SINGLE CELL RNA SEQUENCING TO DETECT DIFFERENTIALLY EXPRESSED GENES AND IDENTIFY SENESCENCE ASSOCIATED GENES

A thesis presented for the degree of

MSc. Data Science and Analytics

Author

Bharvi Dhall

Supervisor

Dr. Katarina Domijan

Co-Supervisor

Dr. Catherine Hurley

Department of Mathematics and Statistics

Maynooth University

August 2019

This thesis is dedicated to my parents

STATEMENT OF ORIGINALITY

I have read and understood the Departmental policy on plagiarism and I certify that work demonstrated in this thesis titled *Single Cell RNA sequencing to detect differentially expressed genes and identify senescence genes* is my own and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. I confirm that:

- Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.
- Sources of all figures or tables has been provided in the document which are not my work.

Name:. . . . .

Student ID:. . . . .

Sign:. . . . .

Date:..

ACKNOWLEDGEMENT

All the way through, working on this dissertation I have received countless support and assistance. I would first like to recognize and thank my supervisor, **Dr. Katarina Domijan**, whose expertise and proficiency helped me to carry out the analysis for this research project in particular. I was not very confident initially as the topic and the background was unfamiliar for me. She guided me and encouraged me to deliver the work done in the thesis. She guided me by providing valuable feedback on my writing skills and advised me to use bookdown for writing and compiling this thesis which helped me develop a new skill to write technical documents.

I would like to give special thanks to my co-supervisor, **Dr. Catherine Hurley** without whom this thesis wouldn't have been completed. She helped me throughout, helping me with my R skills and trying her best to help me and Simran with this thesis. She was very responsive and resolved every query that I had quickly. She also organized the RStudio set up for me as I was having space issues with my laptop and provided me with a part of R-code to integrate multiple datasets. She suggested new ways to target the problems and helped me become better in my skills than I was on Day 1 of the thesis. I am thankful to both of my supervisors for actively participating in weekly project meetings and explaining me my doubts and suggesting me ways to improve the research work.

I would also like to acknowledge **Dr. Mark W Robinson** for providing me with this research topic and clearing all the domain-specific doubts. He helped in the labeling of clusters without which the results of the thesis would not have been achieved.

I am grateful to **Dr. Caroline Brophy** for guiding and inspiring me to work on my writing and time management skills. In spite of a busy schedule, she always took some time whenever I was seeking help.

ACKNOWLEDGEMENT

I am thankful to **Dr. Rob Kitchin** for helping me develop my intellectual, philosophical and writing skills. I would also like to thank **Dr. Peter Mooney** and **Dr. Martin Charlton** for providing good exposure to different applications of Data Science.

A special thanks to **Janice Love and Anthony Waldron** for installing R packages and uploading the big datasets on RStudio Server and helping me on short notices.

I would like to express gratitude to Maynooth University for making Library, labs and other facilities available and for conducting sessions for Writing Skills.

At last, I would like to thank my family and friends for their encouragement, patience, and motivation.

Contents

ABSTRACT

Cells have specialised functions which are responsible for healthy functioning of different organs. Normal cells in our body grow, repair, replicate and die whereas some cells in our body cease to divide and often lead to loss of function. This condition of the cell when it ceases to divide is known as cellular senescence and some specific genes are responsible for such cell conditions. In order to study the human body at the microscopic level, Single Cell RNA Sequencing is used which tries to capture the activity of thousands of genes in a single cell.

This thesis provides a detailed description of methods and discusses the implementation of a Single Cell RNA Sequencing Pipeline to analyse such datasets. It provides a good knowledge and understanding to get familiar with the workflow of such pipelines which can be used for analysing any single cell genomics dataset. The pipeline was used to analyse datasets namely Liver and Bone Marrow to find differentially expressed genes and detect the presence of senescent genes. The results conclude that presence and location of senescent genes can be detected using such analysis.

NOMENCLATURE

Table 1: *Table for Acronyms*

S.No	Abbreviations	Expanded_Form
1	UMI	Unique Molecular Identifier
2	QC	Quality Control
3	HVG	Highly Variable Genes
4	PCA	Principal Component Analysis
5	PCs	Principal Components
6	SNN	Shared Nearest Neighbors
7	KNN	k-Nearest Neighbors
8	t-SNE	t-Distributed Stochastic Neighbor Embedding
9	RNA	Ribonucleic Acid
10	DNA	Deoxyribonucleic Acid
11	RNA-seq	RNA-Sequencing
12	scRNA-seq	single cell RNA Sequencing
13	DE	Differential Expression
14	UMAP	Uniform Manifold Approximation and Projection
15	v2	version2
16	v3	version3

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Purpose and Motivation

A cell is the basic structural unit of life and the human body is made up of millions of cells which are responsible for carrying out various functions and life processes. With time, these processes in the human body begin to depreciate and become inefficient, thereby causing sickness and impacting the health of human beings. In early centuries, the health care system was more of a sick-cure system where a person was cured once they got any illness. With the advancement in healthcare and medicine, this approach has now been shifted to detect the cause of underlying disease instead to make our health care system disease proof. Identification of senescence-associated transcriptional profiles is one such example.

Cellular senescence is a phenomenon in which some cells cease to divide due to an irreversible cell cycle arrest[1]. This phenomenon was first described in paper titled *The serial cultivation of human diploid cell strains* (Hayflick and Moorhead,1961). This project aims to identify cell populations which undergo cellular senescence and to reduce the number of genes to find differentially expressed genes (Refer section3.7) and thus investigating the genes that are associated with cellular senescence. The scope of the project also extends to find out the genes co-expressed with senescent genes with the help of gene-coexpression analysis.

The two Ribonucleic Acid (RNA) sequencing (seq) datasets namely liver and bone marrow were analysed to achieve the goals of this project. RNA-seq analysis was performed using the

[1] Please refer https://biologydictionary.net/cellular-senescence/ for more detailed explanation

RNA-seq pipeline using package Seurat version 3.0 to replicate the results from paper *Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations*(MacParland et al., 2018). The same pipeline was followed to analyse the bone marrow dataset and to replicate results obtained from paper *Single Cell Analyses of Human Bone Marrow* (Oetjen et al.,2018).

The paper titled *Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations* (MacParland et al.,2018) follows the RNA-seq pipeline to identify the 20 distinct populations of cells in the human liver and labels them. We analysed the liver dataset using RNA-seq pipeline to replicate their results and then differential expression analysis was performed to detect the differentially expressed genes in each cluster (every cluster represents a distinct cell population). The differentially expressed genes were used to identify senescent cells [2].

[2] The cells associated with cellular senescence are also known as senescent cells

1.2 Outline of the Thesis

This thesis has been structured into five chapters and each chapter has multiple subsections. Two appendices have been included in the end to demonstrate the R-code used to support the analysis.

Chapter1: Introduction . Chapter 1, provides an introduction to the research topic and provides an overview to different chapters in the thesis

Chapter2: Background . Chapter 2, explains the biological background associated with the research topic. It also discusses the datasets, the packages used to perform the analysis and the challenges associated with it. This chapter also highlights the similar work performed in literature.

Chapter3: Methodology . The functions and packages used in RNA-Seq workflow are different from typical functions used in statistics. Thus, this chapter explains the detailed methodology implemented in the analysis of datasets. The functions used by Seurat package were studied and the underlying methodology used by these functions has been discussed briefly in this chapter.

Chapter4: Results . Chapter 3, demonstrates the results obtained after analysis of liver and bonemarrow dataset.

Chapter5: Conclusion and Future Scope . Chapter 5, concludes the results obtained from this project and discusses the future score of this project.

Chapter 2

BIOLOGICAL BACKGROUND

This chapter discusses some basic terminologies and highlights some life processes which are required to develop a good understanding of biological research areas and datasets. It also gives an introduction to the datasets and packages used for analysis of biological datasets.

The cell is the basic structural unit of all living organisms and human body consists of trillions of cells, and these cells have specialised functions which are responsible for healthy functioning of different organs. The nucleus of the cell contains sub-cellular structures known as chromosomes which are responsible for transferring genetic material. Each chromosome has tightly packed Deoxyribonucleic Acid (DNA).

Information stored in DNA is used to synthesize a protein which is responsible for the functioning of a cell. Gene is a part of DNA on a chromosome which is transcribed to RNA, which is then translated to amino acids which gets folded to synthesise proteins. This process of synthesizing proteins from DNA is known as Gene expression (Crick,1970).

Figure2.1represents Gene expression. It gives a pictorial representation of various stages in conversion of DNA to protein[1].

[1] This image has been downloaded from http://bio.academany.org/2017/labs/BioRiiDL_2017/sreejith/images/assignments/dogma.png

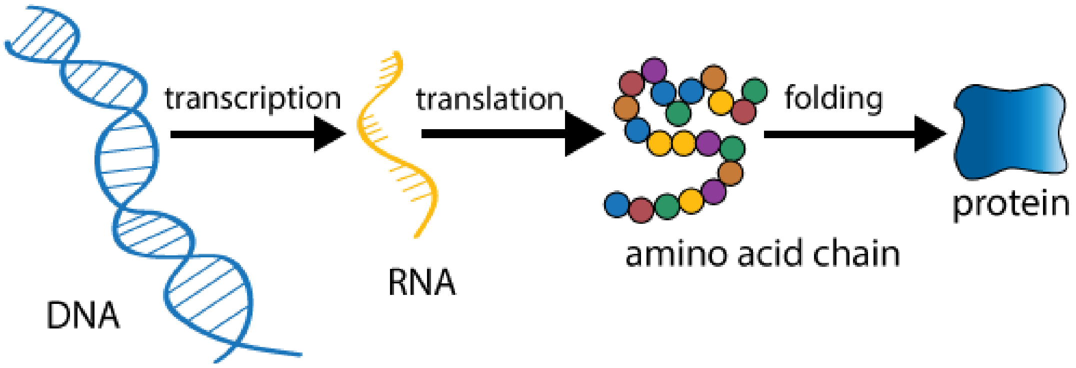


Figure 2.1: *Gene expression*

The knowledge of gene expression helps us to find and target various problems in the human body as the mutation of cells cause diseases and we can study the difference in normal and mutated cells by their gene expression.

Cells can accomplish diverse functions with identical DNA in part by controlling the quantity of RNA which is produced from each gene. For example, although cells throughout the human body have essentially the same DNA, of all genes only some specific genes are turned on which leads to different appearances of the cell (Schug et al.,2005), thereby allowing tissues to perform diverse functions.

The activity of thousands of genes is measured at once to create a global picture of functioning in cells. This is known as expression profiling (Metsis et al.,2004). This is done to find out which cells are actively dividing or how these cells are reacting to a treatment. The data required for carrying out such analysis is generated using various Transcriptomics technologies (Lowe et al.,2017) like DNA-Microarrays and RNA-Sequencing. In this project, we will be using RNA-Sequencing data for analysing the genes associated with cellular senescence.

Cellular senescence is the phenomenon associated with ageing of cells due to which cells are no longer able to divide and this can occur because of damaged DNA (Hayflick and Moorhead, 1961). Normal cells in our body grow, repair, replicate and die whereas senescent cells cease to divide and often lead to loss of function and are found in organs with chronic diseases. Study of senescent cells is an undiscovered and fascinating topic in the field of biotechnology and immunology as the removal of senescent cells from the human body may slow down or reverse ageing (Keizer,2017). Presence of senescent cells, when tested on mice, lead to physical dysfunction whereas the removal of these cells extended their lifespan and restored health (Pan et al.,2017).

2.1 Dataset Description and availability

This section illustrates the layout of biological datasets and provides an overview of the datasets used for analysis in this project. It highlights the challenges involved with using these datasets 2.1.1and also provides an introduction to sparse matrices2.1.2.

The datasets used in this project are generated using single-cell RNA sequencing (scRNA-seq). scRNA is a detailed study of the gene expression as it tries to capture the activity of thousands of genes in a single cell. It has provided a lot of useful insights into the field of cancer genomics (Hwang, Lee, and Bang,2018) and is a new domain with ongoing research. scRNA-seq is used to study the functions of a cell and its heterogeneity (Papalexi and Satija,2018). It treats each cell as an individual sample. A gene expression matrix obtained after scRNA-seq has genes as rows (features) and cells as columns (cases). Each cell in the matrix gives counts per sample for a specific gene.

Biological matrices are represented differently than the matrices in statistics. **These matrices are popularly known as count matrices or gene-expression matrices. Here the rows represent features (p) and columns represent cases (n). Thus, the gene expression matrix is a pXn dimentional matrix. Each cell in the matrix represents the number of reads mapping to each gene for each sample.** These matrices usually have a lot of zeros, which can occur if data is not captured properly or a particular gene is not expressed in that cell.

Table 2.1 represents the layout of the Gene Expression matrix. This is a fake data matrix for illustration purposes.

Liver Dataset:
The dataset contains gene expression profiling of 8444 cells obtained from liver grafts of five healthy neurologically deceased donors (NDD). The data acquired is a 20007 X 8444 matrix with 20007 genes and 8444 cells [2].

[2] The liver dataset used for analysis can be downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115469ty

Table 2.1: *Sample Gene expression matrix*

Genes	Sample.1	Sample.2
Gene 1	1	0
Gene 2	0	0
Gene 3	0	1
Gene 4	1	0

Bone Marrow Dataset:

The bone marrow dataset is obtained from twenty volunteers, that is, 10 males and 10 females with ages ranging from 24 to 84 years old and median age of 57 years. The scRNA-seq was performed using 10X Genomics Single Cell 3 Solution, version 2. Files from multiple donors were merged to obtain the gene-expression matrix with row counts and the dimensions of the matrix thus obtained were 33694 X 90653 and a Seurat object was created2.3.

The dataset has 33694 genes and 90653 cells which will be pre-processed and filtered for analysis[3].

2.1.1 Challenges with Datasets

This section outlines the challenges associated with working with large datasets.

The scRNA data is recorded at a single-cell resolution, thus the size of the count matrices may vary from one dataset to another. The matrices used to store biological information are large and working with these datasets is computationally challenging. Therefore it is important to have a good computational power to carry out analysis using these datasets. To make it easier to work with these datasets the information is stored using sparse matrices (Koenker, Ng, et al., 2003). Section2.1.2provides a background to sparse matrices.

Since both of the datasets used in the analysis were large, the calculations were time-consuming. These datasets have a number of features (p) much higher than the number of cases (n), (p>>n), thus it is important to reduce the number of features in these datasets (Refer3.4).

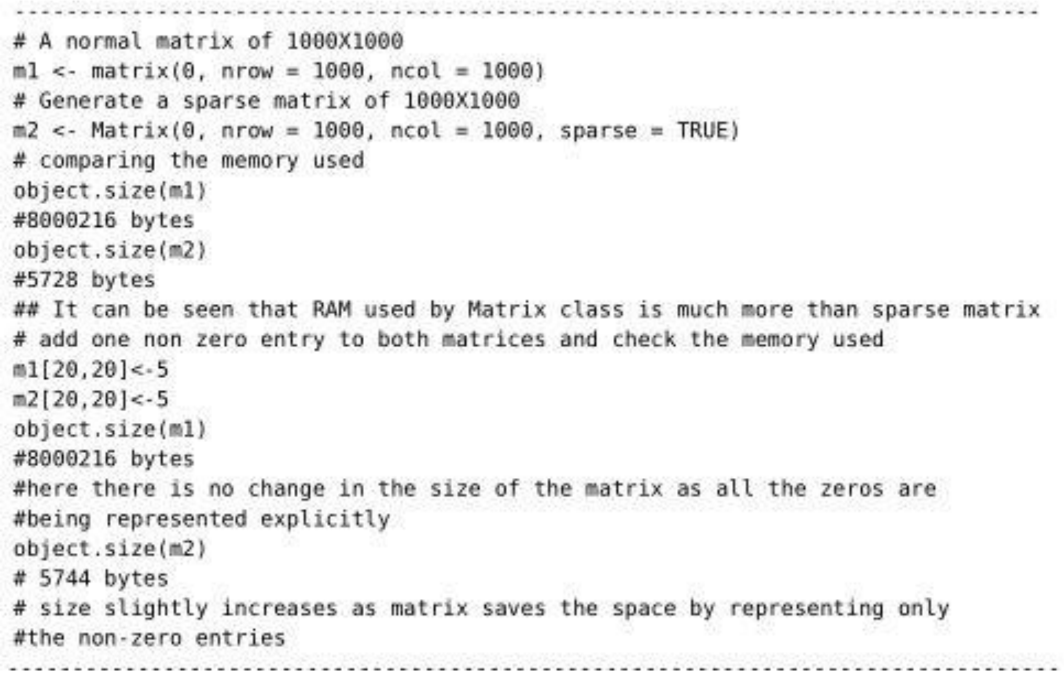
Biological datasets have different terminologies and the results of the functions are difficult to interpret without prior background knowledge.

[3] The bone marrow dataset used for analysis can be downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120221

2.1.2 Sparse Matrices

This section provides an introduction to sparse matrices and illustrates its advantages. The matrices obtained from biological data are usually big and have many zeros. The sparse matrices[4] are used when most of the elements in the dataset are 0. Sparse Matrix saves a lot of space in the memory by representing only the non-zero entries. It is computationally efficient compared to the dense matrices and makes calculations faster. A matrix is called dense when most of the elements in the matrix are non-zero. The below example has been created using R (Team et al.,2013) to understand the advantages of sparse matrices compared to dense matrices.

Figure2.2demonstrates the implementation of sparse matrices in R and it also compares the memory used by a sparse matrix and a normal matrix.



```
# A normal matrix of 1000X1000
m1 <- matrix(0, nrow = 1000, ncol = 1000)
# Generate a sparse matrix of 1000X1000
m2 <- Matrix(0, nrow = 1000, ncol = 1000, sparse = TRUE)
# comparing the memory used
object.size(m1)
#8000216 bytes
object.size(m2)
#5728 bytes
## It can be seen that RAM used by Matrix class is much more than sparse matrix
# add one non zero entry to both matrices and check the memory used
m1[20,20]<-5
m2[20,20]<-5
object.size(m1)
#8000216 bytes
#here there is no change in the size of the matrix as all the zeros are
#being represented explicitly
object.size(m2)
# 5744 bytes
# size slightly increases as matrix saves the space by representing only
#the non-zero entries
```

Figure 2.2: *Comparison of memory used in a Sparse and Normal Matrix*

Addition of a single non-zero observation to the dataset reveals that the space in the memory occupied by the sparse matrix will increase while the space occupied by the normal matrix will remain unchanged. Thus, both the datasets in this project have been stored as a sparse matrix for the ease of calculations.

[4] For documentation please refer https://www.rdocumentation.org/packages/Matrix/versions/1.2-17/ topics/sparseMatrix

2.2 RNA-Sequencing Workflow

This section explains the steps involved in the analysis of RNA-seq datasets. The implementation of this workflow has been discussed in Chapter3and Chapter4.

A typical RNA-Seq analysis consists of five steps. There are a variety of softwares and environments that can be used for the analysis of RNA-Seq data, however, the steps taken in an analysis workflow are typically analogous and will follow the same procedure as discussed in this project. The stages in the workflow are pre-processing of raw counts, normalization, Dimensionality Reduction, Clustering and Labelling, Differential Expression Analysis and Pathway Enrichment Analysis. Figure [5] 2.3describes the workflow commonly used for analysing RNA-Seq Data (MacParland et al.,2018).

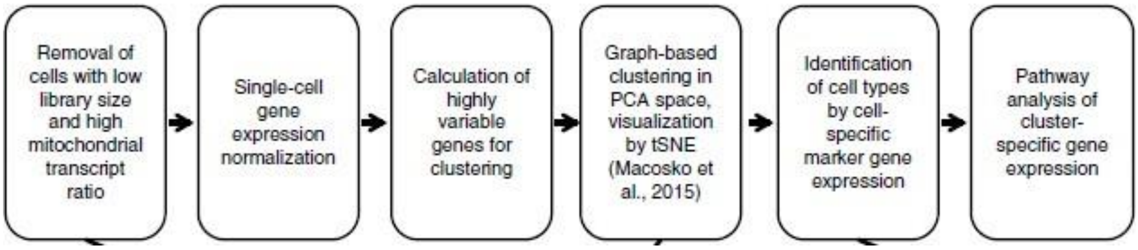

Removal of cells with low library size and high mitochondrial transcript ratio
Single-cell gene expression normalization
Calculation of highly variable genes for clustering
Graph-based clustering in PCA space, visualization by tSNE (Macosko et al., 2015)
Identification of cell types by cell-specific marker gene expression
Pathway analysis of cluster-specific gene expression

Figure 2.3: *RNA-Seq Workflow*

The raw counts in the dataset are filtered to remove low quality and dying cells. The resultant matrix is then normalized to make data comparable and prevent false biological conclusions(Steinhoff and Vingron,2006). Once, normalized matrix is achieved then analysis is performed. Dimensionality reduction techniques are used to reduce the number of genes (features) and to find out important features. This is done with the help of various techniques like Principal Component Analysis (Pearson,1901) and t-SNE (Maaten and Hinton,2008). Then clustering is done on the results of PCA and the resultant clusters are labelled (Refer Chapter 3and Chapter4). The differentially expressed genes also known are marker genes are then located3.

The two datasets used in this project for implementation of RNA-seq workflow or RNA-seq pipeline are liver and bone marrow. The dataset provided for the analysis of liver data was already normalized whereas the dataset for the analysis of data obtained from bone marrow had raw counts and was filtered and pre-processed for further analysis.

[5] This image has been downloaded from https://www.nature.com/articles/s41467-018-06318-7

2.3 Overview of Bioconductor Ecosystem

Bioconductor (Huber et al.,2015) is a collection of packages in R(Team et al.,2013) mainly used for genomics study and analysis[6][7] . It hosts some widely popular workflows which are used to process, analyse and visualize scRNA-seq datasets. Every package in the bioconductor comes with a vignette[8]. Bioconductor also offers tutorials for certain packages.

2.3.1 Seurat Package

In this project, **Bioconductor package Seurat version 3.0 has been used** for RNA-seq analysis. Seurat (Stuart et al.,2018) is a toolkit for quality control, analysis, and exploration of the single-cell RNA sequencing data. 'Seurat' aims to enable the users to identify and interpret sources of heterogeneity from the single-cell transcriptomic measurements, and to integrate diverse types of single cell data. Seurat is developed and maintained by the Satija lab[9].

Seurat package has been used for the two datasets as it is flexible and comes with fine documentation of its functions along with their implementation tutorials. The paper (MacParland et al.,2018) referred for the analysis of data obtained from human liver is analysed using Seurat version 2. The Seurat package supports improved methods of normalization and removes any variations that occur due to technical faults (Stuart et al.,2018). It provides a flexible framework for multiple dataset integration and was used to integrate various datasets for bone marrow data. Seurat has an advantage in dealing with biological data as it automatically saves data as a sparse matrix.

2.3.2 Seurat Object

Seurat package stores all the information of dataset and the analysis results as a Seurat object. A Seurat object is created with the raw counts which contain various slots which will store not only the raw input data but also results from various computations. Figure2.4represents the function to create a Seurat object[10].

[6] Please refer to the book supported by Bioconductor team for detailed documentation https://osca.bioconductor.org/

[7] The official website for Bioconductor provides a list of online courses for analysis of genomics data. Please refer https://www.bioconductor.org/help/course-materials/

[8] A vignette refers to a document listing functions of a particular package

[9] For Seurar documentation please refer https://satijalab.org/seurat/

[10] Please refer documentation https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/ topics/CreateSeuratObject

```
-------------------------------------------------------------------
CreateSeuratObject(counts, project = "SeuratProject", assay = "RNA")
--------------------------------------------------------------------
```

Figure 2.4: *Creation of Seurat object*

Here, the counts argument refers to the unnormalized data such as raw counts and the project sets a name for the Seurat object and assay gives the name of assay corresponding to input data, in our case RNA. Seurat objects have been created for both the datasets. *R code for the object creation and downstream analysis has been provided in the Appendix A*

A Seurat object named pbmc and dall were created for liver and bone marrow datasets respectively. Figure2.5contains the R commands with its output to illustrate various slots used by the Seurat object to store the metadata created.

```
-------------------------------------------------------------------
#pbmc
An object of class Seurat
20007 features across 8444 samples within 1 assay Active
assay: RNA (20007 features) #slotNames(pbmc)
 [1]  "assays"          "meta.data" "active.assay" "active.ident""graphs" "neighbors"
      "reductions" "project.name"
 [9] "misc"             "version"         "commands"     "tools"
#class(pbmc)
[1] "Seurat" attr(,"package")
[1] "Seurat"
--------------------------------------------------------------------
```

Figure 2.5: *Seurat object (pbmc) and its slots*

Chapter 3

METHODOLOGY

This section of the thesis describes the workflow used for carrying out scRNA-seq analysis on both the datasets. It also highlights the underlying methods adopted by various functions. This chapter has been divided into further sub-sections which represent different steps in the scRNA-seq workflow. Analysis of scRNA datasets follows a similar pipeline for all datasets however the computational functions used may vary from one package to another[1].

The raw counts obtained after capturing of RNA-sequences are pre-processed, filtered (Refer section3.1) and normalised to obtain normalized counts (Refer section3.2). The normalized counts are used to obtain highly variable genes to select the genes of interest (Refer section 3.3). Then, the gene expression matrix with only selected highly variable genes is used as an input to calculate principal components (Refer section3.4). This is done to reduce the number of dimensions of high-dimensional data. Clustering is performed on the reduced dimensions (Refer section3.5) and is visualized (Refer section3.6.1). The clusters thus formed are investigated to find the differentially expressed genes (Refer section3.7) which are used to label the clusters and identify clusters associated with senescent genes.

3.1 Pre-Processing of Data

The new advancements in the field of health and medicine are dependent on the data being analysed. It is essential to ensure the use of high quality data to carry out such analysis. Data recorded for scRNA-seq captures rna sequences in a cell using various sequencing technologies

[1] The functions used in this project are from package Seurat

(Pareek, Smoczynski, and Tretyn,2011), but computational and technological limitations often lead to capture of low quality data (improper or low mRNA reads in a cell). Haque et al. (2017) reveals that even with the high RNA-seq protocols some mRNA sequences were not captured in the cell.To overcome this limitation, **low quality cells are filtered before analysis to ensure that technical effects do not distort the downstream analysis results. The problematic cells which have low library size and cell coverages are removed**. A library is the total number of reads aligned to each cell (total sum of counts across all genes). Cell coverage is the average number of expressed genes in a cell(average number of genes with non-zero counts). **Low-quality / dying cells often exhibit extensive mitochondrial contamination, therefore cells with high mitochondrial genome transcript ratio are also removed from the dataset4.2.1** .

3.1.1 Liver Dataset:

The liver dataset was not filtered as normalized counts were already provided. Cells with a very small library size (<1500) and a very high (>0.5) mitochondrial genome transcript ratio were already removed as high proportions are indicative of poor-quality cells (Ilicic et al.,2016) .The resulting dataset was then normalised3.2to get normalised counts. (Please refer MacParland et al.,2018, to find detailed procedure for filtering of raw counts).

3.1.2 Bonemarrow Dataset:

The raw counts in the Bonemarrow dataset were processed and filtered to remove the unwanted cells. Seurat object stores the number of UMIs [2] (Smith, Heger, and Sudbery,2017) per cell as nCount_RNA, number of features per cell as nFeature_RNA and the fraction of mitochondrial RNA as mt.percent in the metadata slot. A feature-scatter plot is created between mt.percent and nCount_RNA, nFeature_RNA and nCount_RNA to visualize the Quality Control(QC) metrics and feature-feature relationships. This is done to find the cutoff value for carrying out the filtering of cells[3] . The results of this have been presented in section4.

[2] Please refer to the document on https://www.illumina.com/science/sequencing-method-explorer/ kits-and-arrays/umi.html for more information on UMI

[3] This method was suggested by Dr.Mark W Robinson (PhD Immunology) who is an immunologist and health researcher

3.2 Normalization of Data

After filtering of unwanted cells next step is to normalize the data. Measurements from genetically distinct populations may occupy different scales and to make them comparable normalization is performed. The variance in the data tends to depend on the absolute intensity of the data which may lead to false biological conclusions and should be remedied by a normalization method (Evans, Hardin, and Stoebel,2017). By default, **Seurat**(Stuart et al., 2018) **uses a global-scaling normalization method "Log Normalize" that normalizes the gene expression measurements for each cell by the total expression, multiplies this by a scale factor (10,000 by default), and log-transforms the result** (Cole et al.,2019). The liver dataset already had normalized counts and the bone marrow dataset was normalized using function NormalizeData() [4] in R(Team et al.,2013).

3.3 Finding Variable Genes

Variable features are identified after the data has been normalized. Feature selection is an important step when dealing with large datasets as it facilitates improved data quality and speeds up the procedure for analysis (Kursa, Rudnicki, et al.,2010). Here, we will detect genes which are highly variable. **In the case of scRNA-seq data, the variation of genes across cells can be a result of statistical noise rather than biological factors** (Brennecke et al.,2013). **Therefore, it becomes important to identify the subset of genes whose variability in the dataset exceeds the background of statistical noise**.

To find highly variable genes (HVG), genes processing high biological variations are targeted. Gene expression data may have heteroscedasticity (Yip et al.,2017), thus variance cannot be considered as the appropriate factor for determination of HVG.

FindVariableFeatures [5] function in Seurat v3 uses the relationship between variance and mean as the indicator of selecting HVG (Yip, Sham, and Wang,2018). The default setting for selecting the HVG is method="vst" in which mean and variance for each gene is calculated and then log-transformed. A loess curve of polynomials of degree 2 is fit with a span of 0.3 to predict

[4] For documentation please refer https://rdrr.io/cran/Seurat/man/NormalizeData.html

[5] For documentation please refer https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/FindVariableFeatures

the variance of each gene as a function of its mean. The values are then standardised using the below transformation:

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_i}$$

where z_{ij} is the standardized value of feature i in cell j, x_{ij} is the raw value of feature i in cell j, $\bar{x}_i$ is the mean raw value for feature i, and σ_i is the expected standard deviation of feature i.Then, the variance for all standardized values is computed across all cells (Stuart et al.,2018). By default this function selects 2000 genes but this number can be adjusted by the use of argument “nfeatures”3.3.The results have been discussed in section4.

3.4 Dimensionality Reduction

Once the HVG have been selected the next step is to reduce the number of dimensions of the single-cell expression matrices. Genomics data records the activity of thousands of cell or genes which makes the data larger. Larger datasets have computational limitations and are more complex. Dimensionality reduction is important when the number of features(p) are more than the number of cases(n), (p>>n). Dimensionality Reduction methods capture the fundamental structure in the data in fewer dimensions possible (Luecken and Theis,2019). Various Dimensionality Reduction techniques like Principal Component Analysis (PCA) (Pearson,1901), t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton,2008) and Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville,2018) were used to reduce the dimensions and project the data in lower dimensions. This section describes the methodology used for performing dimensionality reduction.

3.4.1 Principal Component Analysis

PCA is a linear feature extraction un-supervised learning technique widely for data with a high number of features (Pearson,1901). It provides fully unsupervised information on the dominant directions of highest variability in the data and can, therefore, be used to investigate similarities between individual samples, or formation of clusters (Ringnér,2008). PCA performs linear mapping of the data to lower dimensional space so that variance can be maximised which is done by calculating eigenvectors from the covariance matrix. Eigenvectors that correspond to the largest eigenvalues are used. (Wold, Esbensen, and Geladi,1987).

The data selected containing the HVG is scaled prior to running PCA using ScaleData[6] function in Seurat v3. The ScaleData functions centres each feature to have a mean of 0 and then scales it by the standard deviation of each feature.

PCA is performed on the selected data using RunPCA [7] function in Seurat v3 and the results are stored in the reductions slot of Seurat Object.

Results of PCA are discussed in section4.1.2. The optimal number of Principal Components (PCs) are picked using the JackStraw[8] and Elbow[9] Plots.

JackStraw Plot was used to determine the optimum number of principal components for clustering. Jackstraw implements a resampling test inspired by the JackStraw procedure. It randomly permutes a subset of the data (1% by default) and reruns PCA, constructing a 'null distribution' of feature scores and thus identifying statistically significant PCs (Chung and Storey,2014).
The JackStraw Plot function [10] was used to visualize JackStraw results (Refer section4.1.2for results).

3.5 Clustering of cells

Seurat v3 performs clustering with the help of FindClusters [11] which is an unsupervised graph based clustering algorithm (Stuart et al.,2018) which calculates k-Nearest Neighbours (k-NN) (Altman,1992; Keller, Gray, and Givens,1985; Ripley and Hjort,1996) and constructs a graph using euclidean distance in PCA space. It then gives weight to the cells on the basis of their Shared Neared Neighbours (SNN) (Ertoz, Steinbach, and Kumar,2002) and the graph is divided into clusters using Louvain algorithm (Blondel et al.,2008).

Clustering is done using the selected PCs from Elbow Plot. FindClusters function implements the above described procedure and selected PCs to cluster cells into distinct populations. The results of clustering have been discussed in4.1.3.

[6] https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/ScaleData
[7] https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/RunPCA
[8] https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/JackStraw
[9] https://www.rdocumentation.org/packages/GMD/versions/0.3.3/topics/elbow
[10] https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/JackStrawPlot
[11] https://www.rdocumentation.org/packages/Seurat/versions/1.4.0/topics/FindClusters

3.6 Dimensionality Reduction Techniques for Visualizations

3.6.1 t-SNE

The results of clustering have been visualised using t-SNE. t-SNE is a non linear projection of data. It creates a Gaussian Probability Distribution that defines relationships between points in a higher dimensional space. It recreates probability distribution using t-distribution in a low dimensional space to prevent crowding (Maaten and Hinton,2008). The paper by Maaten and Hinton (2008) provides a detailed description of this method.

t-SNE has been used for visualization purposes instead of PCA as PCA doesnot does not capture the structure of the data in few dimensions as well as non-linear methods whereas t-SNE preserves the local as well as global structure of the data (Luecken and Theis,2019). **However, the visualizations created by t-SNE amplifies the differences and oversees the connections between cell populations, thereby preserving the local structure at the cost of its global structure** (Luecken and Theis,2019).

RunTSNE[12] (Run t-Distributed Stochastic Neighbor Embedding) function in Seurat v3 has been used to visualize different clusters. The Clusters visualized using t-SNE have been used for differential expression analysis. The results of clustering and t-SNE have been discussed in section4.1.3

3.6.2 UMAP

UMAP is an alternative to t-SNE. Please refer the paper *UMAP: uniform manifold approximation and projection* by McInnes, Healy, and Melville (2018) for detailed description of this visualization technique. In the literature, UMAP is considered better performing in absense of particular biological questions compared to t-SNE due to its speed and ability to scale to large number of cells (Becht et al.,2019; Luecken and Theis,2019). The result by UMAP has been discussed in Chapter4.

We have used t-SNE for labelling of clusters and further downstream analysis for the ease of replication of results by paper *Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations* (MacParland et al.,2018).

[12] For documentation of this function please refer https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/RunTSNE

3.7 Differential Expression Analysis

Differential Expression (DE) Analysis refers to the statistical analysis which is performed on normalised counts to get useful insights in biological data. This downstream analysis can follow a cell and gene-level approach (Luecken and Theis,2019). In this project, we have performed a gene-based approach to find marker genes (also known as differentially expressed genes) in each cluster. Marker genes differentiate one group of cells from another. Differentially expressed genes represent the cluster and are used to annotate it with a meaningful biological label. This label exemplifies the population of cells within each cluster. Clustering is done to partition data, whereas differential expression analysis is used to give biological annotation to each other and ensure its validity.

Differential expression analysis[13] is done by comparing one sample to the other usually a test sample (or mutant group) is compared to a reference group. *In this project since there are no two specific groups or classes, thus, different clusters are considered as different groups and DE analysis is carried out by comparing all the cells in one cluster to all the cells in the specified cluster*. By default, FindMarkers[14] calculates marker genes by comparing two classes, one being all cells in a cluster and the other being all the cells in the dataset. Statistical tests such as the Wilcoxon rank-sum test (Wilcoxon,1945) or the t-test(Student,1908; Box et al.,1987) are often used to rank genes by their difference in expression between these two groups. The null hypothesis implicit in DE tests is that genes have the same distribution of expression values between the two groups. (Refer Section4.1.4for results.)

[13] For tutorials please refer https://satijalab.org/seurat/v3.0/de_vignette.html

[14] For Documentation please refer https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/ topics/FindMarkers

Chapter 4

RESULTS

This section is divided into two sub-sections. Section4.1involves the results obtained from liver dataset [20007 X 8444] and Section4.1discusses the results achieved from bone marrow dataset [33694 X 90653][1].

The methodology discussed in the Chapter3has been implemented to get the results. Results obtained by the analysis of both datasets are similar[2] to the results obtained from the papers *Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations*(MacParland et al.,2018) and *Single Cell Analyses of Human Bone Marrow* (Oetjen et al.,2018).

4.1 Liver Dataset

The liver dataset already had normalized counts, thus, preprocessing, filtering and normalization was not performed on the liver dataset. The normalized counts were used to calculate the Highly Variable Genes (HVG)4.1.1.

The liver dataset was analysed and *the samples were clustered to reveal 21 distinct cell populations in human liver. Differentially expressed genes were calculated for each cluster and were studied to label each cluster* . Clusters with presence of senescence related profiles were detected and gene co-expression analysis was performed to find modules of senescent associated genes.

[1] Note: liver dataset [20007 X 8444] has normalized counts and bone marrow dataset has [33694 X 90653] raw counts

[2] The papers have conducted the analysis using Seurat version 2 and alot of functions have changed from older to newer version. The analysis in this project is performed using Seurat v3.

4.1.1 Highly Variable Genes

The FindVariableFeatures3.3function facilitated the selection of 7000 HVG from the normalized count data. The selected HVG were used for PCA.

Figure4.1labels the top 10 HVG. HVG and non-variable genes have been highlighted with different colouts. In the plot, X-axis function is the mean expression level (mean count of a gene across all cells), and for Y-axis standardized variance is the log(Variance/mean). *Note:All mean/variance calculations are not performed in log-space, but the results are reported in log-space, for detailed explaination of standardized variance refer3.3.* [3]

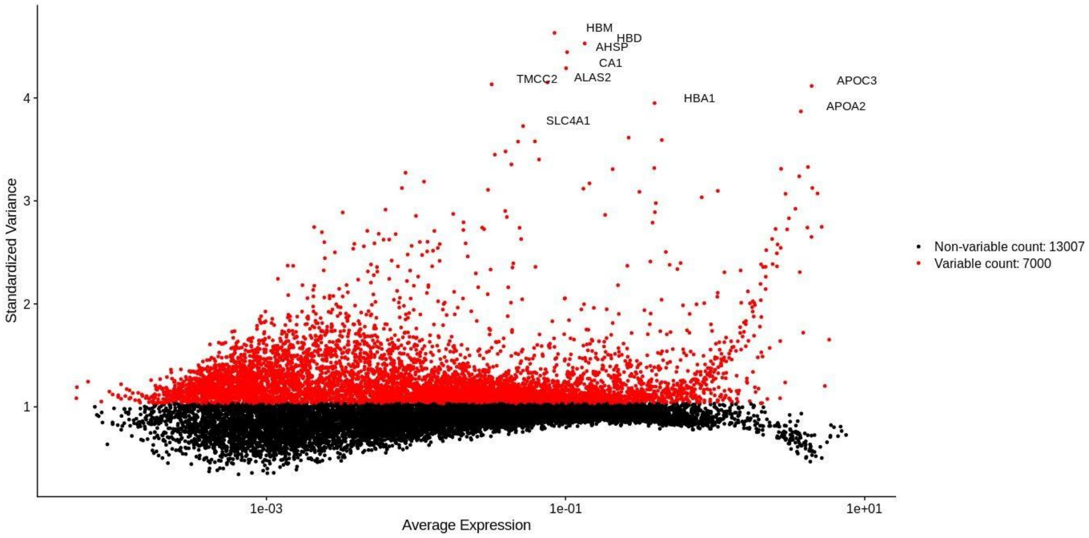


Figure 4.1: *Variable Genes Plot*

By default the FindVariableFeatures function3.3detects top 2000 HVG. For study and labelling purposes (considering a big dataset), the nfeatures argument in the function has been set to 7000 to get 7000 HVG (A similar number was selected by the liver paper). The gene expression matrix is now 7000 (features or genes) X 8444 (cases or cells).

[3] For documentation please refer https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/FindVariableFeatures

4.1.2 Principal Component Analysis:

After computation of HVG, the data was scaled and PCA was performed on the variablegenes. 50 principal components (PC) were obtained from PCA and the first two principal components accounted for most of the variability in the dataset.

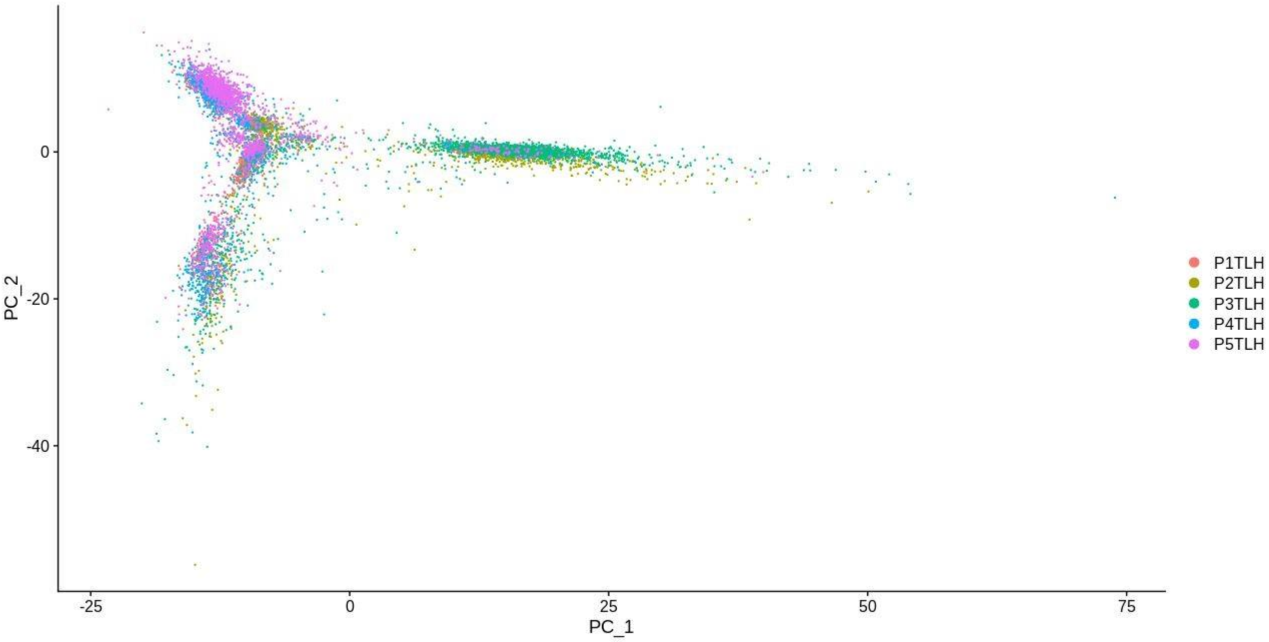


Figure 4.2: *Principal Component Plot of five samples*

Figure4.2plots the first two principal components and colours them by the 5 samples in the dataset. The samples are the 5 liver donors. The plot reveals an overlap between the samples P1TLH, P4TLH and P5TLH. However, sample P3TLH seems seperated from the rest of the clusters.

VizDimLoadings[4] function in Seurat v3 can be used to visualize top genes associated with principal components. Figure4.3visualizes top genes associated with PC1 and PC2 with PC scores on x-axis and genes on y-axis. Larger absolute value of component corresponds to a more important gene.

[4] https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/VizDimLoadings

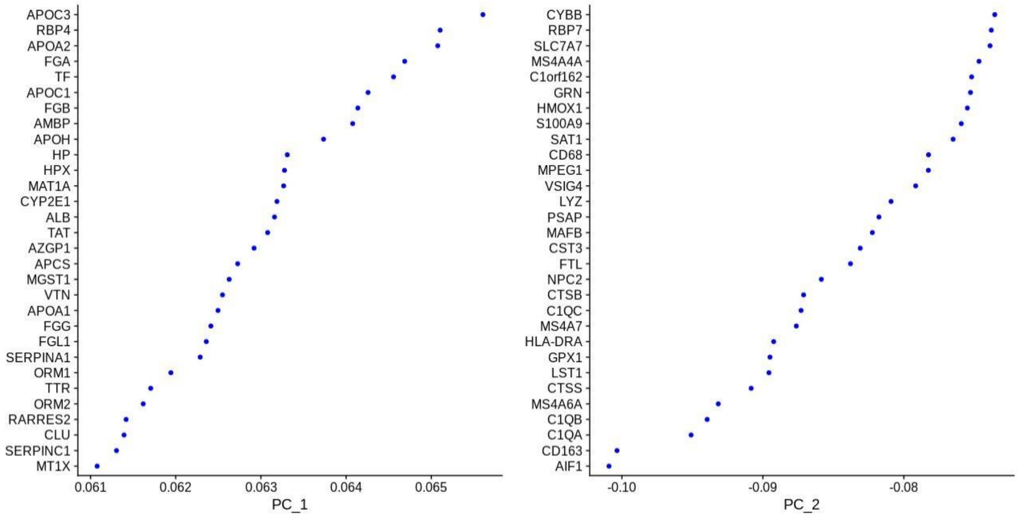


Figure 4.3: *Loadings of PC1 and PC2*

The optimum number of PCs have been Calculated with the help of JackStraw and Elbow plots.

The JackStraw Plot function[5] was used to visualize JackStraw results. Figure4.4demonstrates the distribution of p-values for each PC with a uniform distribution (dashed line) and the PCs with curved lines above the distribution are statistically significant PCs.

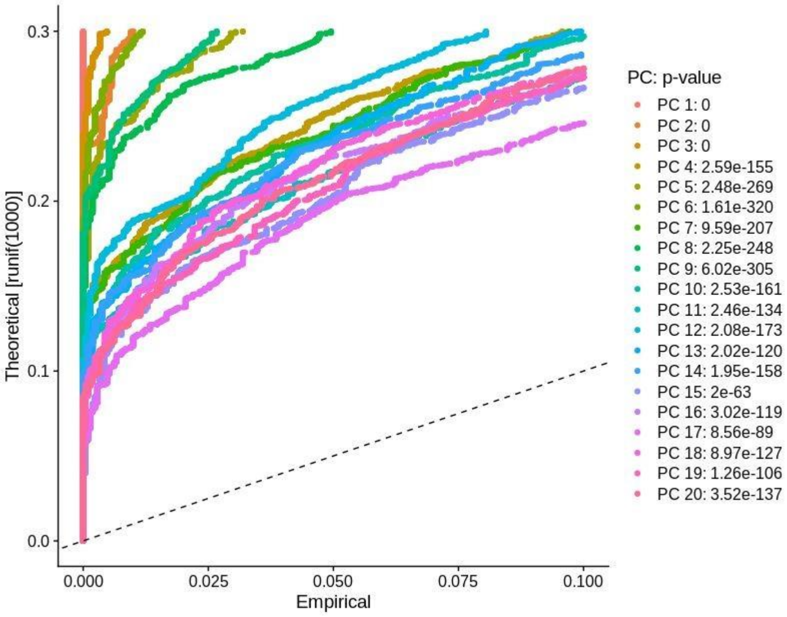


Figure 4.4: *JackStaw Plot*

[5] For documentation please refer https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/ topics/JackStrawPlot

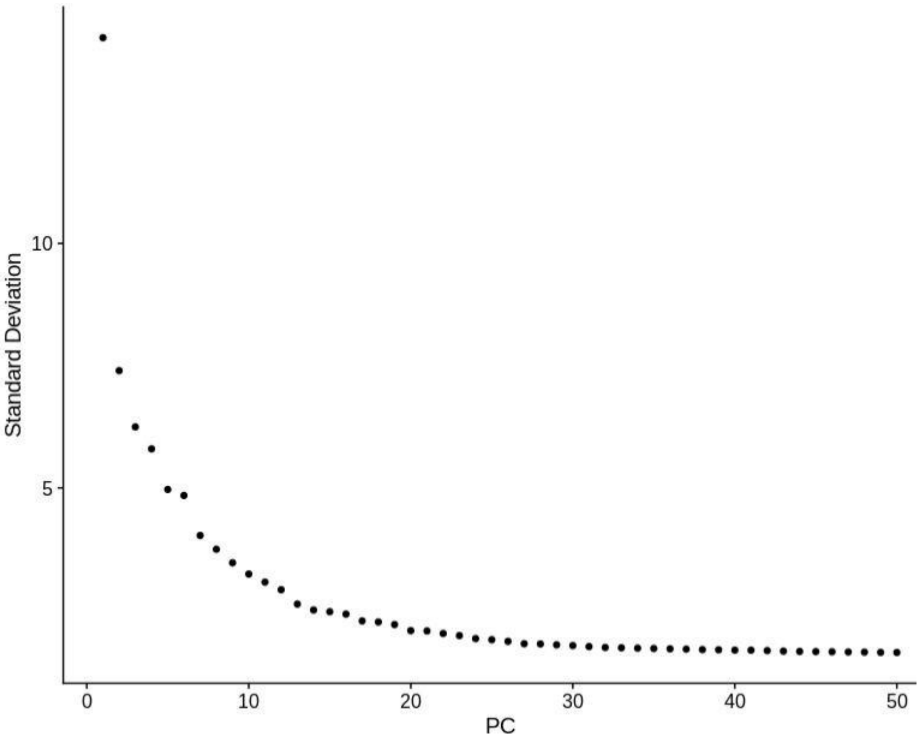


Figure 4.5: *Elbow Plot*

Figure4.5plots all PCs obtained from PCA and is used to select the number of PCs for downstream analysis. *Note:JackStraw Plot can only represent upto 20 PCs, thus, ElbowPlot has been used to visualize all PCs and determine the appropriate number of PCs for further analysis.* Number of PCs are represented on y-axis and their standard deviation on x-axis. 29 Principal Components have been selected for downstream analysis.

4.1.3 Clustering and Visualization

Selected 29 PCs were used as an input for clustering. 21 clusters of distinct cell populations were obtained with the help of graph-based clustering3.5and the results are visualised using t-SNE3.6.1. The clusters are visualized by setting the perplexity parameter to 27 and using the same number of PCs as selected for clustering. Since changing the perplexity parameter may give very different plots, thus, the interpretation of these plots can be misleading (Wattenberg, Viégas, and Johnson,2016). In this project, t-SNE has been used just to visualize the clusters.

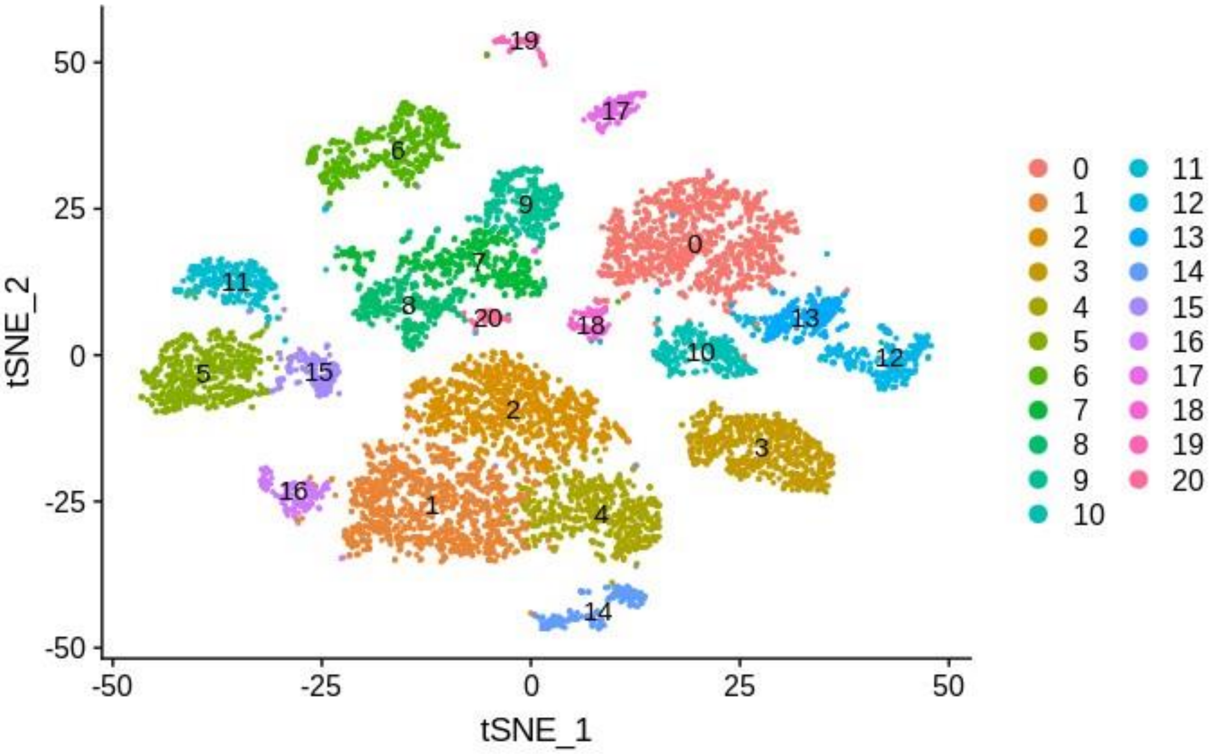


Figure 4.6: *t-SNE projection coloured by clusters*

Figure4.6demonstrates the two dimensional representation of clusters formed. 8444 liver cells are projected by t-SNE projection where each point represents a cell. The plot represents the t-SNE projection of the cells that share same genes grouped by colours obtained from the clustering results. Cluster number is assigned to the various clusters formed.

Figure4.7represents the t-SNE projection of all cells (8444) coloured by the five samples or donors. The plot also highlights a big green island (Sample P3TLH) in the center distinct from all other clusters while some overlap is seen between the other samples[6]. The results of t-SNE are similar to the output obtained by visualization of first two PCs in PCA.

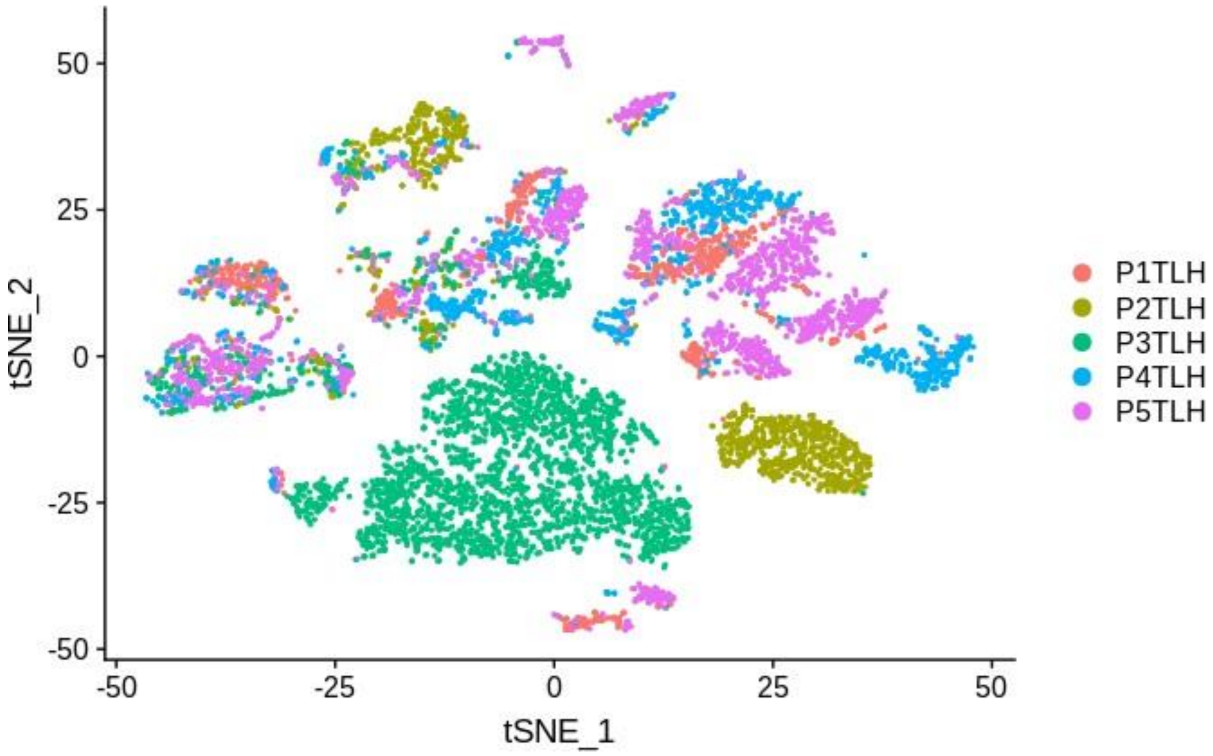


Figure 4.7: *Visualization of t-SNE results by samples*

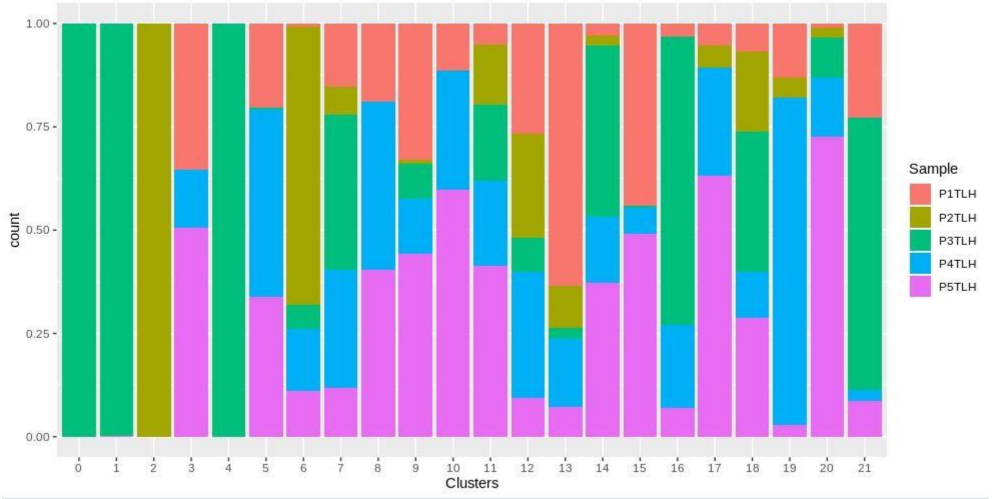


Figure 4.8: *Contribution of cells proportions in each sample to cluster*

[6] The reason for the overlap is unidentified. The plot needs to be discussed with domain expert Dr.Mark W Robinson

Figure4.8represents the proportion of cells that contributed to each cluster. It reveals that cells in cluster 0,1 and 3 are most vulnerable to removal and hence are less likely to have senescent transcriptional profiles wherease the cells which are found in most of the liver samples need to be investigated.

t-SNE plots were used for labelling of the clusters by the different types of cell populations, thereby, using them for locating the clusters associated with senescent genes.

4.1.4 Differentially Expressed genes:

Differentially expressed genes were used for labelling of clusters. 105 differentially expressed genes were selected by extracting top 5 differentially expressed genes in each cluster. Table 4.2 displays the list of top two differentially expressed genes in each cluster.(Since the genes are highly important, the resultant p-values are small and thus are rounded off to 0). The list of genes were studied [7] to label the clusters by different cell populations.

The genes in Table 4.1 (provided at end of section 4.1) have been calculated by FindAllMarkers[8] in Seurat v3. This function is similar to FindMarkers function discussed in Section3.7, it just automates the process for each cluster. By default, differentially expressed or marker genes are calculated by Wilcox Method (Wilcoxon,1945) which identifies differentially expressed genes between two groups of cells using a Wilcoxon Rank Sum test. It is a non-parametric alternative to a two sample t-test and follows a null hypothesis of two groups having same distribution and same median. The p-values are calculated from Wilcoxon Rank Sum Statistic. Please refer the book *Nonparametric statistics for the behavioral sciences*(Siegel,1956) for detailed description.

Table 4.1: *Description for Table 4.2*

Arguments	Description
p_val	p_val (unadjusted)
avg_logFC	log fold-chage of the average expression between the two groups
pct.1	The percentage of cells where the feature is detected in the first group
pct.2	The percentage of cells where the feature is detected in the second group
p_val_adj	Adjusted p-value, based on bonferroni correction using all features

[7] *The list of genes was extracted and provided to Dr.Mark Robinson for evaluation of clusters and to label different clusters*

[8] For documentation please refer https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/ topics/FindAllMarkers

The differential genes extracted were studies and 18 out of 21 clusters were successfully labelled[9]. Figure4.9represents the labels of the clusters for each defined clusters.

The genes in different were studied and the clusters representing cell populations namely Cytotoxic Lymphocyte (NK cells) and B-cells were investigated closely as they might have presence of senescent related profiles[10] .

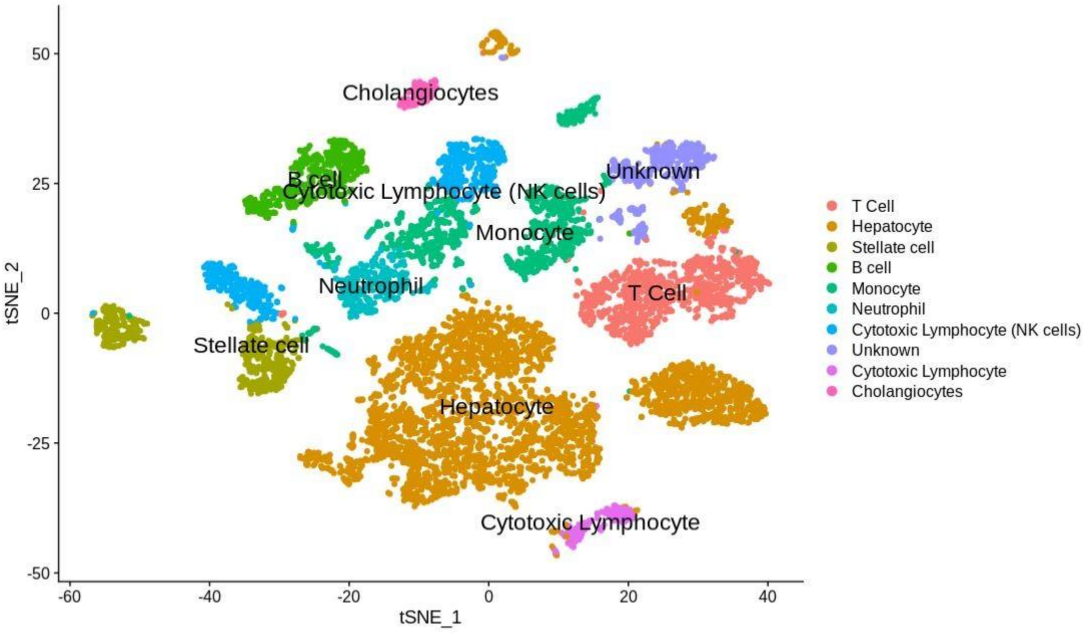


Figure 4.9: *Labelling of the clusters by their biological names*

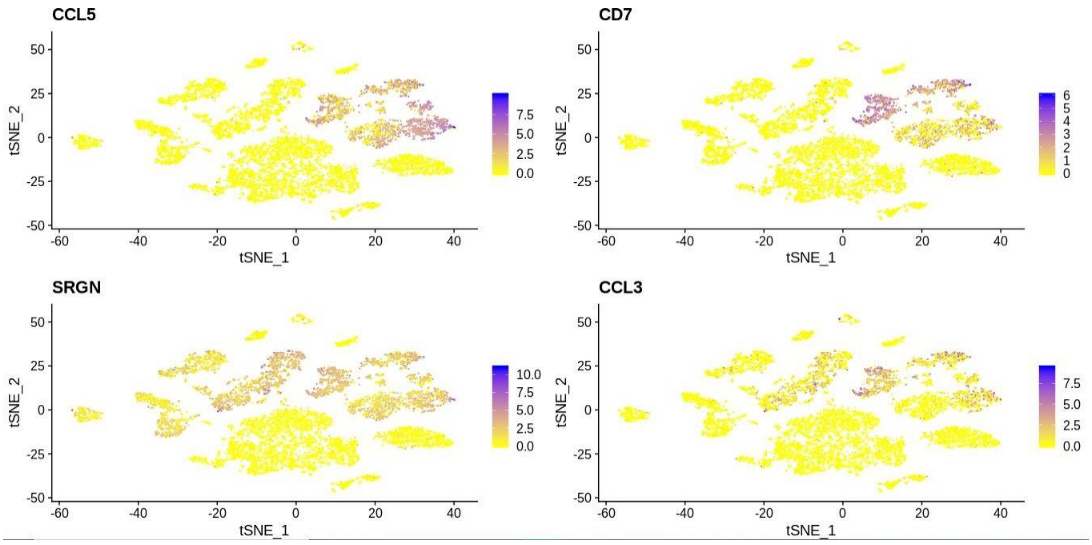


Figure 4.10: *Visualization of distinguishing genes*

[9] *The list of genes was extracted and provided to Dr.Mark Robinson for evaluation of clusters and to label different clusters*

[10] *(as suggested by domain expert Dr.Mark W Robinson)*

The distribution of differentially expressed genes in Cluster 1 can be seen from plot4.10. As observed from the t-SNE projection genes namely CCL5, CD7, SRGN, CCL3 have been differentially expressed in top right corner of the map which corresponds to the cluster of NK cells.

4.1.5 Detection of Senescent Genes

The list of senescence genes was obtained from the paper *Unmasking transcriptional heterogeneity in senescent cells* (Hernandez-Segura et al.,2017). The list was used to extract the genes associated with cellular senescence in the liver dataset and 54 senescent associated transcripts were detected. The list of senescenet genes has been added to Appendix B.

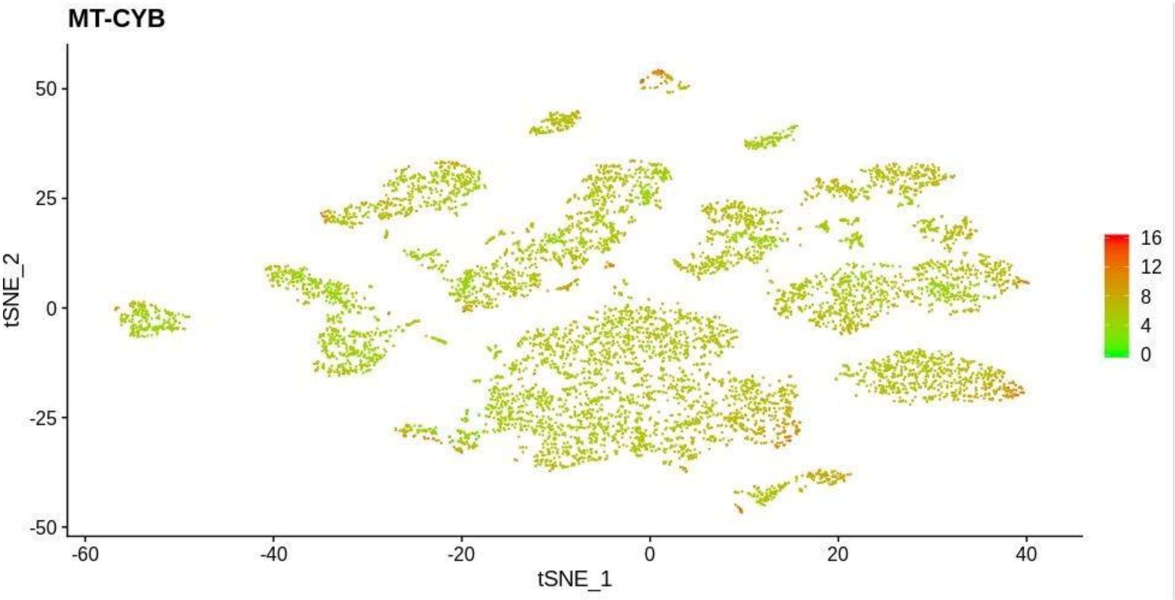


Figure 4.11: *Visualization of MT-CYB*

MT-CYB is a gene associated with senescene. Figure4.11reveals its presence in nearly all the clusters. However, it is low expressed in some clusters and highly expressed in others. The legend describes the expression level of the gene from green being nearly 0 to red being higest level of expression. Figure4.12plots some of the senescent genes to detect their presence. Since each point on the t-SNE projection represents a cell, thus, it can be seen that the senescent genes represented in the plot have occurences in different cells. The dark dots highlight the high expression of that gene.

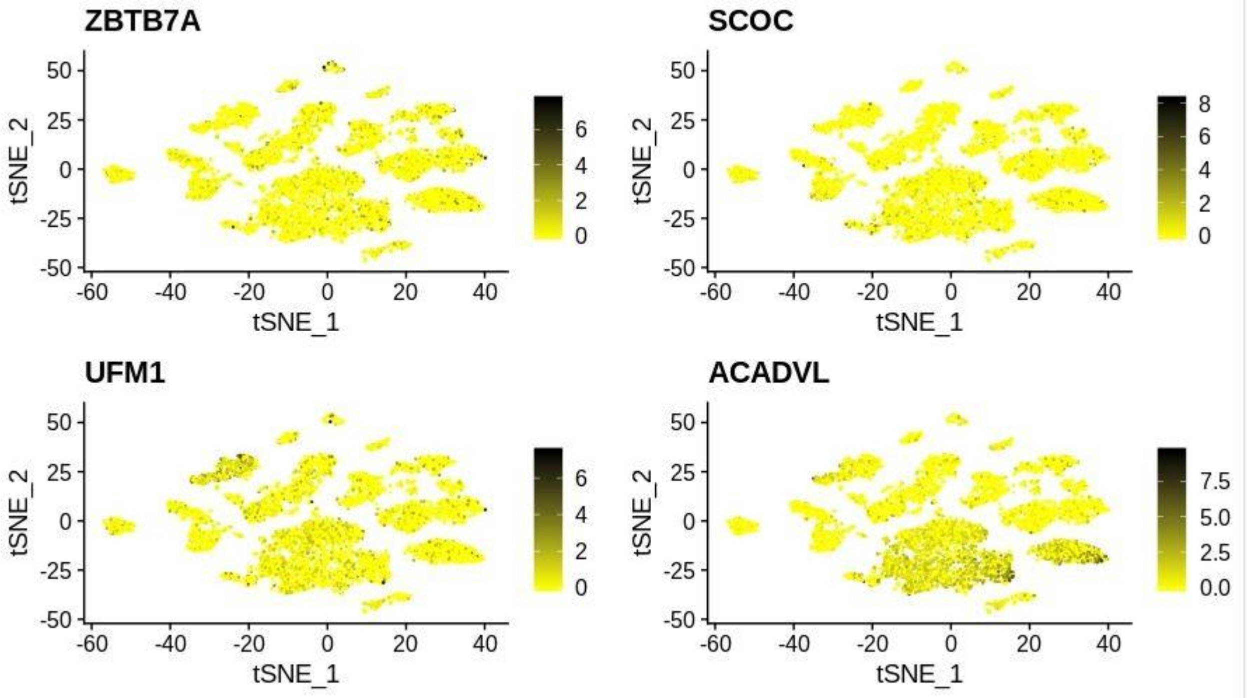


Figure 4.12: *t-SNE projection of some senescent genes*

Table 4.2: *Differentially Expressed genes in each Cluster*

X	p_val	avg_logFC	pct.1	pct.2	p_val_adj	cluster	gene
1	0	1.526	0.669	0.296	0	0	HAMP
2	0	1.548	0.295	0.111	0	0	LINC00467
3	0	2.770	0.674	0.069	0	1	CD2
4	0	2.615	0.847	0.040	0	1	CD3D
5	0	0.403	0.786	0.159	0	2	BCHE
6	0	0.417	0.433	0.101	0	2	GHR
7	0	4.415	0.971	0.224	0	3	SCD
8	0	3.954	0.905	0.138	0	3	HMGCS1
9	0	3.737	0.268	0.128	0	4	RND3
10	0	3.714	0.260	0.135	0	4	ABCD3
11	0	3.705	0.696	0.014	0	5	MGP
12	0	2.392	0.827	0.069	0	5	PRSS23
13	0	10.107	0.561	0.136	0	6	IGLC2
14	0	9.572	0.549	0.090	0	6	IGHG1
15	0	2.204	0.958	0.134	0	7	CD7
16	0	2.102	0.965	0.237	0	7	CMC1
17	0	3.965	0.851	0.196	0	8	HLA-DPB1
18	0	3.337	0.829	0.166	0	8	HLA-DPA1
19	0	6.020	0.807	0.011	0	9	CD5L
20	0	5.306	0.906	0.034	0	9	MARCO
21	0	5.033	0.992	0.127	0	10	S100A8
22	0	4.358	0.994	0.167	0	10	S100A9
23	0	4.257	0.988	0.071	0	11	GNLY
24	0	3.863	0.270	0.031	0	11	PTGDS
25	0	2.940	0.901	0.040	0	12	CLEC1B
26	0	3.068	0.997	0.121	0	12	DNASE1L3
27	0	1.565	0.888	0.056	0	13	CYP2A7
28	0	1.671	0.374	0.019	0	13	CYP3A7
29	0	6.908	0.719	0.262	0	14	AKR1C1
30	0	6.852	0.456	0.252	0	14	HPR
31	0	2.511	0.623	0.005	0	15	MS4A1
32	0	2.337	0.769	0.090	0	15	LTB
33	0	2.095	0.764	0.045	0	16	FGFBP2
34	0	2.017	0.819	0.121	0	16	CD3D
35	0	2.272	0.957	0.006	0	17	FXYD2
36	0	2.284	0.299	0.021	0	17	SCGB3A1
37	0	2.546	0.982	0.104	0	18	STMN1
38	0	1.392	1.000	0.245	0	18	HMGB2
39	0	13.288	0.945	0.025	0	19	HBD
40	0	10.932	0.945	0.011	0	19	AHSP
41	0	4.112	0.875	0.019	0	20	ACTA2
42	0	3.953	0.844	0.004	0	20	COL1A1

4.2 Bonemarrow Dataset

This section contains results of the analysis conducted on Bone marrow dataset. The dataset was filtered4.2.1and normalized to obtain normalized counts. The normalized counts were used to find HVG. This dataset was analysed using a similar pipeline to liver dataset.

The raw counts were filtered and the number of cells after removal of low-quality cells was reduced to 76645 from 90653. Remaining 76645 cells were used for final analysis (The results after filtering of cells are exactly same as the paper *Single Cell Analyses of Human Bone Marrow* (Oetjen et al.,2018).

4.2.1 Pre-Processing of Data

The plot4.13has been created to visualize QC metrics and feature-feature relationships to find the optimal cut-off value for filtering of cells.

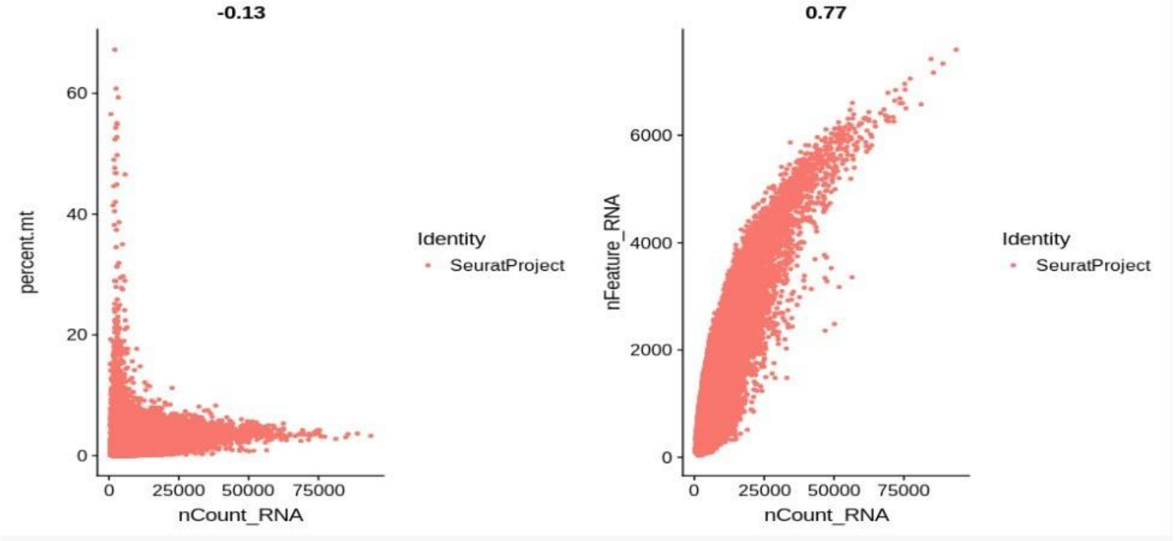


Figure 4.13: *Feature Scatter Plot*

Figure4.13represents fraction of mitochondrial RNA (x-axis) and number of features (x-axis) verses number of UMIs per cell (y-axis), this plot assists in deciding optimum cutoff value for percent.mt and nFeatureRNA. Pearson correlation between the two features is displayed above the plot. Cells with a very small library size (<500) and a very high (>8%) mitochondrial genome transcript ratio were removed(Ilicic et al.,2016).After elimination of low quality cells the remaining cells were used for analysis.

4.2.2 Clustering Results

After filtering the of cells HVG were selected using the default number of features(2000). The filtered gene expression matrix with 2000 (genes or features) and 76645 (cases or cells) was used as an input to perform PCA. 15 PCs were selected from elblow plot5.3for downstream analysis. Selected PCs were used for clustering and 21 distinct cell populations in Bone marrow were revealed. The clusters have been projected using t-SNE4.14and UMAP4.15. Both the plots show similar results as Clusters 0,1,3,6 form an island and are distinct from rest of the clusters.

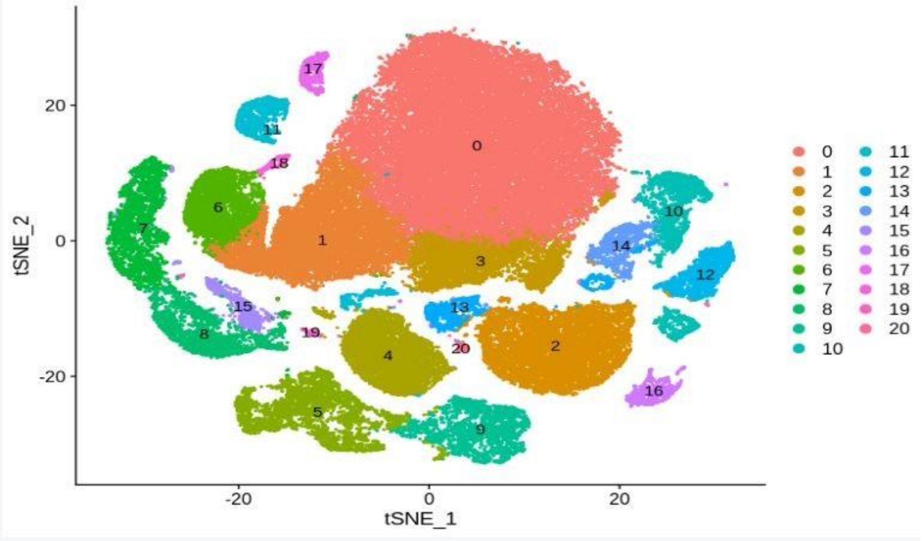


Figure 4.14: *t-SNE Projection of Clusters*

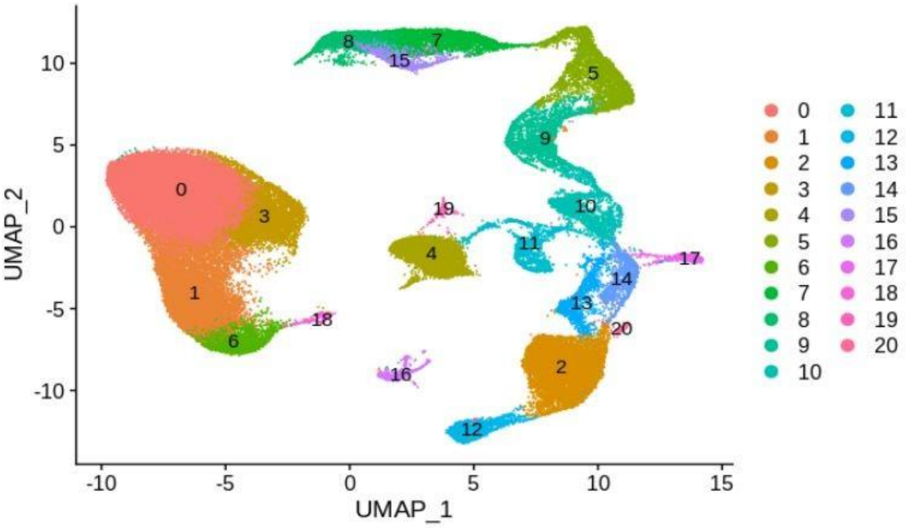


Figure 4.15: *UMAP Projection of Clusters*

The clusters obtained after analysing the bone marrow dataset are similar to the clusters obtained in paper *Single Cell Analyses of Human Bone Marrow* (Oetjen et al.,2018). The difference in results is due to the difference in packages used in both the analysis. The analysis in the paper was performed using Seurat v2 and the analysis in this project has been carried out using v3 of Seurat.

Differential expression analysis was performed on each cluster and a list of differentially expressed genes has been extracted. (Refer Appendix B)

Chapter 5

Conclusion and Future Scope

5.1 Discussion

This thesis provides an introduction to a wide range of methods used in the analysis of a genomics dataset. With a wide range of genomics data being generated, there is a need for analyzing such datasets to discover the new biological insights. Analysing these datasets is challenging as it requires both biological background and appropriate computational skills. This thesis presented an overview of a scRNA-seq pipeline which can be easily understood and implemented.

The pipeline has been implemented on the liver dataset first and revealed 21 distinct cell populations and the presence of senescent genes. Co-expression analysis of the senescent genes has not been completed yet due to the time restriction and the results for finding genes co-expressed with senescent genes are unfinished. CEMI tool (Sa Tavares Russo et al.,2018) has been used to find the co-expressed genes modules and 3 such modules (groups or cluster of co-expressed genes) were detected (Refer Appendix B). The interpretation of modules is not presented as it needs to be discussed with the domain expert. Bone marrow dataset was pre-processed, normalized and analyzed to reveal 20 distinct cell populations, but the cell populations are yet to be labeled.

5.2 Conclusion

This thesis presented the workflow used for carrying out scRNA-seq analysis highlighting the underlying methods adopted by various functions. It discusses a new format to store large matrices (Sparse Matrix Format) and highlights the scRNA-seq analysis supported by Seurat package. The visualizations created using t-SNE were used to locate the genes. Three different methods for dimensionality reduction have been discussed which are flexible and easy to understand and implement.

The liver dataset was analyzed and the samples were clustered to reveal 21 distinct cell populations and the presence of 54 senescent genes in the human liver. Differentially expressed genes were calculated for each cluster. In collaboration with the biological expert, Dr Mark W. Robinson the clusters were successfully given biological labels. However, a few clusters were still left unlabelled and require a more in-depth investigation of genes to identify the cell populations represented by such clusters.

The raw counts from the bone marrow dataset were filtered, normalized and analyzed to identify 20 distinct cell clusters, but the clusters are yet to be labeled.

5.3 Future Work

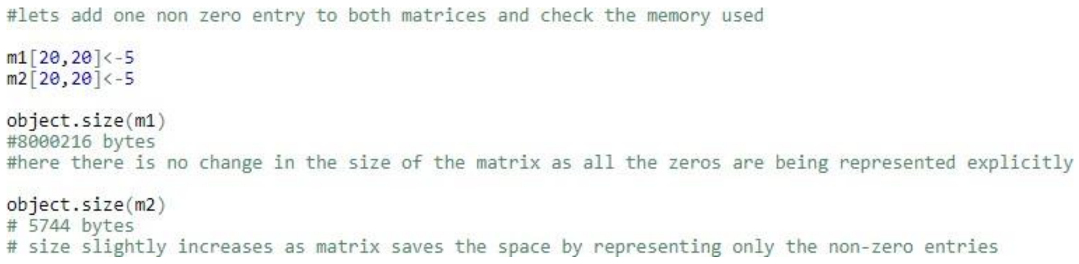
ScRNA-seq is a new advancement and has numerous unexplored areas. The scope of this project is vast and can be extended to achieve undiscovered results. I would like to propose the below extensions to this project:

- Seurat version 3.0 was used to analyze the datasets in this project, but there are multiple packages in R like ascent, liger and single cell experiment which can be used for the same. The same analysis can be carried out using these packages and the results can be compared.
- Packages like wgcna and CEMItool in R are used for gene-coexpression network analysis which can be used to detect genes correlated with the senescent genes. These packages were not fully used due to the lack of time, but can be used to perform network analysis.
- Differential Expression analysis can also be performed with various different packages and a study similar to the paper *Evaluation of methods for differential expression analysis on multi-group RNA-seq count data* (Tang et al.,2015) using the two datasets

Appendix A

This section contains some plots for datasets.

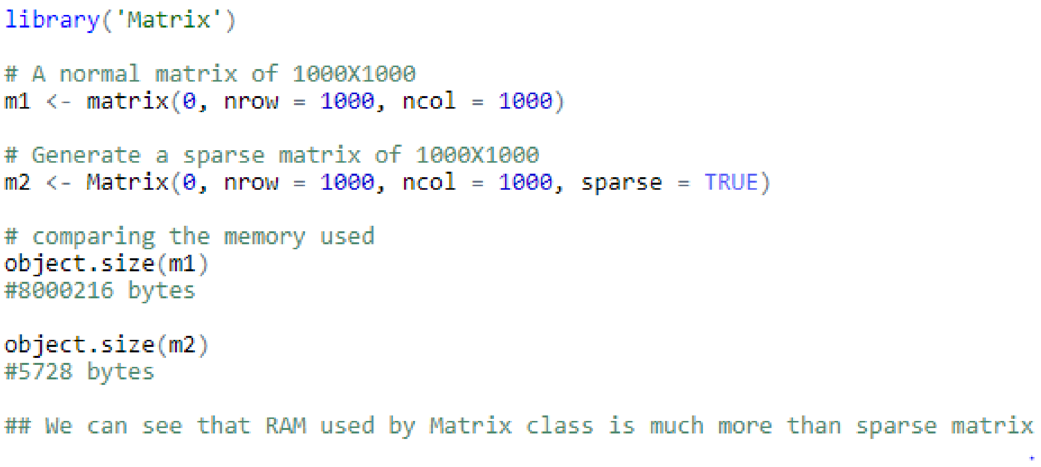
- Bone marrow dataset



```
#lets add one non zero entry to both matrices and check the memory used

m1[20,20]<-5
m2[20,20]<-5

object.size(m1)
#8000216 bytes
#here there is no change in the size of the matrix as all the zeros are being represented explicitly

object.size(m2)
# 5744 bytes
# size slightly increases as matrix saves the space by representing only the non-zero entries
```

Figure 5.1: *Bone marrow dataset at a glance*

- Liver Dataset



```
library('Matrix')

# A normal matrix of 1000X1000
m1 <- matrix(0, nrow = 1000, ncol = 1000)

# Generate a sparse matrix of 1000X1000
m2 <- Matrix(0, nrow = 1000, ncol = 1000, sparse = TRUE)

# comparing the memory used
object.size(m1)
#8000216 bytes

object.size(m2)
#5728 bytes

## We can see that RAM used by Matrix class is much more than sparse matrix
```

Figure 5.2: *Glimpse of Liver Dataset*

- Elbow plot for Bone marrow dataset

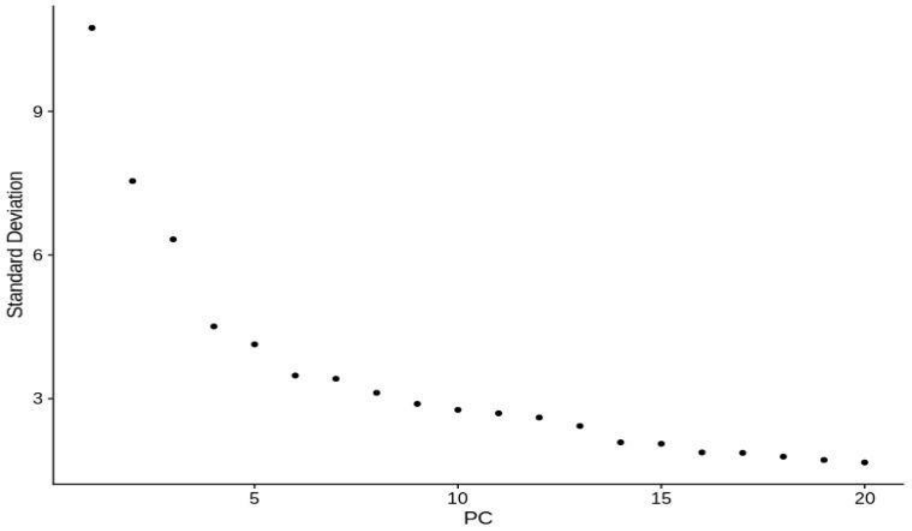


Figure 5.3: *Elbow Plot for Bone Marrow Dataset*

- List of Senescent Genes found in Liver Dataset

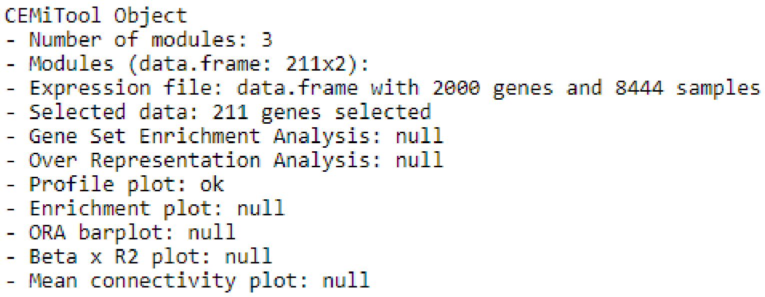
```
-----------------------------------------------------------------------------
[1] "CDKN2A"   "CDKN1A"  "NFIA"      "EFNB3"   "SPATA6"  "GSTM4"   "MEIS1"
"PATZ1"    "USP6NL"

[10] "ASCC1"    "CREBBP"  "ZC3H4"    "ARID2"    "ICE1"     "PDS5B"    "SPIN4"
"TRDMT1"  "STAG1"

[19] "RHNO1"    "PCIF1"    "CNTLN"    "KCTD3"    "SMO"      "GDNF"     "PLK3"
"TSPAN13" "CCND1"

[28] "P4HA2"    "SLC10A3" "ZBTB7A"   "SCOC"     "UFM1"     "B4GALT7" "ACADVL"
"POFUT2"   "TAF13"

[37] "NOL3"     "ADPGK"    "DDA1"     "ZNHIT1"   "CHMP5"    "TOLLIP"   "KLC1"
"TMEM87B" "BCL2L2"

[46]  "SUSD6"    "DYNLT3"   "RAI14"    "FAM214B" "PDLIM4"   "DGKA"     "PLXNA3"
```

CHAPTER 5. CONCLUSION AND FUTURE

"MT-CYB"

- CEMItool results

Three modules were detected using CEMItool5.4. Due to long list of genes in modules, it has not been published (The csv file for module has been uploaded with R-scripts to moodle.)



```
CEMiTool Object
- Number of modules: 3
- Modules (data.frame: 211x2):
- Expression file: data.frame with 2000 genes and 8444 samples
- Selected data: 211 genes selected
- Gene Set Enrichment Analysis: null
- Over Representation Analysis: null
- Profile plot: ok
- Enrichment plot: null
- ORA barplot: null
- Beta x R2 plot: null
- Mean connectivity plot: null
```

Figure 5.4: *R results highlighting 3 modules*

- Differentially expressed genes (top gene in each cluster) in Bone Marrow Dataset

Table 5.1: *Differentially Expressed genes in each Cluster*

X	cluster	genes.1	genes.2	genes.3	genes.4	genes.5	genes.6	genes.7
1	0	LTB	CCL5	S100A9	LTB	CD74	HBM	GNLY
2	1	HBD	HBA2	PRDX2	SPINK2	IGLL1	LST1	CST3
3	2	MPO	HBA2	IGKC	ITM2C	STMN1	HBB	HBA2
4	3	LTB	CCL5	S100A9	LTB	CD74	HBM	GNLY
5	4	HBD	HBA2	PRDX2	SPINK2	IGLL1	LST1	CST3
6	5	MPO	HBA2	IGKC	ITM2C	STMN1	HBB	HBA2
7	6	LTB	CCL5	S100A9	LTB	CD74	HBM	GNLY
8	7	HBD	HBA2	PRDX2	SPINK2	IGLL1	LST1	CST3
9	8	MPO	HBA2	IGKC	ITM2C	STMN1	HBB	HBA2
10	9	LTB	CCL5	S100A9	LTB	CD74	HBM	GNLY
11	10	HBD	HBA2	PRDX2	SPINK2	IGLL1	LST1	CST3
12	11	MPO	HBA2	IGKC	ITM2C	STMN1	HBB	HBA2
13	12	LTB	CCL5	S100A9	LTB	CD74	HBM	GNLY

14	13	HBD	HBA2	PRDX2	SPINK2	IGLL1	LST1	CST3
15	14	MPO	HBA2	IGKC	ITM2C	STMN1	HBB	HBA2
16	15	LTB	CCL5	S100A9	LTB	CD74	HBM	GNLY
17	16	HBD	HBA2	PRDX2	SPINK2	IGLL1	LST1	CST3
18	17	MPO	HBA2	IGKC	ITM2C	STMN1	HBB	HBA2
19	18	LTB	CCL5	S100A9	LTB	CD74	HBM	GNLY
20	19	HBD	HBA2	PRDX2	SPINK2	IGLL1	LST1	CST3
21	20	MPO	HBA2	IGKC	ITM2C	STMN1	HBB	HBA2

Bibliography

Altman, NS (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* **46**(3), 175–185.

Becht, E, L McInnes, J Healy, CA Dutertre, IW Kwok, LG Ng, F Ginhoux, and EW Newell (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* **37**(1), 38.

Blondel, VD, JL Guillaume, R Lambiotte, and E Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**(10), P10008.

Box, JF et al. (1987). Guinness, Gosset, Fisher, and small samples. *Statistical science* **2**(1), 45–52.

Brennecke, P, S Anders, JK Kim, AA Kołodziejczyk, X Zhang, V Proserpio, B Baying, V Benes, SA Teichmann, JC Marioni, et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods* **10**(11), 1093.

Chung, NC and JD Storey (2014). Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics* **31**(4), 545–554.

Cole, MB, D Risso, A Wagner, D DeTomaso, J Ngai, E Purdom, S Dudoit, and N Yosef (2019). Performance assessment and selection of normalization procedures for single-cell RNA-seq. *Cell systems* **8**(4), 315–328.

Crick, F (1970). Central dogma of molecular biology. *Nature* **227**(5258), 561.

Ertoz, L, M Steinbach, and V Kumar (2002). A new shared nearest neighbor clustering algorithm and its applications. In: *Workshop on clustering high dimensional data and its applications at 2nd SIAM international conference on data mining*, pp.105–115.

Evans, C, J Hardin, and DM Stoebel (2017). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics* **19**(5), 776–792.

Haque, A, J Engel, SA Teichmann, and T Lönnberg (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine* **9**(1), 75.

Hayflick, L and PS Moorhead (1961). The serial cultivation of human diploid cell strains. *Experimental cell research* **25**(3), 585–621.

Hernandez-Segura, A, TV de Jong, S Melov, V Guryev, J Campisi, and M Demaria (2017). Unmasking transcriptional heterogeneity in senescent cells. *Current Biology* **27**(17), 2652–2660.

Huber, W, VJ Carey, R Gentleman, S Anders, M Carlson, BS Carvalho, HC Bravo, S Davis, L Gatto, T Girke, R Gottardo, F Hahne, KD Hansen, RA Irizarry, M Lawrence, MI Love, J MacDonald, V Obenchain, AK Ole's, H Pag'es, A Reyes, P Shannon, GK Smyth, D Tenenbaum, L Waldron, and M Morgan (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **12**(2), 115–121.

Hwang, B, JH Lee, and D Bang (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine* **50**(8), 96.

Ilicic, T, JK Kim, AA Kolodziejczyk, FO Bagger, DJ McCarthy, JC Marioni, and SA Teichmann (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome biology* **17**(1), 29.

Keizer, PL de (2017). The fountain of youth by targeting senescent cells? *Trends in molecular medicine* **23**(1), 6–17.

Keller, JM, MR Gray, and JA Givens (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics* (4), 580–585.

Koenker, R, P Ng, et al. (2003). SparseM: A sparse matrix package for R. *Journal of Statistical Software* **8**(6), 1–9.

Kursa, MB, WR Rudnicki, et al. (2010). Feature selection with the Boruta package. *J Stat Softw* **36**(11), 1–13.

Lowe, R, N Shirley, M Bleackley, S Dolan, and T Shafee (2017). Transcriptomics technologies. *PLoS computational biology* **13**(5), e1005457.

Luecken, MD and FJ Theis (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology* **15**(6).

Maaten, Lvd and G Hinton (2008). Visualizing data using t-SNE. *Journal of machine learning research* **9**(Nov), 2579–2605.

MacParland, SA, JC Liu, XZ Ma, BT Innes, AM Bartczak, BK Gage, J Manuel, N Khuu, J Echeverri, I Linares, et al. (2018). Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nature communications* **9**(1), 4383.

McInnes, L, J Healy, and J Melville (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Metsis, A, U Andersson, G Baurn, P Ernfors, P Lnnerberg, A Montelius, M Oldin, A Pihlak, and S Linnarsson (2004). Whole-genome expression profiling through fragment display and combinatorial gene identification. *Nucleic acids research* **32**(16), e127–e127.

Oetjen, KA, KE Lindblad, M Goswami, G Gui, PK Dagur, C Lai, LW Dillon, JP McCoy, and CS Hourigan (2018). Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. *JCI insight* **3**(23).

Pan, J, D Li, Y Xu, J Zhang, Y Wang, M Chen, S Lin, L Huang, EJ Chung, DE Citrin, et al. (2017). Inhibition of Bcl-2/xl with ABT-263 selectively kills senescent type II pneumocytes and reverses persistent pulmonary fibrosis induced by ionizing radiation in mice. *International Journal of Radiation Oncology* Biology* Physics* **99**(2), 353–361.

Papalexi, E and R Satija (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology* **18**(1), 35.

Pareek, CS, R Smoczynski, and A Tretyn (2011). Sequencing technologies and genome sequencing. *Journal of applied genetics* **52**(4), 413–435.

Pearson, K (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572.

Ringnér, M (2008). What is principal component analysis? *Nature biotechnology* **26**(3), 303.

Ripley, BD and N Hjort (1996). *Pattern recognition and neural networks*. Cambridge university press.

Sa Tavares Russo, P de, GR Ferreira, LE Cardozo, MC Burger, R Arias-Carrasco, SR Maruyama, TDC Hirata, DS Lima, FM Passos, KF Fukutani, M Lever, JS Silva, V Maracaja-Coutinho, and HTI Nakaya (2018). CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics* **19**(56), 1–13.

Schug, J, WP Schuller, C Kappen, JM Salbaum, M Bucan, and CJ Stoeckert (2005). Promoter features related to tissue specificity as measured by Shannon entropy. *Genome biology* **6**(4), R33.

Siegel, S (1956). *Nonparametric statisticsfor the behavioral sciences*. Tech. rep.

Smith, T, A Heger, and I Sudbery (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research* **27**(3), 491–499.

Steinhoff, C and M Vingron (2006). Normalization and quantification of differential expression in gene expression microarrays. *Briefings in bioinformatics* **7**(2), 166–177.

Stuart, T, A Butler, P Hoffman, C Hafemeister, E Papalexi, WM Mauck, M Stoeckius, P Smibert, and R Satija (2018). Comprehensive integration of single cell data. *bioRxiv*. eprint: https://www.biorxiv.org/content/early/2018/11/02/460147.full.pdf.

Student (1908). The probable error of a mean. *Biometrika*, 1–25.

Tang, M, J Sun, K Shimizu, and K Kadota (2015). Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. *BMC bioinformatics* **16**(1), 360.

Team, RC et al. (2013). R: A language and environment for statistical computing.

Wattenberg, M, F Viégas, and I Johnson (2016). How to use t-SNE effectively. *Distill* **1**(10), e2.

Wilcoxon, F (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1**(6), 80–83.

Wold, S, K Esbensen, and P Geladi (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems* **2**(1-3), 37–52.

Yip, SH, PC Sham, and J Wang (2018). Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Briefings in Bioinformatics*. eprint: http://oup.prod.sis.lan/bib/advance-article-pdf/doi/10.1093/bib/bby011/24122377/bby011.pdf.

Yip, SH, P Wang, JPA Kocher, PC Sham, and J Wang (2017). Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic acids research* **45**(22), e179–e179.