in each dataset cell. The function generates the Seurat Object by organizing and combining these dissimilar components. This object is subsequently utilized in other research projects like as quality control, normalization, clustering, dimensional reduction, and differential gene expression analysis (Slovin et al., 2021). This made it possible for us to concentrate on data on single-cell gene expression from several cancer datasets. Each dataset was transformed into a Seurat Object independently, meeting the requirements of at least five cells and two hundred characteristics.

Seurat included a quality control step to verify the general quality of the datasets. We omitted this step as we used filtered datasets from the 10X database and thought their quality was already good.

Normalization

We modified the gene expression data in the Seurat object using the normalizing approach in order to enhance the distribution's symmetry and lessen the impact of extraordinarily high expression levels. To ensure data comparability and stabilize variance, we applied log-normalization to normalize the dataset's Seurat Object (Stuart et al., 2019). It became simpler to compare the expression values between cells by compressing the data and lowering the dynamic range of expression. In our investigation, we performed log normalization using Seurat's NormalizeData() tool (Hrault et al., 2021). This procedure was crucial to producing a well standardized and reliable dataset, which enabled us to do perceptive downstream analysis and evaluate the results. By employing log normalization, we were able to effectively address the issues brought about by varying expression levels while also ensuring that the data was appropriately scaled for further examination.

Identification of high variable features

We took the top 3000 high variable characteristics out of the dataset after normalizing it (Fig.1). Finding highly variable genes is similar to highlighting the most important characteristics within a large dataset. By focusing on these highly variable genes, the analysis is made simpler rather than having to sort through all the genes. These genes are crucial for distinguishing between various cell types and reflect notable cell-to-cell changes. To enable a more effective and comprehensible downstream analysis, including PCA, grouping, and visualization (UMAP), we concentrated on the highly variable ones. Following the log-normalization of the scRNA-seq data, we used Seurat's FindVariableFeatures() function with the selection method set to variance stabilizing transformation (VST) to identify genes that exhibit significant variation in their expression levels among different cells (Salazar-Martín et al.,2023).

Scaling data

We initially employed log-normalization to control the expression distribution's complexity and stabilize variance. We utilized scaleData to equalize expression within cells following log-normalization in an attempt to lessen technical variations (Stuart et al., 2019). This phase ensured that each gene is expressed consistently in all cells, which was necessary for precise analyses like clustering. Scaling made it easier to find patterns that have biological significance by increasing the consistency of the gene contribution to variability. Reducing technological noise and improving the representation of actual biological differences in gene expression levels was our aim while scaling. Both scaling and log-normalization helped to clean up the data in a complementary way, improving analytical precision and revealing more about the interactions between genes and the behavior of cells. As such, our methodical approach ensured that data quality was optimized for robust biological interpretations. To scale our gene expression data inside individual cells using the single-cell RNA sequencing (scRNA-seq) dataset, we utilized Seurat's scaleData function (Stuart et al., 2019). This function aims to achieve centered and scaled gene expression levels by applying mathematical procedures to each cell. Typically, these actions include dividing each cell by the standard deviation (or another scaling factor) after obtaining the mean. To ensure more accurate downstream analysis and less technical variability, the function standardizes the gene expression statistics.

2.2 Dimensionality reduction

Principal Component Analysis (PCA)

In specifically, we performed linear dimensionality reduction using Principal Component Analysis (PCA) on the top 3000 variable features that were found in the preceding stage. This was a critical step that made it possible to simplify the dataset without losing any of its vital biological information. Through PCA, we were able to identify main components that represent the most important causes of variance in gene expression. We reduced the linear dimensionality of the data using the Principal Component Analysis (PCA) method provided by Seurat's RunPCA() function (Zhong et al., 2021). The top 3000 variable characteristics of every dataset that we had previously extracted were the targets of the PCA calculation. Elbow plot (Fig.2) was used to determine the optimal number of main components to retain (Table2) for further analysis, such as clustering, visualization, or extra dimensionality reduction approach (UMAP). With regard to the number of PCs in a dataset, the elbow plot displays the standard deviation. The standard deviation values on the plot demonstrated a steady drop with each new principle component. My goal was to locate this curve's "elbow," or the point at which the standard deviation dropped noticeably more slowly.

Clustering

It was crucial to categorize cells based on comparable expression profiles after applying principal component analysis (PCA) to decrease the dimensionality of our single-cell RNA sequencing dataset and identify the most meaningful principal components. Through the technique of clustering, we were able to classify cells into distinct groups or clusters according to the patterns of gene expression. Finding and differentiating between cell groups with similar expression profiles or biological characteristics was the aim of this clustering step. Using reduced dimensions (derived from PCA) to cluster cells based on expression similarity, we tried to find innate cellular groupings or groups within the dataset (Fig.3). Several mathematical calculations are used in this clustering method, which makes use of Seurat's FindNeighbors() function (Grabski et al.,2023).

The FindClusters() function was used following the creation of a neighborhood network using FindNeighbors() (Grabski et al.,2023). This function divides cells into distinct clusters based on their proximity to each other in

the reduced-dimensional space. We implemented the Louvain algorithm as a clustering tool on the generated neighborhood graph. The Louvain algorithm optimizes the network structure to identify cohesive communities or cell clusters. Maximizing modularity, a measure of the quality of cluster assignments made inside the network, is the aim of clustering algorithms. Modularity is a measure of the number of intra-cluster connections relative to the number of inter-cluster connections. The resolution input is used as a tuning option by Seurat's FindClusters() function to alter the cluster identification granularity (Grabski et al.,2023). We individually changed the resolution parameter for each dataset to create six distinct clusters in each set (Table2). This change was intended to result in a more complete data segmentation, which would enable a more comprehensive analysis of both datasets. These six clusters from each dataset will serve as the foundation for our future studies, allowing us to look more closely at the complexity both inside and across the datasets.

Cancer Type	Dataset No.	Number of cells	Number of genes	Number of PCs	Resolution parameter	Remark
Breast	1	2518	15946	12	0.3	Training data
	2	4169	19673	9	0.3	Training data
	3	4895	20227	12	0.15	Validation data
Lung	1	3830	18053	10	0.15	Training data
	2	6195	18067	12	0.2	Training data
	3	2521	19708	14	0.1	Validation data

Table 2. Number of cells, genes, number of PCs and the resolution parameter for cluster formation in the datasets after converting into seurat object.

Uniform Manifold Approximation and Projection (UMAP)

After dividing cells into groups based on the levels of gene expression using clustering approaches that came after PCA, we searched for a more complex and nuanced representation of the data. While PCA was successful in decreasing dimensions and capturing key sources of variation, UMAP provided an alternate perspective by highlighting non-linear interactions and preserving both local and global patterns within the data. Unlike PCA, UMAP excels at capturing the complex relationships and multidimensional structures found in high-dimensional data. It offers a more thorough depiction of the cellular landscape by projecting the cells into a low-dimensional space while preserving their intrinsic structure and relationships with one another based on comparable gene expression patterns. After clustering, we used UMAP to provide a more comprehensive picture of the cellular heterogeneity inside and between the identified clusters. We were able to investigate and depict the interactions between cell groups in a more thorough and physiologically meaningful way thanks to our non-linear reduction approach. For our scRNA-seq datasets, we used Seurat's RunUMAP() tool to perform UMAP (Uniform Manifold Approximation and Projection) (Massier et al., 2023). There are several downstream investigations that may be conducted using the reduced-dimensional coordinates obtained by UMAP. These include grouping, identifying marker genes, and studying cellular transitions and heterogeneity. We assigned numbers ranging from 0 to 7 to each of the eight clusters, so naming them.

2.3 Marker genes identification

We discovered the marker genes inside the clusters to gain a better understanding of the distinct characteristics and identities of different cell types. Determining the genetic fingerprints that characterize each cluster's identity and function required first identifying and then removing marker genes from within each cluster. Important insights into the many biological processes and traits of different cell populations in our dataset were obtained from these processes. To find specific markers in our dataset that are required for certain groups or clusters, we utilized Seurat's FindConservedMarkers() tool (Prazanowska et al.,2023). Marker genes serve as genetic fingerprints that distinguish different cell types or subgroups based on patterns of gene expression. These markers provided precise cell identification, revealing biological variety, providing functional insights, and guiding more research. The FindConservedMarkers() function searches for genes that have differential expression in a particular cluster relative to other clusters or groups, often assessing one group against all other.

2.4 Correlation based network dataframe analysis

The first step in our procedure was to separate and collect the gene expression data unique to each cluster in the dataset we were examining. In order to analyze the distinct gene expression patterns of various cell groupings, or clusters, we isolated the gene activity profiles that are particular to those clusters. We concentrated on unique genes, known as marker genes, that were strongly expressed and connected to each cluster. We condensed our data by retaining only this particular gene information for each cluster and eliminating the others in order to have a better understanding of these crucial genes. Therefore, we were

interested in understanding the interactions or cooperative mechanisms among these important genes within each cluster. We looked at the degree of relationship between these filtered genes inside the clusters to investigate this. This required figuring out a metric known as correlations, which gave us insight into the strength of the connections between these genes and whether or not their activity altered simultaneously.

In order to create the correlation, the rcorr() function from the R Hmisc library was utilized (Alexa et al., 2022). For pairwise comparisons between the marker genes inside each cluster, this function effectively computed the Pearson correlation coefficients(r)and their corresponding values.

$$r = \frac{\sum(A_i-\bar{A})((B_i-\bar{B})}{\sqrt{\Sigma(A_i-\bar{A})^2\Sigma(B_i-\bar{B})^2}}$$

For the 'i-th' sample, the expression levels of Gene A and Gene B are, respectively, A_i and B_i.

The means (averages) of Gene A and Gene B's expression levels across all samples are represented by the numbers $\bar{A}$ and $\bar{B}$.

To determine if a correlation coefficient is significant, the function used the following formula to get the t-statistic:

$$t = r \times \sqrt{\frac{n-2}{1-r^2}}$$

r = Pearson correlation coefficient

n = Number of genes

The cumulative distribution function (or CDF) of the t-distribution is used in the following mathematical equation to get the $p-value$ from the t-statistic:

$$p-value = P(T \geq |t| \; or \; T \leq -|t|\,)$$

Where:

$|t|$ = Absolute value of the calculated t-statistic.

T = Random variable following the t-distribution with $n-2$ degrees of freedom.

$P(T \geq |t| \; or \; T \leq -|t|\,)$ = Probability of observing a t-value as extreme as $|t|$ or more extreme in a two-tailed test.

We chose a threshold of $p < 0.05$ for significance across all datasets in order to provide a strict standard. We were able to find statistically significant gene correlations within each cluster by using this level.

In the initial phase, we constructed a correlation matrix depicting pairwise correlations among genes inside a certain cluster, where the correlation coefficient 'r' between. We make sure that no gene is linked with itself by setting the diagonal elements of the correlation matrix to zero in order to remove self-correlations. In order to distinguish non-correlation values, a placeholder value of 25 was used to replace the upper triangle members while preserving diagonal symmetry. The gene correlation matrix was then transformed using the 'melt()' function into a long or melted format, which preserved information about gene pairings and their correlation values while transforming rows and columns into a two-column format. This produced a new data frame in which each row represents a gene pair ('variable1' and 'variable2') together with the corresponding significance value. After using filtering procedures to eliminate superfluous items, such as self-correlations and placeholder values, a revised dataset emphasizing noteworthy correlations was produced. The method then moved on to the next step, which was building the correlation network. First, to find gene pairings with substantial correlations, the correlation data was filtered using a threshold of.5. From this filtered correlation data, an adjacency list useful for building the network was subsequently produced. Every row in this list denoted an edge between two genes, indicating a strong connection between these gene pairs that served as the foundation for the network. Following the generation of adjacency lists derived from the filtered correlation data, we transformed these lists into structured data frames, creating network data frames essential for our cancer complexity analysis. These network data frames were specifically structured to encode the relationships

between genes exhibiting significant correlations within the context of cancer datasets. We analysed network dataframes of different clusters of all the cancer datasets to understand the complexity in the datasets.

All additional clusters in the specific cancer dataset have been processed using the aforementioned processes. To create networks based on gene correlation for every cluster, we used many R packages, such as reshape2, tibble, and dplyr.

3. Results and Discussion

3.1 Variable features and Principal components identification

We figured out top 3000 variable features in each cancer dataset. We mentioned top 5 high variable features in each cancer dataset in the Table3. We used Seurat's FindVariableFeatures() function with the selection method set to variance stabilizing transformation (VST) to identify genes that exhibit significant variation in their expression levels among different cells. We then used these top 3000 variable features in the PCA for the number of PCs calculation. We reduced the linear dimensionality of the data using the Principal Component Analysis (PCA) method provided by Seurat's RunPCA() function (Zhong et al., 2021).

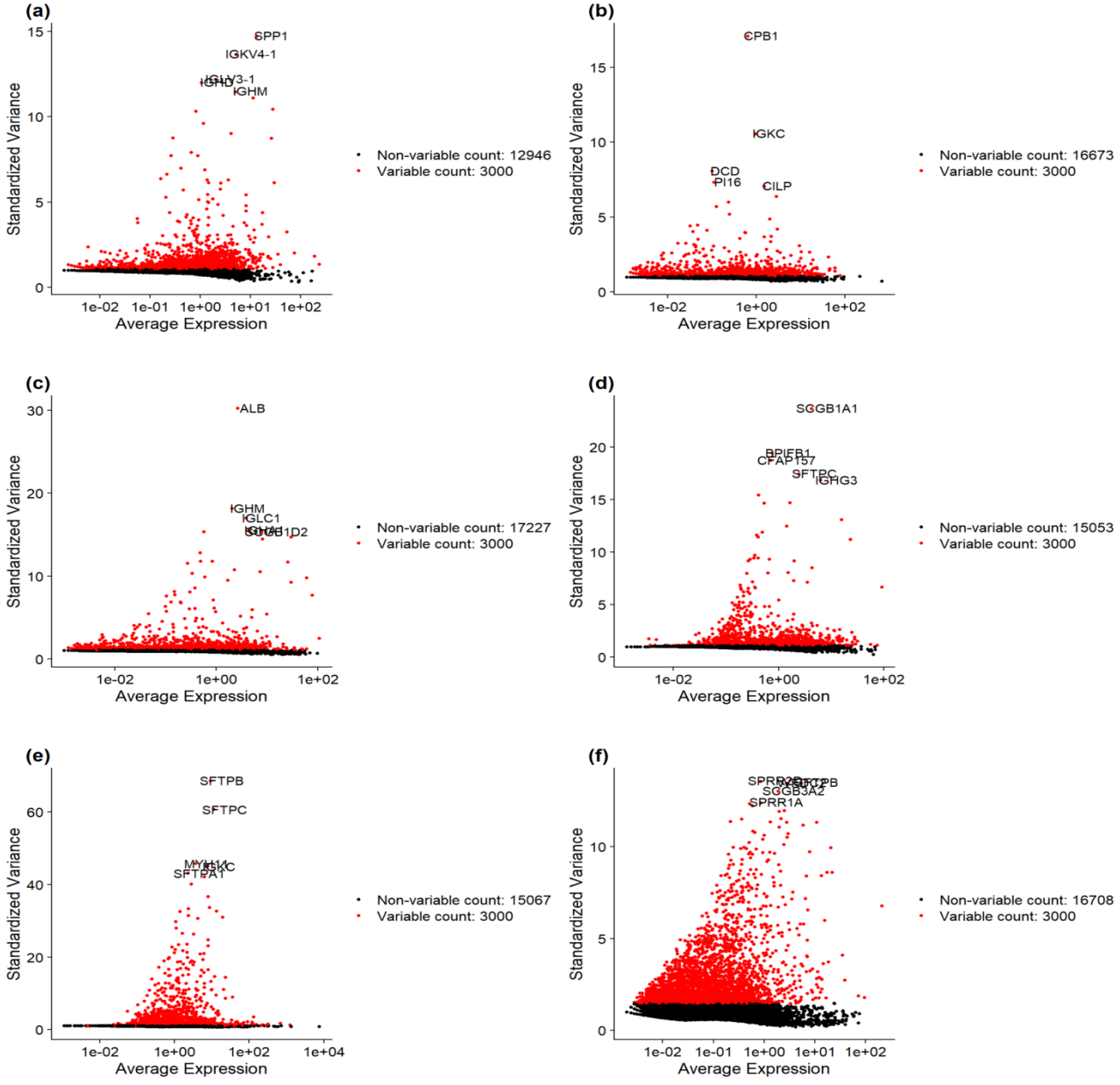

Fig.1: Plots showing high variable features(genes) present in the datasets. (a) Breast cancer Dataset 1 (b) Breast cancer Dataset 2 © Breast cancer Dataset 3 (d) Lung cancer Dataset 1 (e) Lung cancer Dataset 2 (f) Lung cancer Dataset 3

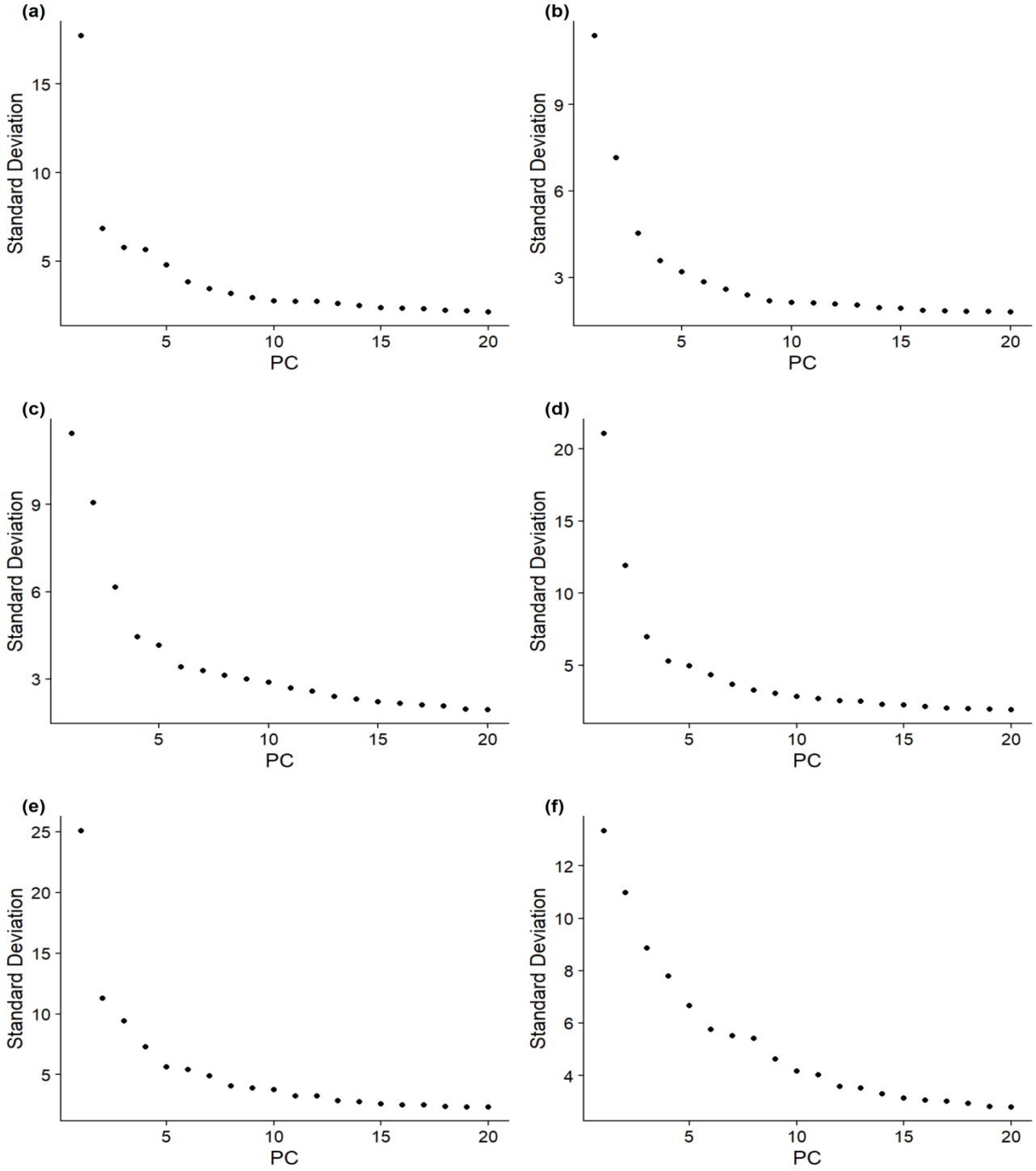

Fig.2: Elbow plots showing PC wise standard deviations present in the datasets. (a) Breast cancer Dataset 1 (b) Breast cancer Dataset 2 © Breast cancer Dataset 3 (d) Lung cancer Dataset 1 (e) Lung cancer Dataset 2 (f) Lung cancer Dataset 3

We used Elbow plot (Fig.2) to determine the optimal number of main components to retain (Table2) for further analysis, such as clustering, visualization, or extra dimensionality reduction approach (UMAP). With regard to the number of PCs in a dataset, the elbow plot displays the standard deviation. The standard deviation values on the plot demonstrated a steady drop with each new principle component. My goal was to locate this curve's "elbow," or the point at which the standard deviation dropped noticeably more slowly. The number of PCs we figured out in each dataset are mentioned in the Table2.

Cancer Type	Dataset No.	1	2	3	4	5	Remark
Breast	1	SPP1	IGKV4-1	IGLV3-1	IGHD	IGHM	Training data
	2	CPB1	IGKC	DCD	PI16	CILP	Training data
	3	ALB	IGHM	IGLC1	IGHA1	SCGB1D2	Validation data
Lung	1	SCGB1A1	BPIFB1	CFAP157	SFTPC	IGHG3	Training data
	2	SFTPB	SFTPC	MYH11	IGKC	SFTPA1	Training data
	3	SPRR2D	SFTPB	WFDC2	SCGB3A2	SPRR1A	Validation data

Table 3. Top 5 high variable genes present in the datasets

3.2 Cluster visualization

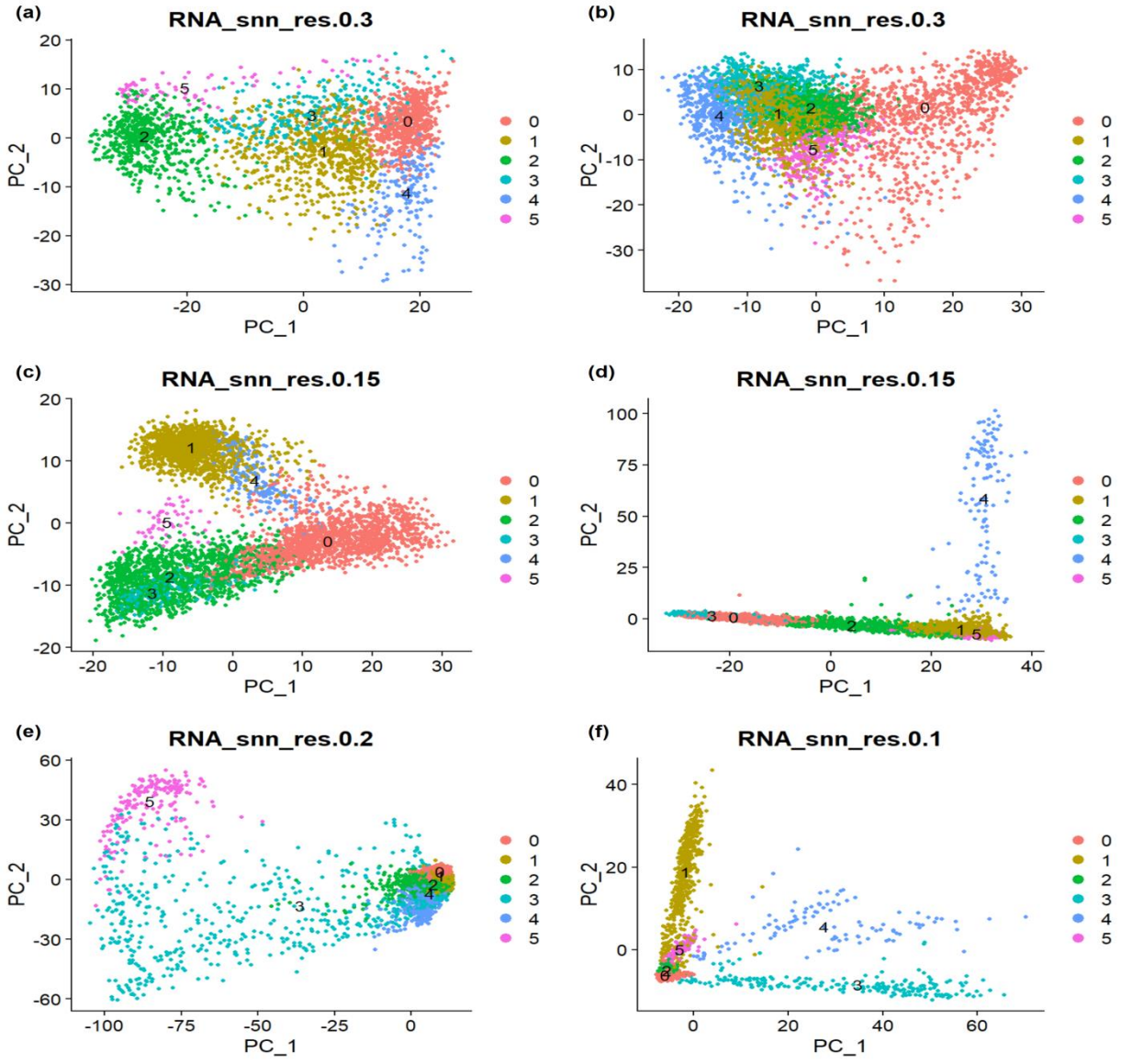

Fig.3: Plots showing six different clusters obtained after performing PCA in the datasets. (a) Breast cancer Dataset 1 (b) Breast cancer Dataset 2 © Breast cancer Dataset 3 (d) Lung cancer Dataset 1 (e) Lung cancer Dataset 2 (f) Lung cancer Dataset 3

The FindClusters() function was used following the creation of a neighborhood network using FindNeighbors() (Grabski et al.,2023). This function divides cells into distinct clusters based on their proximity to each other in the reduced-dimensional space. The resolution input is used as a tuning option by Seurat's FindClusters() function to alter the cluster identification granularity (Grabski et al.,2023). We individually changed the resolution parameter for each dataset to create six distinct clusters in each set (Table2). This change was intended to result in a more complete data segmentation, which would enable a more comprehensive analysis of the datasets. These six clusters from each dataset served as the foundation for our future studies, allowing

us to look more closely at the complexity both inside and across the datasets. Fig.3 provides us the visualization of clusters obtained after performing PCA technique.

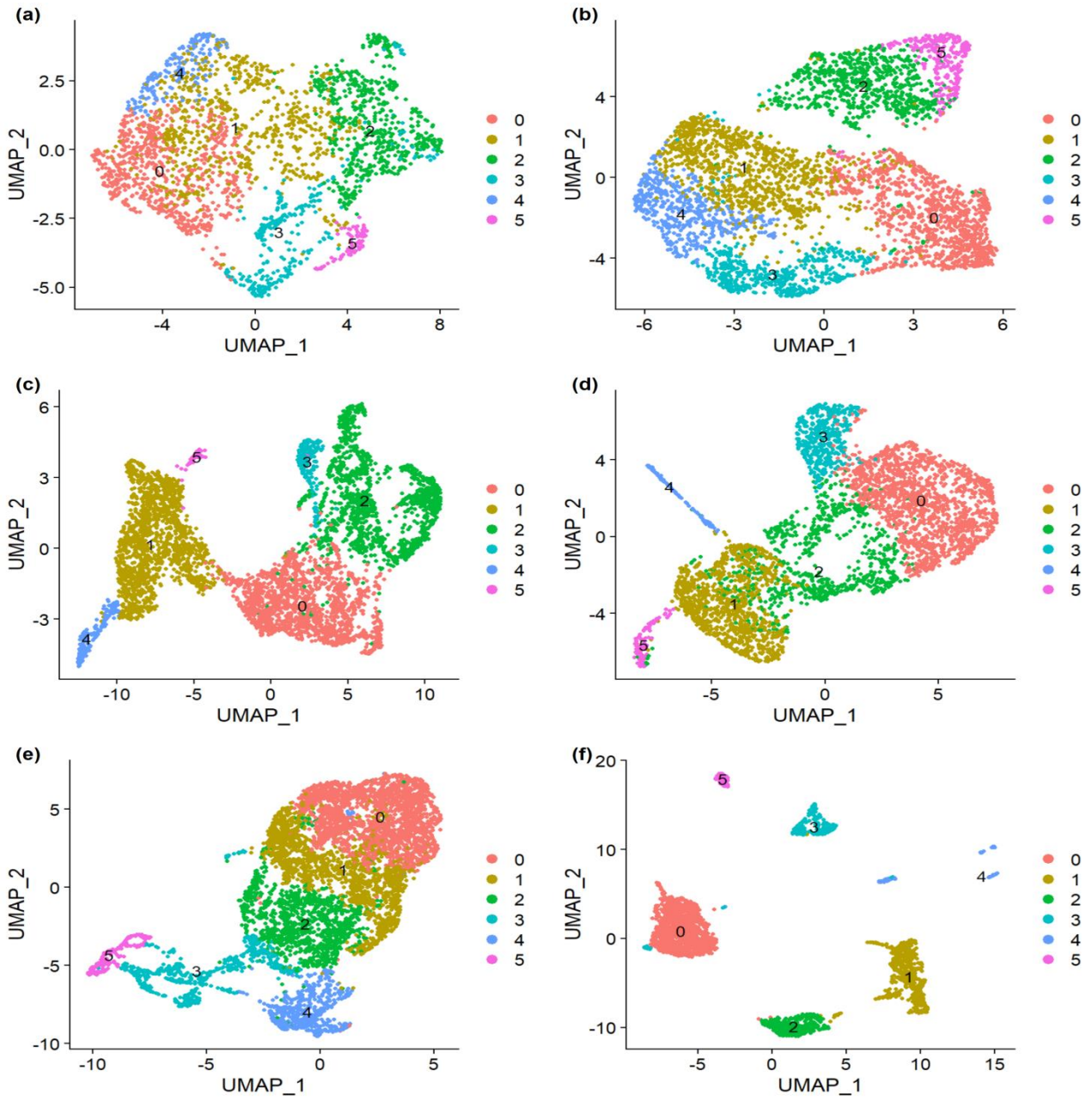

Fig.4: Plots showing six different clusters obtained after performing UMAP in the datasets. (a) Breast cancer Dataset 1 (b) Breast cancer Dataset 2 © Breast cancer Dataset 3 (d) Lung cancer Dataset 1 (e) Lung cancer Dataset 2 (f) Lung cancer Dataset 3

We used UMAP to provide a more comprehensive picture of the cellular heterogeneity inside and between the identified clusters. We were able to investigate and depict the interactions between cell groups in a more thorough and physiologically meaningful way. For our scRNA-seq datasets, we used Seurat's RunUMAP() tool to perform UMAP (Uniform Manifold Approximation and Projection) (Becht et al., 2019). Fig.4 provides us the visualization of clusters obtained after performing UMAP.

3.3 Marker genes identification

To find specific markers in our dataset that are required for certain groups or clusters, we utilized Seurat's FindConservedMarkers() tool (Prazanowska et al.,2023). Marker genes serve as genetic fingerprints that distinguish different cell types or subgroups based on patterns of gene expression. These markers provided

precise cell identification, revealing biological variety, providing functional insights, and guiding more research. The FindConservedMarkers() function searches for genes that have differential expression in a particular cluster relative to other clusters or groups, often assessing one group against all others. We determined the number of marker genes in each cluster of every dataset and figured out the result that the number of marker genes obtained in the lung cancer datasets were pretty high as compared to the breast cancer datasets(Table4).

In order to fully describe and comprehend the intricate web of gene expression found in biological datasets, marker genes are essential. By acting as markers or signatures connected to particular cell types or subpopulations, these genes let scientists recognize and distinguish between various cell types. The quantity of marker genes found in datasets derived from various tissues or illnesses can shed light on the intrinsic heterogeneity or variety present in those samples. More marker genes are frequently indicative of a more complex and diverse cellular makeup. Variations in the microenvironment or the existence of unique cell types or subtypes could be the cause of this variety. In our analysis, a higher number of marker genes indicated a greater heterogeneity or diversity in the cellular composition of the lung cancer samples.

Cancer Type	Dataset No.	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Remark
Breast	1	1896	292	2962	647	2000	2013	Training data
	2	1267	165	374	565	470	471	Training data
	3	1865	1317	1114	1396	466	1250	Validation data
Lung	1	3085	3726	534	2441	4302	4248	Training data
	2	629	434	258	2741	493	4586	Training data
	3	2814	3148	2064	4133	3500	3254	Validation data

Table 4. Number of marker genes present in the identified clusters of the datasets

We showed the number of marker genes present in the clusters of different datasets of breast and lung cancer by using bar plot(Fig.5). These bar plots helped us to compare the marker genes distribution visually in different datasets. We observed that the marker genes in lung cancers were comparatively higher than the breast cancer datasets.

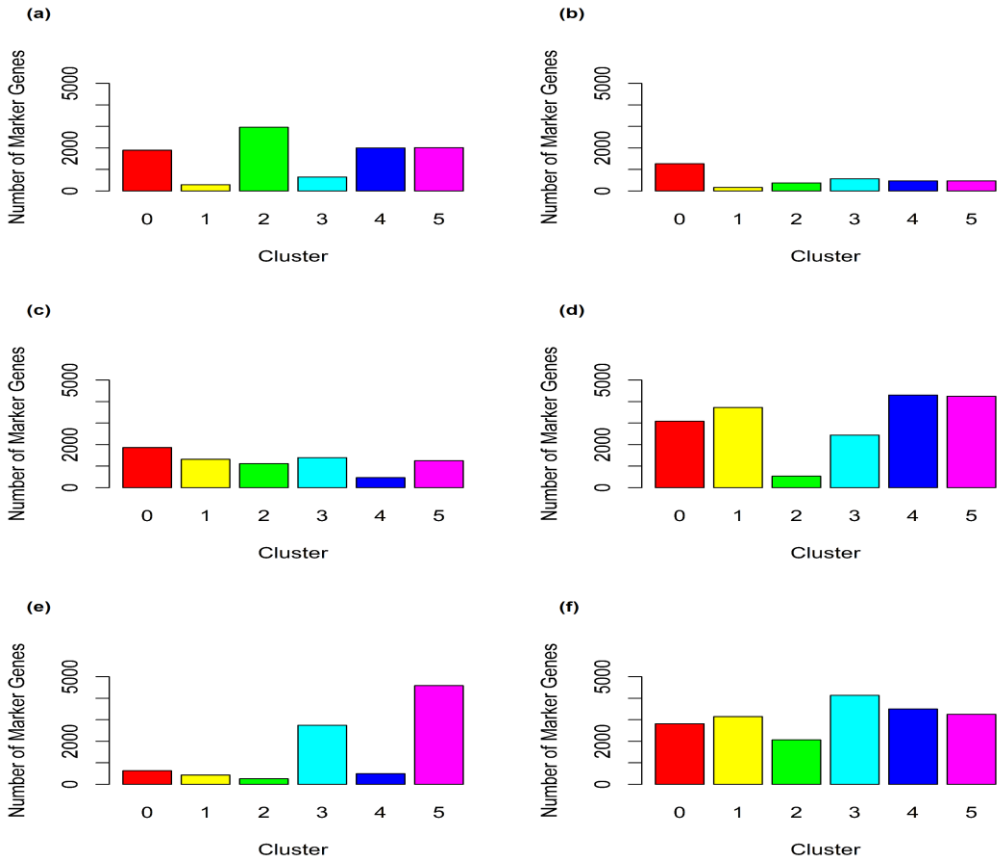

Fig.5: Bar Plots showing distribution of marker genes in the six different clusters in each dataset. (a) Breast cancer Dataset 1 (b) Breast cancer Dataset 2 © Breast cancer Dataset 3 (d) Lung cancer Dataset 1 (e) Lung cancer Dataset 2 (f) Lung cancer Dataset 3

3.4 Gene correlation networks analysis

We formed gene correlation network dataframe for each cluster at gene correlation threshold of 0.5 as described in the methodology section. These network dataframes highlight genes with an absolute correlation value greater than 0.5 and depict connections between genes through edges. The presence or absence of edges in the network dataframe indicates whether there is a correlation between the respective genes. This approach enables the visualization and exploration of relationships within and between gene clusters, offering valuable insights into the co-expression patterns and potential functional associations among genes in our analysis. We determined how many genes(Table4) and edges(Table5) were in each cluster's network by counting them in the network dataframe. A noteworthy finding surfaced during analysis: compared to the breast cancer datasets, the lung cancer datasets showed noticeably more edges and genes.

In particular, a higher number of edges in lung cancer networks indicates a more complex interaction and relationship between genes. The molecular landscape of lung cancer may exhibit greater levels of coordination and potential regulatory connections, as indicated by the links between genes becoming increasingly complicated. Larger gene counts and a more extensive network of edges in the lung cancer networks suggest that the disease include a variety of molecular subtypes, complex signaling pathways, or higher levels of heterogeneity among lung cancer samples. This increased complexity might have an impact on our comprehension of the biology underlying lung cancer and could guide future research into the disease's molecular underpinnings.

Cancer Type	Dataset No.	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Remark
Breast	1	102	83	356	105	49	1584	Training data
	2	4	5	31	30	190	35	Training data
	3	69	82	114	75	29	425	Validation data
Lung	1	266	49	221	422	1294	1334	Training data
	2	20	16	138	2536	570	981	Training data
	3	358	648	519	2919	3819	2800	Validation data

Table 5. Number of genes found in the networks of six different clusters of each dataset

Cancer Type	Dataset No.	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Remark
Breast	1	191	102	999	257	44	6674	Training data
	2	2	3	41	34	552	46	Training data
	3	83	116	154	119	17	338	Validation data
Lung	1	985	99	839	2018	18295	1321	Training data
	2	14	11	632	118070	6102	7269	Training data
	3	1085	4501	3506	15117	22507	7404	Validation data

Table 6. Number of edges found in the networks of six different clusters of each dataset

We analysed six distinct single-cell RNA sequencing (scRNA-seq) cancer datasets where three datasets were related to breast cancer and three to lung cancer gathered from 10X Genomics Database. Among the datasets of each cancer type, one was validation dataset and the other two were training datasets. The results obtained in the training datasets were verified using the validation datasets. We thus compared the lung and breast cancer datasets by analysing them visually and also analytically.

4. Conclusion

Finally, our work explored the molecular subtleties of lung and breast malignancies using sophisticated single-cell RNA sequencing (scRNA-seq) analysis. We processed six different datasets from the 10X Genomics Database, three for lung and three for breast cancer, using a strict analytical approach. We used a rigorous preprocessing, quality control, normalization, Principal Component Analysis (PCA) to reduce dimensionality, clustering, and Uniform Manifold Approximation and Projection (UMAP) for visualization. In order to understand the intricate landscape of gene expression in these malignancies, we subsequently discovered marker genes and investigated gene correlation networks. A noteworthy finding was that the gene correlation networks of lung cancer datasets had more edge and marker genes than those of breast cancer datasets. This discrepancy points to a more complex molecular profile and higher variation among the lung cancer samples. The presence of many cell types or subpopulations, potentially indicating different cancer subtypes or differences in the tumor microenvironment, is suggested by the abundance of marker genes. Our

comprehension was deepened by the use of UMAP for visualization, which made clear the intricate linkages and multidimensional structures present in high-dimensional data. By using this method, we were able to project cells into a low-dimensional environment while maintaining global and local patterns based on similarity in gene expression. The generated visuals offered a detailed depiction of the variety of cells both inside and amongst the detected clusters. The intricacy of lung cancer datasets was highlighted even further by our examination of gene correlation networks. These networks' greater edge counts indicate more complex gene-to-gene interactions and relationships. Numerous molecular subtypes, intricate signaling pathways, or greater variability amongst lung cancer samples could all be responsible for this complexity. The results indicate that lung cancer, as represented by our datasets, might have a more complex genetic landscape than breast cancer.

In conclusion, by offering a thorough examination of the gene expression patterns in lung and breast tumors, our study adds to the expanding corpus of knowledge pertaining to cancer biology. The discovered gene correlation networks and marker genes provide information about the molecular mechanisms and heterogeneity that underlie these malignancies. This data may have implications for personalized medicine by facilitating the creation of focused treatment plans that are customized to each patient's unique molecular signature. To improve clinical outcomes and expand our knowledge of cancer biology, it will be imperative to do additional research on the identified marker genes and intricate gene interactions.

5. References

1. Alexa, E., Cobo-Diaz, J. F., Renes, E., FO, T., Kilcawley, K., Mannion, D., ... & Alvarez-Ordóñez, A. (2022). Environmental sources along natural cave ripening shape the microbiome and metabolome of artisanal blue-veined cheeses.
2. Grabski, I. N., Street, K., & Irizarry, R. A. (2023). Significance analysis for clustering with single-cell RNA-sequencing data. *Nature Methods*, *20*(8), 1196-1202.
3. Imodoye, S. O., Adedokun, K. A., & Bello, I. O. (2024). From complexity to clarity: unravelling tumor heterogeneity through the lens of tumor microenvironment for innovative cancer therapy. *Histochemistry and Cell Biology*, 1-25.
4. Ji, F., & Sadreyev, R. I. (2019). Single-Cell RNA-seq: Introduction to Bioinformatics Analysis. *Current protocols in molecular biology*, *127*(1), e92.
5. Massier, L., Jalkanen, J., Elmastas, M., Zhong, J., Wang, T., Nono Nankam, P. A., ... & Mejhert, N. (2023). An integrated single cell and spatial transcriptomic map of human white adipose tissue. *Nature Communications*, *14*(1), 1438.
6. Prazanowska, K. H., & Lim, S. B. (2023). An integrated single-cell transcriptomic dataset for non-small cell lung cancer. *Scientific Data*, *10*(1), 167.
7. Salazar-Martín, A. G., Kalluri, A. S., Villanueva, M. A., Hughes, T. K., Wadsworth, M. H., Dao, T. T., ... & Edelman, E. R. (2023). Single-Cell RNA Sequencing Reveals That Adaptation of Human Aortic Endothelial Cells to Antiproliferative Therapies Is Modulated by Flow-Induced Shear Stress. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *43*(12), 2265-2281.
8. Slovin, S., Carissimo, A., Panariello, F., Grimaldi, A., Bouché, V., Gambardella, G., & Cacchiarelli, D. (2021). Single-cell RNA sequencing analysis: a step-by-step overview. *RNA Bioinformatics*, 343-365.
9. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., ... & Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, *177*(7), 1888-1902.
10. Vrahatis, A. G., Tasoulis, S. K., Maglogiannis, I., & Plagianakos, V. P. (2020). Recent machine learning approaches for single-cell rna-seq data analysis. *Advanced Computational Intelligence in Healthcare-7: Biomedical Informatics*, 65-79.
11. Zhong, R., Chen, D., Cao, S., Li, J., Han, B., & Zhong, H. (2021). Immune cell infiltration features and related marker genes in lung cancer based on single-cell RNA-seq. *Clinical and Translational Oncology*, *23*, 405-417.