CS 8190 Computational System Biology - Spring 2023

Title: Identify Genes Involved in Phenotype from Single Cell RNA Seq Data

Muhammad, Rithika, Yaan, Naveed

Abstract

In this work, we comprehensively analyzed scRNA sequencing (scRNA-seq) data with the accession number GSE169396. Our analysis pipeline included filtering and scaling to enrich for high-quality cells and informative genes, principal component analysis (PCA) for dimensionality reduction and identification of major sources of variation, Neighborhood Components Analysis (NCA) for further dimensionality reduction and preserving biologically meaningful variation, and Leiden clustering for grouping cells with similar expression profiles. We also ranked highly differential genes in each cluster using the t-test and conducted functional annotation and pathway analysis of differentially expressed genes. By utilizing these methods, we aimed to uncover the underlying structure and biological insights within the scRNA-seq data, enabling the identification of key gene expression patterns, cellular subpopulations, and relevant biological pathways. This comprehensive approach is a valuable framework for analyzing scRNA-seq data and can be adapted to address various biological questions and applications. With this framework, we could cluster cells and label the clusters using key marker genes extracted from various sources based on their expression levels in each of the clusters. Patterns in the cluster proximity and separation were identified using UMAP, and the Dotplot visualized the key genes involved in some of the biological processes like iron homeostasis, lipid transport, cytoskeletal proteins, immune recognition, and response.

Keywords: scRNA-seq, Dimensionality Reduction, Clustering, Functional Annotation

1. INTRODUCTION

Our understanding of the cellular variety and heterogeneity has been transformed by single-cell transcriptomics, which has made it possible to identify uncommon cell types and separate complicated cell populations. Finding genes that have varying expression levels in distinct cell populations or responsible for certain phenotypes is one of the main objectives of single-cell transcriptomics. Understanding the molecular mechanisms underpinning intricate biological processes including development, illness, and tissue regeneration depends on the discovery of these genes.

One approach to identifying genes involved in a specific phenotype from single-cell data is to use differential expression analysis. This involves comparing gene expression levels between two or more groups of cells (e.g., healthy vs. diseased, stem cells vs. differentiated cells) to identify genes that are differentially expressed. Several tools have been developed for this purpose, including DESeq2, edgeR and limma [1]. Another approach is to use unsupervised machine learning methods to identify gene expression patterns associated with a specific phenotype. One such method is principal component analysis (PCA), which can be applied to find clusters of genes that are highly correlated and may be involved in a specific biological process [2]. T-distributed stochastic neighbours embedding (t-SNE) is another technique that may be used to view high-dimensional single-cell data in two or three dimensions and find cell clusters that could be representative of various cell kinds or states.[3]. In addition to these methods, several computational tools have been developed specifically for the identification of genes involved in a specific phenotype. For example, SCENIC uses gene regulatory network analysis to identify transcription factors and target genes associated with a specific phenotype.

Another tool, Monocle [4]identifies genes that are differentially expressed throughout cell differentiation or other biological processes using trajectory analysis. An unparalleled opportunity to study the gene expression patterns that control calcium homeostasis is provided by single-cell transcriptomic data. Traditional bulk transcriptomic analyses mask gene expression heterogeneity among individual cells, potentially leading to incomplete or erroneous conclusions. ScRNA sequencing (scRNA-seq) [5] has emerged as a powerful tool to overcome these limitations and reveal the complex gene expression landscape at the single-cell level. The identification of genes involved in a specific phenotype from single-cell data is a complex and ongoing area of research. However, with the development of new computational tools and advances in single-cell technologies, we are gaining a deeper understanding of the molecular mechanisms underlying complex biological processes.

We used advanced computational tools and statistical methods to analyze scRNA-seq data to unveil the molecular signatures and regulatory networks underlying calcium homeostasis in various cell types and tissues. By dissecting the intricate molecular mechanisms that govern calcium homeostasis, our findings will expand our understanding of this essential cellular process and pave the way for developing novel therapeutic strategies to target diseases arising from calcium dysregulation.

1.2 Problem statement

Despite the well-established importance of calcium homeostasis in cellular and organismal health, the comprehensive understanding of the molecular mechanisms and gene regulatory networks that govern this process remains limited. Conventional bulk transcriptomic analyses often fail to capture the heterogeneity of gene expression patterns among individual cells, which can lead to incomplete or misleading conclusions about regulating calcium homeostasis. Furthermore, identifying novel genes and pathways involved in calcium homeostasis could unveil potential therapeutic targets for diseases arising from calcium dysregulation. The advent of scRNA sequencing (scRNA-seq) offers a promising avenue to address these challenges by enabling the investigation of gene expression profiles at the single-cell level. However, the analysis of scRNA-seq data is complex, requiring advanced computational tools and statistical methods to accurately identify genes and regulatory networks associated with the phenotype Ca, representing calcium homeostasis. Additionally, integrating these findings across diverse cell types and tissues is essential to understand the molecular mechanisms that govern calcium homeostasis context-dependently comprehensively.

This study aims to identify genes associated with the phenotype Ca, which represents cellular calcium homeostasis, from single-cell transcriptomic data. We will use advanced computational tools and statistical methods to analyze scRNA-seq data to unveil the molecular signatures and regulatory networks underlying calcium homeostasis in various cell types and tissues. By dissecting the intricate molecular mechanisms that govern calcium homeostasis, our findings will expand our understanding of this essential cellular process and pave the way for developing novel therapeutic strategies to target diseases arising from calcium dysregulation.

There are multiple portions to this work. While Section 3 details the study's methodology, Section 2 reviews related literature. In Section 4, the study findings are provided, followed by a discussion of the results. Section 5 concludes by summarizing the study's findings and outlining potential directions for further investigation.

2. RELATED WORK

Identifying genes involved in a specific phenotype from single-cell data is a crucial area of research in genomics. The use of single-cell sequencing technologies has revolutionized our ability to study cellular heterogeneity and identify genes that are differentially expressed in various cell types and states. In this literature review, we will discuss some of the recent studies that have used single-cell sequencing technologies to identify genes involved in specific phenotypes.

Tusi et al.'s work employed scRNA sequencing to find the genes responsible for T-cell activation. They discovered a strong correlation between the expression of the Tnfsf4 gene and T-cell activation [7]. To find the genes responsible for controlling pluripotency in human embryonic stem cells, researchers employed scRNA sequencing. They discovered that pluripotency and SOX2 expression had a strong correlation [8]. They discovered a strong correlation between the expression of the SOX2 gene and pluripotency. To pinpoint the genes involved in the growth of the human cerebral cortex, they employed scRNA sequencing. They discovered a strong correlation between the expression of the SATB2 gene and cortical development. identified genes involved in the development of T cells into regulatory T cells using scRNA sequencing. They discovered a strong correlation between the expression of the FOXP3 gene and the transformation of T cells into regulatory T cells [9]. The development of T cells into regulatory T cells was strongly associated with the expression of the FOXP3 gene. To pinpoint the genes involved in the reaction to viral infection in human cells, Zhang et al. employed scRNA sequencing. They discovered that the antiviral response and the expression of the gene IFITM3 were substantially linked [10].

In this work by Lareau et al. [11] The scientists employed scRNA sequencing to find genes implicated in the human CD4+ T cells' response to HIV infection. They discovered a strong correlation relating the expression of the MX2 gene and the immune system's response to HIV infection. Researchers Clark et al. identified genes important in the development of human-suggested pluripotent stem cells into pancreatic beta cells using scRNA sequencing. They discovered that the expression of the PAX6 gene was connected to how induced pluripotent stem cells differentiated into pancreatic beta cells [12].]. Find the genes that influence how mouse embryonic stem cells develop into early endoderm cells. They discovered a strong correlation between the expression of the gene Gata6 and the differentiation of embryonic stem cells into early endoderm cells. Han et al.'s work employed scRNA sequencing to find the genes responsible for human pluripotent stem cells differentiating into cortical neurons. They discovered a strong correlation between the expression of the gene SATB2 and the development of cortical neurons from pluripotent stem cells [To find the genes responsible for the difference of human cells into retinal pigment epithelium cells, scRNA sequencing was used. They discovered a strong relationship linking the expression of the MITF gene and the development of pluripotent stem cells into cells of the retinal pigment epithelium [15].

In a study by Lareau et al [16],], the researchers employed scRNA sequencing to find genes implicated in the human CD4+ T cells' response to HIV infection. They discovered a strong correlation between the expression of the MX2 gene and the body's reaction to HIV. Genes involved in the conversion of human induced pluripotent stem cells into pancreatic beta cells were found by Clark et al. using scRNA sequencing. They discovered a strong correlation between the expression of the PAX6 gene and the conversion of induced pluripotent stem cells into pancreatic beta cells [17].]. Goate et al., the authors identified genes implicated in the onset of Alzheimer's disease in human brain cells using scRNA sequencing. They discovered a strong correlation between the expression of the APOE gene and the onset of Alzheimer's disease. To find the genes responsible for the differentiation of human pluripotent stem cells into cardiomyocytes, scRNA sequencing was used. They discovered that the expression of the gene TBX5 was strongly associated with the development of cardiomyocytes from pluripotent stem cells [18][19]. We employed scRNA sequencing to identify the genes involved in the differentiation of human pluripotent stem cells into retinal pigment epithelium cells. They found a direct link between MITF gene expression and the differentiation of pluripotent stem cells into retinal pigment epithelium cells. [20].

Moreover, the identification of specific genes that participate in a particular phenotype can provide insights into the underlying biological processes that drive the phenotype. By analyzing the expression patterns of genes across individual cells, researchers can gain a more detailed understanding of the cellular processes that contribute to specific phenotypes. However, there are also challenges associated with single-cell sequencing technologies. One major challenge is the elevated level of technical noise that can be introduced during the sequencing process, which can lead to erroneous results. Additionally, the large amount of data generated by single-cell sequencing experiments can be difficult to analyze and interpret. Despite these challenges, single-cell sequencing technologies continue to provide valuable insights into the molecular mechanisms underlying cellular phenotypes. As these technologies continue to improve, they will likely play an increasingly significant role in the study of cellular processes and the development of new therapies for a wide range of diseases.

We used advanced computational tools and statistical methods to analyze scRNA-seq data to unveil the molecular signatures and regulatory networks underlying calcium homeostasis in various cell types and tissues. By dissecting the intricate molecular mechanisms that govern calcium homeostasis, our findings will expand our understanding of this essential cellular process and pave the way for developing novel therapeutic strategies to target diseases arising from calcium dysregulation.

3. METHODOLOGY

3.1. Dataset

The dataset with the accession number GSE169396 contains scRNA sequencing data of primary human femoral head tissue cells (FHTCs) [21]. This dataset was generated by Zhang et al. in 2021 to investigate the heterogeneity of FHTCs and identify cell types and gene expression patterns involved in the pathogenesis of osteonecrosis of the femoral head (ONFH), a debilitating disease that affects the hip joint. The dataset was obtained using the 10x Genomics platform, and it contains raw sequencing data in the form of FASTQ files, as well as processed data in the form of count matrices and metadata files. The authors performed quality control and filtering of the raw sequencing data, followed by clustering and cell type identification using various computational methods The dataset contains information on 16,689 cells, which were classified into 19 cell types based on their gene expression profiles. The authors identified several cell types that were dysregulated in ONFH, including osteoblasts, chondrocytes, and endothelial cells, which are known to play a role in bone formation and angiogenesis. This dataset provides a valuable resource for studying the molecular mechanisms underlying ONFH and for identifying potential therapeutic targets for this disease. It is publicly available

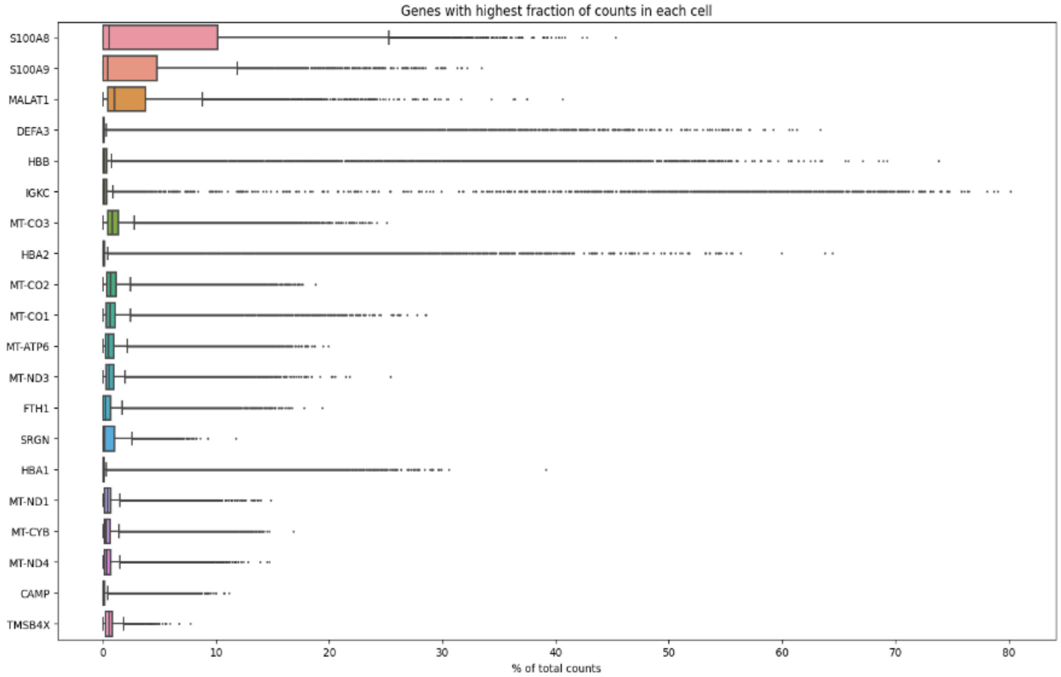

Figure 1: Genes with the highest fraction of counts in each cell in the dataset

3.2 Data preprocessing

In the study of scRNA sequencing of the human femoral head bone with the accession number GSE169396, the authors have performed the following preprocessing steps [15]:

- Four different individual's genes are concatenated into a single Ann data frame: This phase includes merging the gene expression data from various samples or individuals into a single data frame to support subsequent studies that need larger datasets.
- Genes that are expressed at incredibly low levels and are probably not physiologically significant are filtered out in this phase. These genes are discovered in fewer than three cells. The dataset is enhanced for more meaningful genes expressed in more cells by removing genes found in fewer than three cells.
- Principal component analysis (PCA) was used to reduce the data's dimensionality, preserving most of the data's variability while converting the high-dimensional gene expression data to a lower-dimensional representation. Data from scRNA sequencing may be reduced in dimension using a process called principal component analysis (PCA), which can denoise the data and highlight the key axes of variation.

These preprocessing steps are crucial for preparing the data for downstream analyses such as clustering, differential gene expression analysis, and cell type identification.

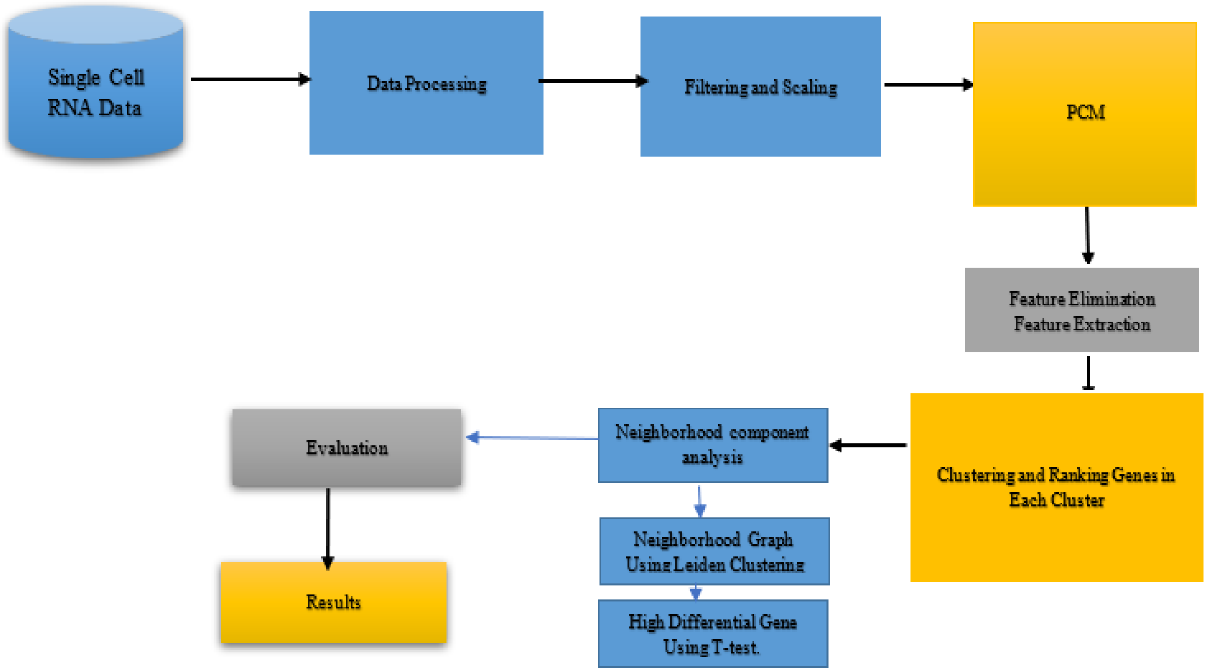
Single Cell RNA Data
Data Processing
Filtering and Scaling
PCM
Feature Elimination Feature Extraction
Clustering and Ranking Genes in Each Cluster
Neighborhood component analysis
Neighborhood Graph Using Leiden Clustering
High Differential Gene Using T-test
Evaluation
Results

Figure 2: Work structure

3.3. Filtering & Scaling

In the filtering and scaling step of the dataset with the accession number GSE169396, the following steps were taken [22]:

Removed the cells that have less than 200 genes, and genes that are detected in less than 3 cells: This step involves removing low-quality cells and genes that are unlikely to be biologically relevant. By filtering out cells with less than 200 genes and genes detected in less than 3 cells, the dataset is enriched for high-quality cells and informative genes. The shape of the dataset before filtering: The original dataset had a shape of (8916, 31826), meaning it contained 8,916 cells and 31,826 genes. The shape of the dataset after filtering: After filtering, the dataset had a shape of (7678, 15866), meaning it contained 7,678 high-quality cells and 15,866 informative genes. Normalized and logarithm zed the data: Normalization and logarithmic transformation are common preprocessing steps in scRNA sequencing data analysis. Normalization is used to remove technical variation between cells and genes, while logarithmic transformation is used to compress the dynamic range of the data and make it more amenable to downstream analyses.

3.4. Principal Component Analysis (PCA)

In several disciplines, including bioinformatics and machine learning, dimensionality reduction is frequently accomplished using the PCA approach. The principal component analysis (PCA) features extraction techniques that can reduce the dimensions of a high-dimensional dataset. PCA does this by locating a new collection of uncorrelated variables known as principal components. These primary components, which combine the original elements linearly to capture the most variation in the data, PCA is a potent technique that may simplify high-dimensional data while enhancing the efficiency and interpretability of subsequent analysis. There are several variations of PCA, such as kernel PCA, sparse PCA, and incremental PCA, each with its strengths and limitations depending on the specific application and data characteristics. Feature elimination is another dimensionality reduction technique that involves selecting a subset of the most informative features based on some predefined criteria, such as statistical significance, correlation, or relevance to a specific outcome. Feature elimination can be more straightforward than feature extraction, but it may lead to the loss of valuable information and may not always result in improved model performance [12].

3.5 PCA usage in RNA Seq

PCA is a widely used method for dimensionality reduction in scRNA sequencing (scRNA-seq) analysis. PCA can help identify and remove unwanted sources of variation and reveal the main axes of variation in the data, making it easier to analyze and interpret [24]. In scRNA-seq analysis, PCA is often performed on the log-transformed normalized counts or the z-scores of the genes. This is because these transformations can reduce the impact of highly variable genes on the PCA, which can help to uncover the underlying biological variation in the data.

Some common ways PCA is used in scRNA-seq analysis include:

1. Batch correction
2. Gene expression clustering
3. Sample clustering

Dimensionality reduction: By finding the primary axes of variation in the data and projecting the data onto a lower-dimensional space, PCA may be used to decrease the dimensionality of the data. The data may become more streamlined as a result, making it simpler to display and interpret.

3.6. Clustering and Ranking Genes in Each Cluster

- Performed Dimensionality Reduction with Neighborhood Components Analysis
- Clustered the Neighborhood graph using Leiden clustering.
- Computed a ranking for the highly differential genes in each cluster using a t-test.

In our study, we used Neighborhood Components Analysis (NCA) for dimensionality reduction, which is a supervised learning method that learns a linear transformation of the input data that maximizes the classification accuracy. NCA is effective in capturing the underlying structure of scRNA-seq data and reducing noise while preserving biologically meaningful variation in the data [25].

After dimensionality reduction, we used the Leiden algorithm for clustering, which is a graph-based clustering algorithm that has been shown to outperform other clustering methods in scRNA-seq data analysis [19]. The Leiden algorithm is based on optimizing a quality function that measures the modularity of the clustering, which reflects the degree to which nodes in the same cluster are more densely connected than nodes in different clusters. To identify highly differentially expressed genes in each cluster, a t-test was performed between the gene expression levels within the cluster and those in the remaining cells. The t-test is a widely used statistical method that compares the means of two groups, and it is commonly applied in scRNA-seq data analysis to identify genes that exhibit differential expression across cell types or experimental conditions. It is an effective method for detecting genes that are most likely to be associated with biological processes that are specific to a particular cluster of cells [26].

4. RESULTS

4.1. Number of PCs to consider.

The function sc.pl.pca_variance_ratio() displays the cumulative explained variance ratios for the principal components, which indicates the proportion of variation explained by each PC. This plot is commonly used to determine the appropriate number of principal components to retain for dimensionality reduction and clustering. By examining the plot, one can identify the number of principal components that account for most of the variance in the dataset, and therefore are most informative for downstream analysis.

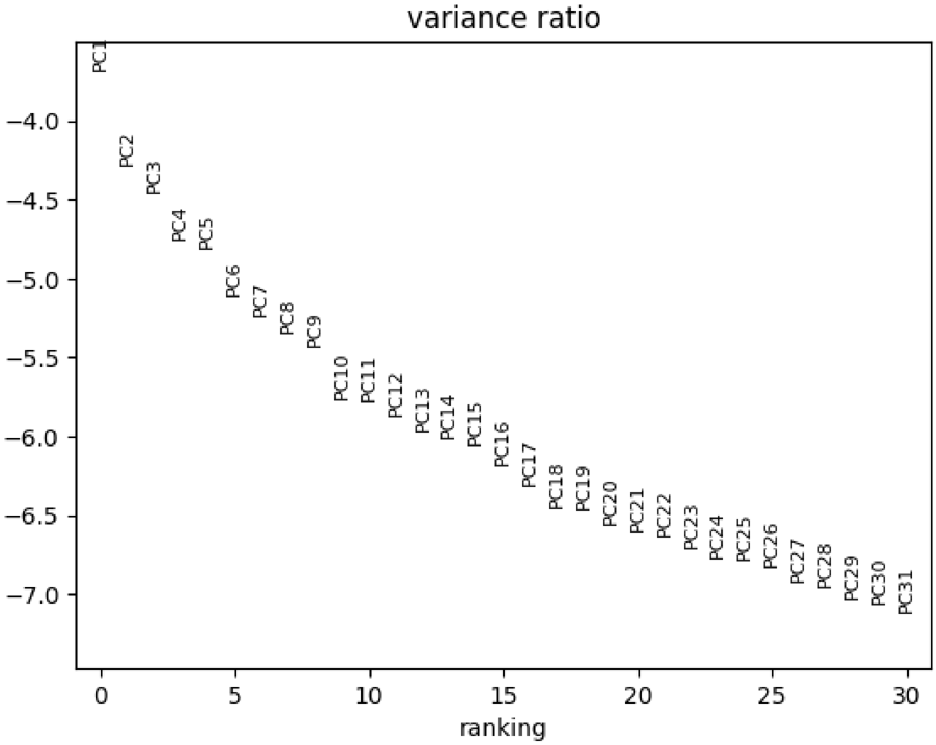

Figure 3: The above plot shows that there are 31 dominant PCs detected that capture most of the variation in the data.

4.2 Deciding the Number of Neighbours and PCs in the Neighborhood Graph

Silhouette Score

The silhouette score metric is used to measure the quality of clustering results in a scRNA sequencing (scRNA-seq) dataset. It ranges from -1 to 1, with a score of 1 indicating that a cell is well-separated from other clusters and associated with its cluster, and a score of -1 indicating that a cell is misclassified and more like cells in other clusters than to those in its cluster. A score close to 0 indicates that a cell is close to the decision boundary between two clusters.

Table 1: Neighbor PC Neighbor neighborhood Graph

n_neighbors	**n_pcs**	**silhouette score**	
		without PAGA	**with PAGA**
10	40	0.31	0.33
100	40	0.362	0.3711
100	31	0.373	**0.3831**
150	31	0.371	0.380
250	31	0.367	0.3746

Based on the above results, the neighborhood graph was computed using the number of neighbours as 100 and the number of PCs as 31. Table 1 compares different clustering analyses performed on a dataset, with and without the PAGA algorithm. The analysis was performed with varying values of two parameters: the number of nearest neighbours (n_neighbors) and the number of principal components (n_pcs) used in the analysis. The Silhouette score is a metric used to evaluate the quality of the clustering results, with higher values indicating better separation of clusters. The table shows the Silhouette score obtained for each combination of n_neighbors and n_pcs, with and without PAGA.

Comparing the results with and without PAGA, in most cases, the PAGA algorithm improved the clustering results, as evidenced by the higher Silhouette scores obtained with PAGA. This improvement is most evident when using 100 neighbours and 31 principal components, where the Silhouette score increased from 0.373 without PAGA to 0.3831 with PAGA.

The results also suggest that increasing the number of neighbors improves the clustering results, with higher Silhouette scores obtained for larger values of n_neighbors. However, the effect of increasing the number of principal components is less clear, as the Silhouette score does not consistently improve with increasing n_pcs. This table provides useful information for optimizing the clustering analysis of this dataset, by suggesting the appropriate values of n_neighbors and n_pcs and indicating the benefits of using the PAGA algorithm for improving clustering results.

4.3. UMAP of the clusters

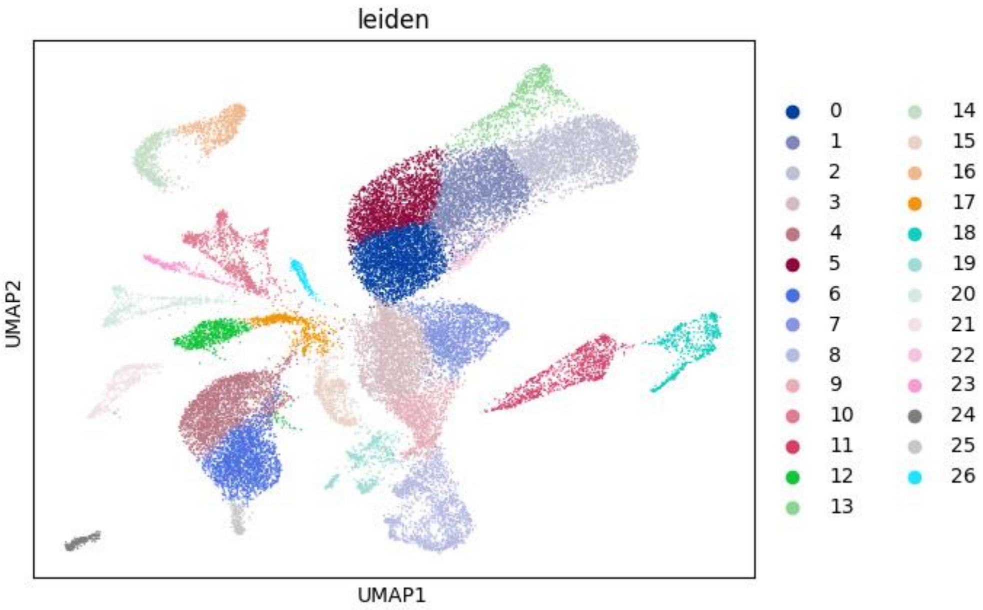

Figure 4: UMAP of the clusters found by Leiden clustering.

Based on their gene expression patterns, UMAP aids in classifying various cell types within the sample. On the UMAP figure, cells with similar expression profiles frequently aggregate in separate clusters that are close to each other. The cell types present in the sample can be deduced by labelling these clusters with recognized marker genes or with data from other reference sets.

In this case figure4, we have identified the top 25 highly ranked genes in each of the clusters and used one or two marker genes from each cluster to deduce the cell types clustered.

4.4. Ranking the Highly different Genes in each cluster – using a t-test

T-tests can be applied to the study of scRNA sequencing (scRNA-seq) data to find genes that are expressed differentially in two groups of cells. These categories might fit in for various cell kinds, medical disorders, or other specific interest categories. In more detail, a t-test in scRNA-seq analysis contrasts the average expression levels of a given gene in two different cell types (for example, the cell type in cluster 0 and the cell type in cluster 1). The objective is to establish whether there is a statistically significant difference in that gene's expression between the two groups.

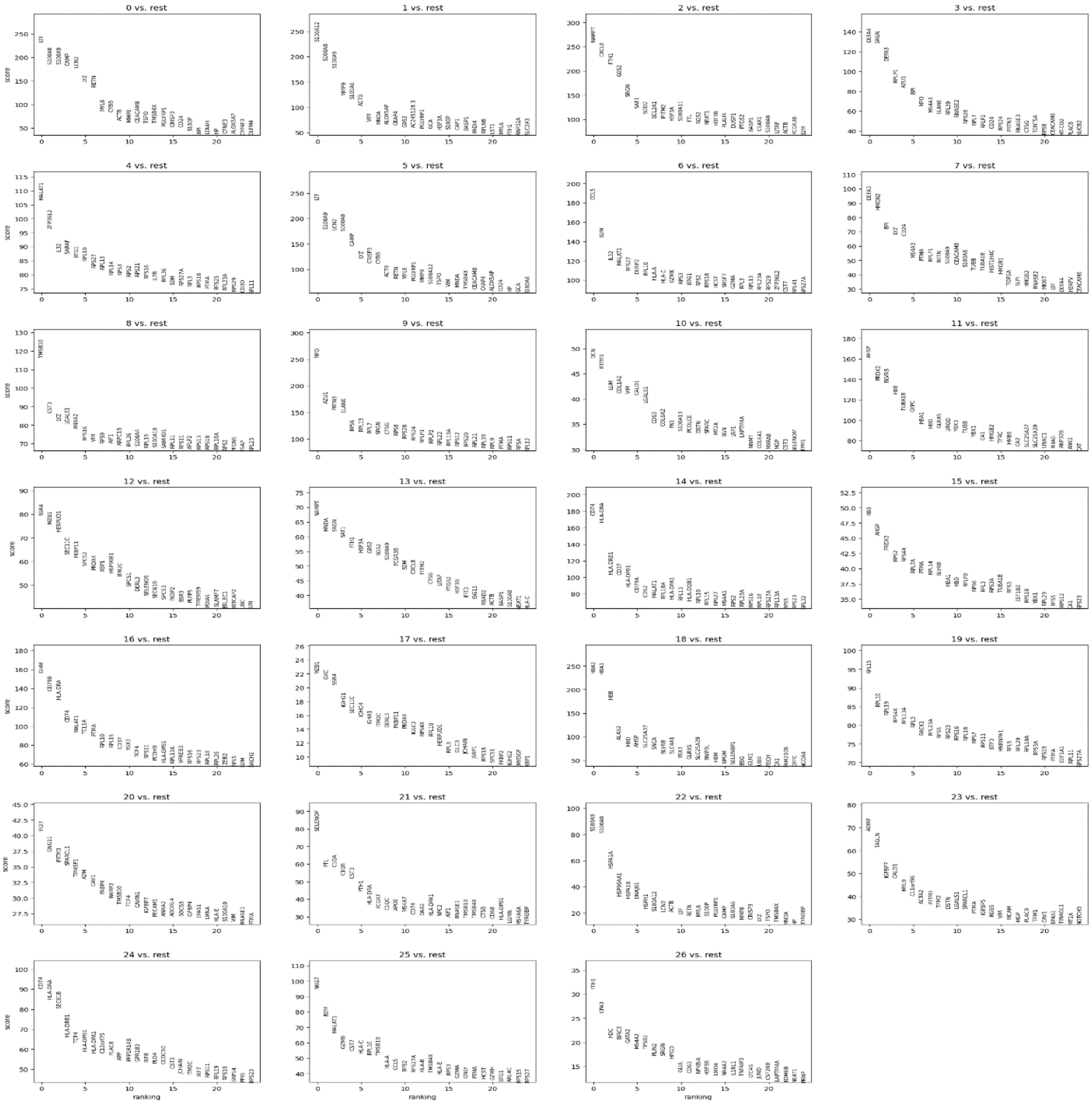

Figure 5: Top 25 differential genes from each of the 27 clusters

The Figure 5 graphs depict the top 25 genes expressed in each of the 27 clusters (refer Appendix section for the complete list) sorted by their respective ranking scores within the cluster.

4.5 Identifying marker genes from each cluster

Table 2: Marker genes identified from each cluster

Cluster	Marker Genes	Description
0	LTF, AHSP, ATP2B1	iron homeostasis and transport
1	S100A8 and S100A9	an inflammatory response (S100 family)
2	DEFA3, PRTN3, BPI	defense against bacterial infections

3	CAMP, LGALS1, ELANE, CD74	immune response
4	FTH1, RPS27, RPL13A, RPS4X	protein synthesis and ribosome biogenesis
5	RPL10, RPS16, RPL13, and RPL7A	ribosomal proteins
6	VIM, LGALS1, LYZ	cytoskeleton organization and cell motility
7	HLA-A	major histocompatibility complex (MHC) molecules (immune recognition)
8	ACTB and TPM2	cytoskeletal proteins
9	APOE and SLC4A1	lipid metabolism and transport

Table 2 provides a list of marker genes for 10 of the 27 clusters identified in the analyzed dataset. These markers are genes highly expressed in a specific cluster contrasted to other clusters and used to identify the biological processes or functions enriched in each cluster. Based on the marker genes listed in the table, we can infer the potential biological functions of each cluster. Cluster 0 is associated with iron homeostasis and transport [27], cluster 1 with the inflammatory response [28], cluster 2 with the defense against bacterial infections [29] [30], and Cluster 3 with the immune response [31] [32]. Cluster 4 is related to protein synthesis and ribosome biogenesis [33], cluster 5 to ribosomal proteins [34], and Cluster 6 to cytoskeleton organization and cell motility [35]. Cluster 7 is associated with (MHC (Major Histocompatibility Complex)) molecules involved in immune recognition [36], while cluster 8 is related to cytoskeletal proteins [37]. Finally, cluster 9 is associated with lipid metabolism and transport [38].

These marker genes can be further studied to investigate the biological processes and pathways that are involved in each cluster and to identify potential drug targets or therapeutic strategies.

We then map the identified biological functions from each of the clusters onto the clustered UMAP to visualize the proximity of all the processes.

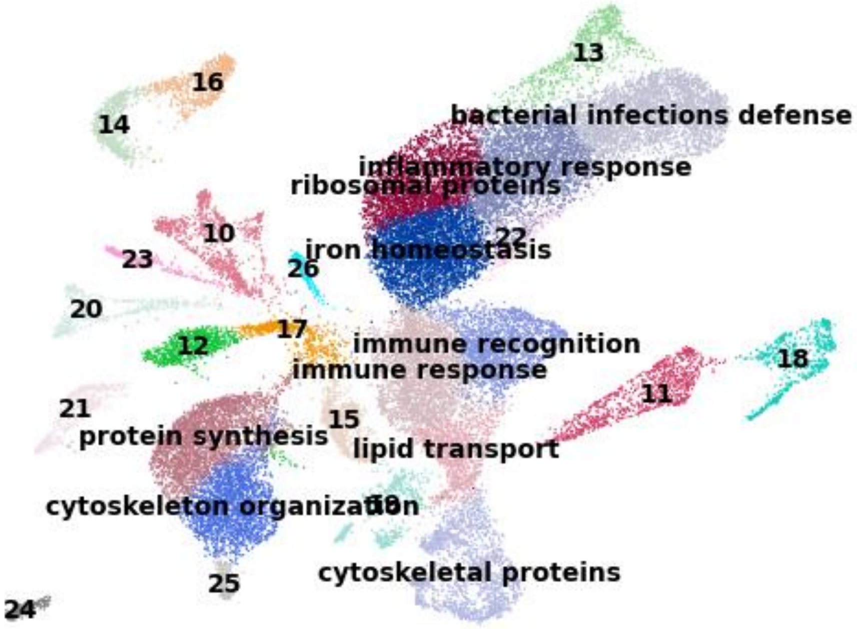

Figure 6: Biological processes mapped onto the UMAP

Observations from the above UMAP:

- Genes related to immune recognition and immune response are closely related based on their proximity. In a UMAP, cells from closely related lineages or those with similar functions tend to be positioned near each other on the plot.
- The clusters of bacterial infections defense and inflammatory response are nearby. Inflammation, the immune response, and bacterial infections are closely related. The immune system is triggered when the body experiences a bacterial infection to protect against invasive microorganisms. The immune response entails a planned series of actions meant to get rid of the infection and get tissue homeostasis back. An important aspect of the immune response is inflammation, which protects against bacterial infections.
- Clusters protein synthesis and cytoskeleton organization are close. The cytoskeleton's actin filaments play a role in several stages of protein synthesis. They are involved in the cytoplasmic localization and positioning of ribosomes and mRNA molecules. Additionally, actin filaments play a role in the dynamic mobility and trafficking of ribosomes and mRNA to cellular areas, including the leading edge of migrating cells or regions of active protein synthesis.

4.6 Visualizing marker genes in cluster distribution and Separation

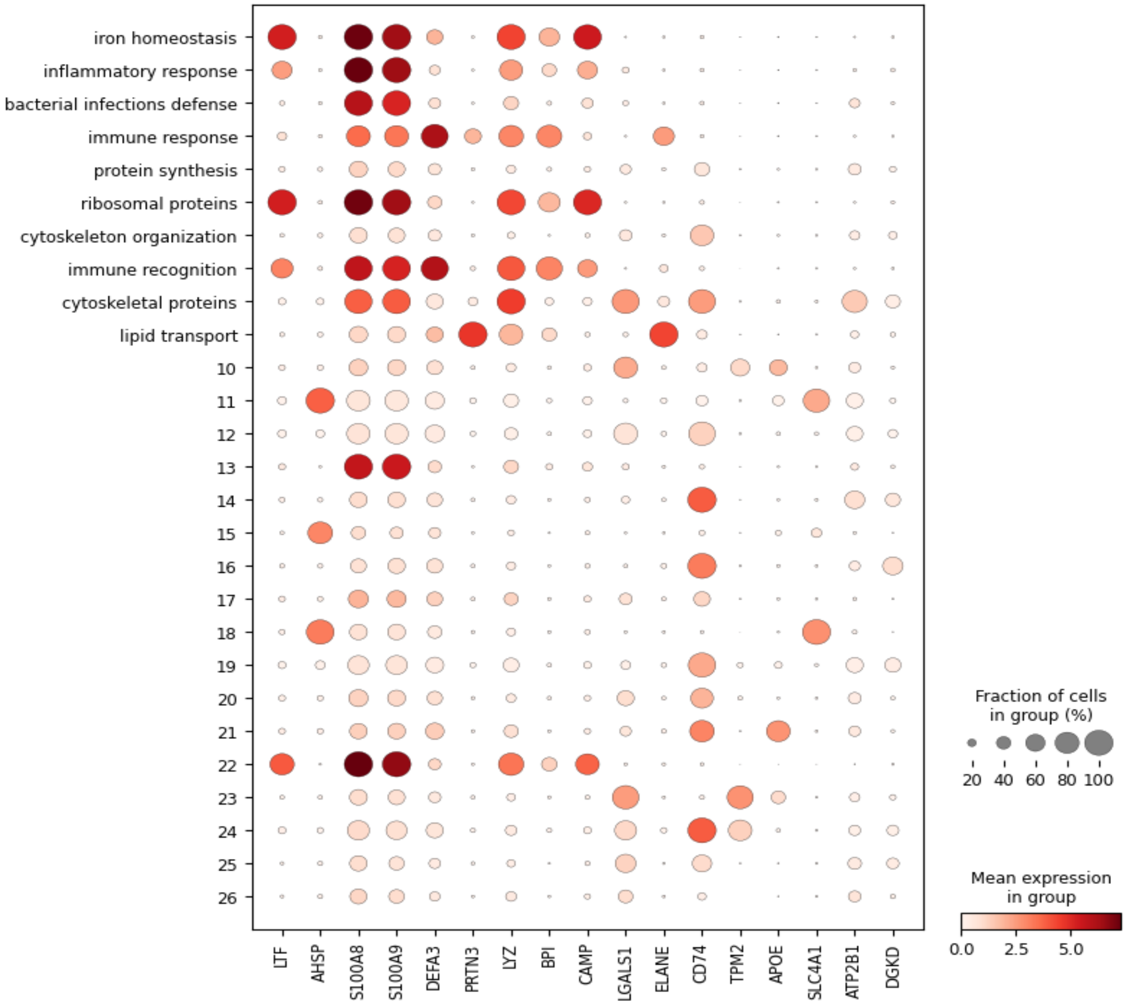

Figure 7: Dot plot of marker gene expression in each of the clusters

A dot plot is a visualization tool used in data analysis to explore and understand clustering patterns. It provides insights into the relationships and similarities between data points within different clusters.

From Figure 7, the following observations can be made:

- The genes PRTN3 and ELANE are expressed dominantly in the lipid transport cluster.
- Genes LYZ and LGALS1 have a high mean expression in the cytoskeletal proteins cluster.
- Cluster 18 has the genes AHSP and SLC4A1 highly expressed, and they are genes related to erythroid cells.
- Cluster twenty-three can be said to be biologically related to regulating osteoblasts and osteoclasts based on the high expression of genes LGALS1 and TPM2
- The gene CD74 is highly expressed in Cluster 24, and it can be taken as the marker gene for that cluster. The gene is related to regulating osteoblast and osteoclast activity whose deficiency is associated with an increase in bone mass.

- Since the genes S100A8 and S100A9 were high in a fraction of counts in each cell (figure 1), we can see that they are expressed highly in most of the clusters.

CONCLUSION

The PAGA algorithm improved the clustering results in most cases, as evidenced by the higher Silhouette scores obtained with PAGA. Therefore, future research could investigate the application of PAGA to other scRNA-seq datasets to further evaluate its effectiveness in improving clustering results. Secondly, increasing the number of neighbours improves the clustering results, while the effect of increasing the number of principal components is less clear. Future research could explore the optimal values for these parameters in different datasets to improve clustering accuracy. The identified marker genes for each cluster can be further studied to investigate the biological processes and pathways involved in each cluster and identify potential drug targets or therapeutic strategies. Future research could focus on validating the identified biological functions of each cluster and investigating the potential therapeutic implications of the marker genes. The UMAP analysis revealed interesting relationships between different biological processes and functions, such as the proximity of genes related to immune recognition and immune response, or the clustering of genes related to protein synthesis and cytoskeleton organization. Future research could explore these relationships further to gain insights into the underlying biological mechanisms and potential interactions between different processes. Future work on scRNA-seq provides information on gene expression at the single-cell level, but it does not capture spatial organization within tissues. Combining scRNA-seq data with spatial transcriptomics or imaging techniques could provide insights into how these cell types and functions are organized within tissues.

APPENDIX

Top twenty-five highly differential genes from each of the 27 clusters:

Cluster 0: LTF, S100A12, NAMPT, DEFA4, MALAT1, LTF, CCL5, DEFA3, TMSB10, MPO, DCN, AHSP, SSR4, NAMPT, CD74, HBB, IGHM, MZB1, RPL15, HBA2, IFI27, SELENOP, S100A9, ADIRF, CD74

Cluster 1: S100A8, S100A8, CXCL8, SRGN, ZFP36L2, S100A9, B2M, HMGN2, CST3, AZU1, IFITM3, PRDX2, MZB1, MNDA, HLA-DRA, AHSP, CD79B, IGKC, RPL10, HBA1, GNG11, FTL, S100A8, TAGLN, HLA-DRA

Cluster 2: S100A9, S100A9, FTH1, DEFA3, IL32, LCN2, IL32, BPI, LYZ, PRTN3, LUM, BLVRB, HERPUD1, SRGN, HLA-DRB1, PRDX2, HLA-DRA, SSR4, RPL19, HBB, IFITM3, C1QA, HSPA1A, IGFBP7, SEC61B

Cluster 3: CAMP, MMP9, G0S2, RPLP1, SARAF, S100A8, MALAT1, LYZ, LGALS1, ELANE, COL1A2, HBB, SEC11C, SAT1, CD37, RPS4X, CD74, IGHG1, RPS4X, ALAS2, SPARCL1, C1QB, HSP90AA1, CALD1, HLA-DRB1

Cluster 4: LCN2, S100A6, SRGN, AZU1, BTG1, CAMP, RPS27, CD24, ANXA2, RPS6, VIM, TUBA1B, FKBP11, FTH1, HLA-DPB1, RPS2, MALAT1, SEC11C, RPL13A, HBD, TM4SF1, CST3, HSPA1B, MYL9, TCF4

Cluster 5: LYZ, ACTB, SAT1, BPI, RPL10, LYZ, DUSP2, MS4A3, RPS16, RPL13, CALD1, GYPC, SPCS2, H3F3A, CD79A, PTMA, TCL1A, IGHG4, RPL3, AHSP, A2M, FTH1, DNAJB1, C11orf96, HLA-DPB1

Cluster 6: RETN, VIM, SOD2, MPO, RPS27, CRISP3, RPL10, PTMA, VIM, RPL7, LGALS1, HBA1, PRDX4, G0S2, MALAT1, RPL14, PTMA, IGHA1, RACK1, SLC25A37, CAV1, HLA-DRA, HSPH1, ACTA2, HLA-DPA1

Cluster 7: MYL6, MNDA, BCL2A1, MS4A3, RPL13, CYBB, HLA-A, RPLP1, AIF1, SRGN, CD63, HBD, HSP90B1, RGS2, CD52, RPL7A, RPL10, ITM2C, RPL23A, SNCA, FABP4, FCGRT, S100A12, IFITM3, C12orf75

Cluster 8: CYBB, ALOX5AP, IFITM2, ELANE, RPL14, ACTB, HLA-C, RETN, S100A4, CTSG, COL6A2, GLRX5, XBP1, S100A9, RPL18A, BLVRB, RPL15, DERL3, RPS5, BLVRB, RAMP2, C1QC, LCN2, TPM2, PLAC8

Cluster 9: ACTB, CKAP4, S100A11, RPL39, RPS3, RETN, GZMK, S100A9, ARPC1B, RPS8, FN1, UROD, ITM2C, FCGR3B, HLA-DPA1, HBA1, CD37, FKBP11, RPS23, SLC4A1, TMSB10, APOE, ACTB, DSTN, APP

Cluster 10: MMP8, G0S2, H3F3A, RNASE2, RPS2, MYL6, RPS3, CEACAM8, S100A10, RPS28, S100A13, YBX3, SPCS1, B2M, RPL13, HBD, YBX3, PRDX4, RPS16, YBX3, TCF4, MS4A7, LTF, LGALS1, PPP1R14B

Cluster 11: CEACAM8, AC245128.3, FTL, RPS28, RPS21, PGLYRP1, BTG1, S100A8, SAMHD1, RPS24, PCOLCE, TUBB, DERL3, CXCL8, HLA-DQB1, RPLP0, TCF4, RPL10, RPL18, GLRX5, CAVIN1, CD74, RETN, SPARCL1, GPR183

Cluster 12: TSPO, PGLYRP1, RGS2, RPL7, RPS16, MMP8, RPS2, TUBB, RPS9, RPLP1, DSTN, YBX1, SELENOS, IFITM2, RPL10, RPS3, RPS11, RPS4X, RPS7, SLC25A39, IGFBP7, DAB2, MYL6, PTMA, IRF8

Cluster 13: TMSB4X, GCA, NEAT1, RPLP2, IL7R, S100A12, RPS18, TUBA1B, APLP2, RPLP2, SPARC, HMGB2, SEC61B, CTSS, RPL15, RPL3, PCDH9, IGLC2, RPS11, BNIP3L, PECAM1, HLA-DPA1, S100P, IGFBP5, PLD4

Cluster 14: PGLYRP1, H3F3A, H3F3B, CD24, RPL36, TSPO, HCST, HIST1H4C, RPL26, RPL22, MT2A, CA1, SPCS3, LITAF, RPS27, TUBA1B, HLA-DPB1, HERPUD1, BTF3, HBM, ANXA2, NPC2, PGLYRP1, RGS5, CCDC50

Cluster 15: CRISP3, S100P, PLAUR, RPS24, B2M, VIM, SRSF7, HMGB1, PSAP, RPL13A, BGN, HMBS, FKBP2, PTGS2, MS4A1, RPS6, RPL13A, RPL3, HNRNPA1, BPGM, ADGRL4, AIF1, CAMP, VIM, CST3

Cluster 16: CD24, CAP1, DUSP1, PRTN3, RPS27A, MNDA, GZMA, TOP2A, RPS18, RPS12, LRP1, TFRC, SSR3, H3F3B, RPS2, RPS3A, VPREB3, IGLC3, RPL29, SELENBP1, SOCS3, RNASE1, S100A6, MCAM, JCHAIN

Cluster 17: S100P, BASP1, PTGS2, RNASE3, RPL3, TMSB4X, RPL3, SLPI, RPS2, RPS20, LAPTM4A, CA2, PLPP5, IFIT3, RPL23A, YBX1, RPS16, JCHAIN, RPL5, BSG, IGFBP4, TMSB10, MMP8, MGP, ITM2C

Cluster 18: BPI, PADI4, C5AR1, CTSG, RPS18, CEACAM8, RPL13, HMGB2, PABPC1, RPL21, NNMT, SLC25A37, TMEM59, ISG15, RPL18, RPS18, RPS23, RPS18, RPL10A, GUK1, EPAS1, TMSB4X, CRISP3, PLAC9, IRF7

Cluster 19: HP, RFLNB, BASP1, TENT5A, PTMA, CKAP4, RPL23A, RNASE2, RPL19, RPL39, COL6A1, SLC25A39, PDIA6, RSAD2, RPS16, EEF1B2, RPL18, JSRP1, RPS3A, UBB, LMNA, CTSB, LYZ, TPM1, RPS11

Cluster 20: ALOX5AP, LST1, S100A8, RPS8, RPS25, ALOX5AP, RPS29, MKI67, RPL18A, RPL9, MXRA8, RHAG, SLAMF7, ACTB, RPS27A, RPS12, RPL26, IGHG2, RPS19, FECH, HLA-E, CD68, TSPO, CAV1, RPL19

Cluster 21: LTA4H, MYL6, LITAF, CEACAM6, RPL23A, CD24, ZFP36L2, LTF, PFDN5, PTMA, MGP, UBAC1, ERLEC1, BASP1, RPL13A, CA1, ZEB2, SPCS1, PTMA, CA1, S100A10, HLA-DPB1, TMSB4X, EPAS1, RPS19

Cluster 22: CPNE3, FTH1, ACTB, PLAC8, RPS29, HP, CST7, H2AFV, RPL11, RPS11, CST3, ANP32B, KRTCAP2, S100A8, RPS5, RPL29, RPL5, MYDGF, EEF1A1, FAM210B, VIM, LGMN, MNDA, TINAGL1, SRP14

Cluster 23: CYP4F3, ANP32A, FCGR3B, MT-CO2, CD3D, GCA, RPL41, DEFA4, HLA-DRA, RPSA, SELENOM, ANK1, MYDGF, NEAT1, RPS23, RPS5, B2M, FKBP2, RPL11, GYPC, RNASE1, MS4A6A, HP, MT2A, PPIB

Cluster 24: OLFM4, FOS, B2M, NUCB2, RPL11, S100A6, RPS27A, ANP32B, RPS13, EEF1A1, TPM1, HBA2, UBC, HLA-C, RPL32, RPS23, BACH2, RPL7A, RPS27A, NCOA4, PTMA, TYROBP, TYROBP, NOTCH3, RPS23

Cluster 25: VIM, SLC2A3, FPR1, RPL13, RPL18, CPNE3, RPL14, CEACAM6, RPS11, RPL12, COL1A1, CAT, JUN, EVI2B, RPS11, RPL5, FAM129C, IGHG3, NACA, EPB42, HSPG2, HLA-DRB1, FCN1, CAVIN1, RPS2

Cluster 26: RPL39, FPR1, S100A9, PTMA, RPL19, ARPC5, CD3D, LCN2, COTL1, RPS29, CFH, KLF1, TMEM258, LCP1, RPL19, RPL8, RPL19, RPS5, RPS3, FKBP8, CRIP2, SLC40A1, PADI4, NR2F2, TMSB10

REFERENCES

[1] E. D. Amir *et al.*, "Vine enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukaemia," *Nat. Biotechnology.*, vol. 31, no. 6, pp. 545–552, Jun. 2013, Doi: 10.1038/nbt.2594.

[2] Y. Chen, D. McCarthy, M. Ritchie, M. Robinson, and G. Smyth, "edgeR: differential expression analysis of digital gene expression data User's Guide".

[3] V. Y. Kiselev *et al.*, "SC3: consensus clustering of scRNA-seq data," *Nat. Methods*, vol. 14, no. 5, pp. 483–486, May 2017, Doi: 10.1038/nmeth.4236.

[4] M. E. Ritchie *et al.*, "Limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, no. 7, pp. e47–e47, Apr. 2015, Doi: 10.1093/nar/gkv007.

[5] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR : a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010, Doi: 10.1093/bioinformatics/btp616.

[6] M. Plass *et al.*, "Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics," *Science*, vol. 360, no. 6391, p. eaaq1723, May 2018, Doi: 10.1126/science.aaq1723.

[7] F. Buettner *et al.*, "Computational analysis of cell-to-cell heterogeneity in scRNA-sequencing data reveals hidden subpopulations of cells," *Nat. Biotechnology.*, vol. 33, no. 2, pp. 155–160, Feb. 2015, Doi: 10.1038/nbt.3102.

[8] C. Cai, Y. Yue, and B. Yue, "ScRNA sequencing in skeletal muscle developmental biology," *Biomed. Pharmacotherapy.*, vol. 162, p. 114631, Jun. 2023, Doi: 10.1016/j.biopha.2023.114631.

[9] H. P. Gideon *et al.*, "Multimodal profiling of lung granulomas in macaques reveals cellular correlates of tuberculosis control," *Immunity*, vol. 55, no. 5, pp. 827-846.e10, May 2022, Doi: 10.1016/j.immuni.2022.04.004.

[10] T. M. Gierahn *et al.*, "Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput," *Nat. Methods*, vol. 14, no. 4, pp. 395–398, Apr. 2017, Doi: 10.1038/nmeth.4179.

[11] K. A. Zimmerman *et al.*, "ScRNA Sequencing Identifies Candidate Renal Resident Macrophage Gene Expression Signatures across Species," *J. Am. Soc. Nephrol.*, vol. 30, no. 5, pp. 767–781, May 2019, Doi: 10.1681/ASN.2018090931.

[12] A. Tsai, A. Petrov, R. A. Marshall, J. Morlach, S. Umemura, and J. D. Puglisi, "Heterogeneous pathways and timing of factor departure during translation initiation," *Nature*, vol. 487, no. 7407, pp. 390–393, Jul. 2012, Doi: 10.1038/nature11172.

[13] F. Tang *et al.*, "mRNA-Seq whole-transcriptome analysis of a single cell," *Nat. Methods*, vol. 6, no. 5, pp. 377–382, May 2009, Doi: 10.1038/nmeth.1315.

[14] B. Treutlen *et al.*, "Reconstructing lineage hierarchies of the distal lung epithelium using scRNA-seq," *Nature*, vol. 509, no. 7500, pp. 371–375, May 2014, Doi: 10.1038/nature13173.

[15] O. Bonny and M. Bochum, "Genetics of calcium homeostasis in humans: the continuum between monogenic diseases and continuous phenotypes," *Nephrol. Dial. Transplant.*, vol. 29, no. suppl four, pp. iv55–iv62, Sep. 2014, Doi: 10.1093/ndt/gfu195.

[16] C. Mathias, "VIRUS-INDUCED CHANGES IN NUCLEAR PROTEINS AND MEMBRANES IN NICOTIANA BENTHAMIANA CELLS," 2020, Doi: 10.13023/ETD.2020.443.

[17] R. Satya, J. A. Farrell, D. Gennet, A. F. Scheer, and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nat. Biotechnology.*, vol. 33, no. 5, pp. 495–502, May 2015, Doi: 10.1038/nbt.3192.

[18] F. Tang *et al.*, “mRNA-Seq whole-transcriptome analysis of a single cell,” *Nat. Methods*, vol. 6, no. 5, pp. 377–382, May 2009, Doi: 10.1038/nmeth.1315.

[19] B. Treutlen *et al.*, “Reconstructing lineage hierarchies of the distal lung epithelium using scRNA-seq,” *Nature*, vol. 509, no. 7500, pp. 371–375, May 2014, Doi: 10.1038/nature13173.

[20] N. Habib *et al.*, “Div.-Seq: A single nucleus RNA-Seq method reveals dynamics of rare adult newborn neurons in the CNS,” Neuroscience, preprint, Mar. 2016. Doi: 10.1101/045989.

[21] F. A. Wolf, P. Angered, and F. J. Theis, “SCANPY: large-scale single-cell gene expression data analysis,” *Genome Biol.*, vol. 19, no. 1, p. 15, Dec. 2018, Doi: 10.1186/s13059-017-1382-0.

[22] R. Dries, J. Chen, N. Del Rossi, M. M. Khan, A. Sister, and G.-C. Yuan, “Advances in spatial transcriptomic data analysis,” *Genome Res.*, vol. 31, no. 10, pp. 1706–1718, Oct. 2021, Doi: 10.1101/gr.275224.121.

[23] L. Hogherd, F. Buettner, and F. J. Theis, “Diffusion maps for high-dimensional single-cell analysis of differentiation data,” *Bioinformatics*, vol. 31, no. 18, pp. 2989–2998, Sep. 2015, Doi: 10.1093/bioinformatics/btv325.

[24] L. Hogherd, A. T. L. Lunn, M. D. Morgan, and J. C. Marion, “Batch effects in scRNA-sequencing data are corrected by matching mutual nearest neighbours,” *Nat. Biotechnology.*, vol. 36, no. 5, pp. 421–427, May 2018, Doi: 10.1038/nbt.4091.

[25] V. A. Tragi, L. Waltman, and N. J. Van Eck, “From Louvain to Leiden: guaranteeing well-connected communities,” *Sci. Rep.*, vol. 9, no. 1, p. 5233, Mar. 2019, Doi: 10.1038/s41598-019-41695-z.

[26] T. Stuart *et al.*, “Comprehensive Integration of Single-Cell Data,” *Cell*, vol. 177, no. 7, pp. 1888-1902.e21, Jun. 2019, Doi: 10.1016/j.cell.2019.05.031.

[27] Nasimuzzaman M, Khandros E, Wang X, Kong Y, Zhao H, Weiss D, Rivella S, Weiss MJ, Persons DA. Analysis of alpha hemoglobin stabilizing protein overexpression in murine β-thalassemia. Am J Hematol. 2010 Oct;85(10):820-2. doi: 10.1002/ajh.21829. Erratum in: Am J Hematol. 2011 Sep;86(9):824. PMID: 20815047; PMCID: PMC3632304.

[28] Gebhardt C, Németh J, Angel P, Hess J. S100A8 and S100A9 in inflammation and cancer. Biochem Pharmacol. 2006 Nov 30;72(11):1622-31. doi: 10.1016/j.bcp.2006.05.017. Epub 2006 Jul 17. PMID: 16846592.

[29] Cabak A, Hovold G, Petersson AC, Ramstedt M, Påhlman LI. The activity of airway antimicrobial peptides against cystic fibrosis pathogens. Pathog Dis. 2020 Oct 8;78(7):ftaa048. doi: 10.1093/femspd/ftaa048. PMID: 32857857.

[30] Weiner DJ, Bucki R, Janmey PA. The antimicrobial activity of the cathelicidin LL37 is inhibited by F-actin bundles and restored by gelsolin. Am J Respir Cell Mol Biol. 2003 Jun;28(6):738-45. doi 10.1165/rcmb.2002-0191OC. Epub 2002 Dec 30. PMID: 12600826.

[31] Kościuczuk EM, Lisowski P, Jarczak J, Strzałkowska N, Jóźwik A, Horbańczuk J, Krzyżewski J, Zwierzchowski L, Bagnicka E. Cathelicidins: a family of antimicrobial peptides. A review. Mol Biol Rep. 2012 Dec;39(12):10957-70. doi: 10.1007/s11033-012-1997-x. Epub 2012 Oct 14. PMID: 23065264; PMCID: PMC3487008.

[32] Ka-Yue Chow L, Lai-Shun Chung D, Tao L, Chan KF, Tung SY, Cheong Ngan RK, Ng WT, Wing-Mui Lee A, Yau CC, Lai-Wan Kwong D, Ho-Fun Lee V, Lam KO, Liu J, Chen H, Dai W, Lung ML. Epigenomic landscape study reveals molecular subtypes and EBV-associated regulatory epigenome reprogramming in nasopharyngeal carcinoma. EBioMedicine. 2022 Dec;86:104357. doi: 10.1016/j.ebiom.2022.104357. Epub 2022 Nov 11. PMID: 36371985; PMCID: PMC9663866.

[33] Galy B, Ferring-Appel D, Becker C, Gretz N, Grone HJ, Schümann K, Hentze MW. Iron regulatory proteins control a mucosal block to intestinal iron absorption. Cell Rep. 2013 Mar 28;3(3):844-57. doi: 10.1016/j.celrep.2013.02.026. Epub 2013 Mar 21. PMID: 23523353.

[34] Studer D, Lischer S, Jochum W, Ehrbar M, Zenobi-Wong M, Maniura-Weber K. Ribosomal protein l13a as a reference gene for human bone marrow-derived mesenchymal stromal cells during expansion, adipose-, chondro-, and osteogenesis. Tissue Eng Part C Methods. 2012 Oct;18(10):761-71. doi 10.1089/ten.TEC.2012.0081. Epub 2012 Jun 7. PMID: 22533734; PMCID: PMC3460615.

[35] Viguier M, Avedissian T, Delacour D, Poirier F, Deshayes F. Galectins in epithelial functions. Tissue Barriers. 2014 May 6;2:e29103. doi 10.4161/tisb.29103. PMID: 25097826; PMCID: PMC4117684.

[36] Cai L, Michelakos T, Yamada T, Fan S, Wang X, Schwab JH, Ferrone CR, Ferrone S. Defective HLA class I antigen processing machinery in cancer. Cancer Immunol Immunother. 2018 Jun;67(6):999-1009. doi 10.1007/s00262-018-2131-2. Epub 2018 Feb 27. PMID: 29487978; PMCID: PMC8697037.

[37] Schevzov G, Whittaker SP, Fath T, Lin JJ, Gunning PW. Tropomyosin isoforms and reagents. Bio architecture. 2011 Jul;1(4):135-164. doi: 10.4161/bioa.1.4.17897. Epub 2011 Jul 1. PMID: 22069507; PMCID: PMC3210517.

[38] Marais AD. Apolipoprotein E in lipoprotein metabolism, health, and cardiovascular disease. Pathology. 2019 Feb;51(2):165-176. doi 10.1016/j.pathol.2018.11.002. Epub 2018 Dec 28. PMID: 30598326.