Single cell RNA-Sequencing and co-expression analysis for determination of highly variable genes and to identify senecence associated profiles

A thesis submitted for the degree of

Masters Degree

by

Bharvi Dhall

Maynooth University
National University of Ireland Maynooth

School of Mathematics and Statistics

Maynooth University

Ireland

August 2019

Contents

List of Figures

List of Tables

Contents

Preface

The abstract is a summary of the whole thesis. Typically this section would have about 250-350 words. Check the rules of your School or University.

Acknowledgement

The abstract is a summary of the whole thesis. Typically this section would have about 250-350 words. Check the rules of your School or University.

Abstract

The abstract is a summary of the whole thesis. Typically this section would have about 250-350 words. Check the rules of your School or University.

Table of Contents

List of figures

The abstract is a summary of the whole thesis. Typically this section would have about 250-350 words. Check the rules of your School or University.

List of tables

The abstract is a summary of the whole thesis. Typically this section would have about 250-350 words. Check the rules of your School or University.

Nomenclature

The abstract is a summary of the whole thesis. Typically this section would have about 250-350 words. Check the rules of your School or University.

Chapter 1

INTRODUCTION

1.1 Purpose and Motivation

Cell is the basic structural unit of life and 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. Therefore, it is important to study and analyse various aspects of human body to keep human population healthy and make it resistant to diseases.

In early centuries, the health care system was more of a sick-cure system where a person was cured once he got any illness. With 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.

This thesis involves the analysis of two RNA sequencing datasets namely liver and bone marrow. This project aims to reduce the number of genes with the help of various dimensionality reduction techniques, to find differentially expressed genes and thus investigating the genes that are associated with cellular senescence.The scope of the project also extends to perform gene co-expression analysis to find which genes are correlated with those differentially expressed genes. Differential gene expression analysis is also

carried out to test if there are any significant differences in gene expressions between different conditions.

****Describe one line about each chapter

1.2 Biological Background

We all know that 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 cell contains sub-cellular structures known as chromosomes which are responsible for transferring genetic material. Each chromosome has tightly packed DNA. Information stored in DNA is used to synthesize a protein which is responsible for functioning of a cell. Gene is a part of DNA on chromosome which is transcribed to RNA, which is then translated to amino acids which get folded to synthesise proteins. This process of synthesizing proteins from DNA is known as gene expression (Crick, 1970) .

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

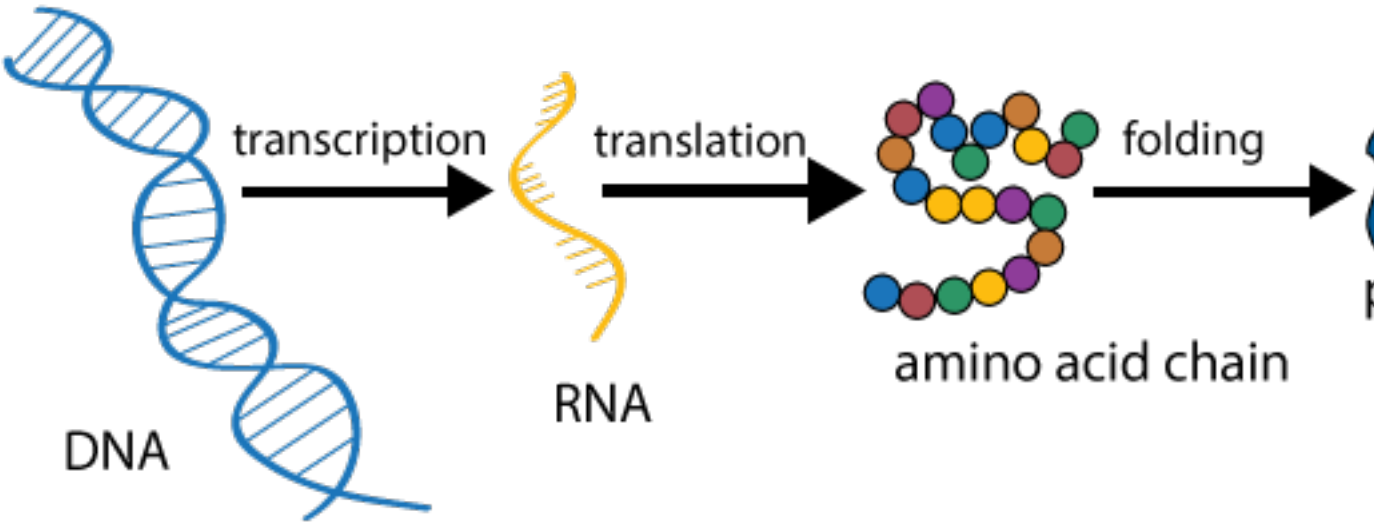

Figure 1.1: *Gene Expression-Conversion of DNA to Protein*

The knowledge of gene expression helps us to find and target various problems in human body as 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 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).

1.3 Dataset Description and availability

The datasets used in this project are generated using scRNA-Seq. ScRNA is a detailed study if geneexpression as it tries to capture the activity of thousands of genes in a single cell.It has provided a lot of useful insights in the field of cancer genomics(Hwang, Lee,

Table 1.1: *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

and Bang, 2018) and is a new domain with ongoing research.Sc RNA-seq is used to study functions of a cell and its heterogeneity (Papalexi and Satija, 2018).It treats each cell like an individual sample. A gene expression matrix obtained after sc RNA with genes as rows (features) and cells as columns (cases).Each row in matrix gives counts per sample for a specific gene. The reads matrix obtained after RNA-Seq are filtered and pre-processed (refer Chapter 3), and once the data is filtered the gene expression matrix is normalised to make data comparable and to prevent false biological conclusions (Steinhoff and Vingron, 2006).The two datasets in this project are liver and bone-marrow. The dataset provided for liver was already normalized whereas the dataset for bone-marrow data had raw-counts and was filtered and pre-processed for further analysis.

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 and columns represent cases. 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.

The table 1.1 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.

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

```
6 x 6 sparse Matrix of class "dgCMatrix"
              P1TLH_AAACCTGAGCAGCCTC_1 P1TLH_AAACCTGTCCTCATTA_1
RP11-34P13.7                         .                        .
FO538757.2                           .                        .
AP006222.2                           .                0.3147595
RP4-669L17.10                        .                        .
RP5-857K21.4                         .                        .
RP11-206L10.9                        .                        .
              P1TLH_AAACGGGAGTAGGCCA_1 P1TLH_AAACGGGGTTCGGGCT_1
RP11-34P13.7                         .                        .
FO538757.2                           .                        .
AP006222.2                           .                        .
RP4-669L17.10                        .                        .
RP5-857K21.4                         .                        .
RP11-206L10.9                        .                        .
[1] 20007  8444
```

Figure 1.2: *First five rows and columns of the liver dataset*

```
> show(dall)
An object of class Seurat
33694 features across 90653 samples within 1 assay
Active assay: RNA (33694 features)
> slotNames(dall)
 [1] "assays"       "meta.data"    "active.assay" "active.ident"
 [6] "neighbors"    "reductions"   "project.name" "misc"
[11] "commands"     "tools"
```

Figure 1.3: *Seurat object created from gene-expression matrix of bone-marrow dataset*

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 (single cell RNA- Sequencing) 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 created @ref{Outline of Bioconductor Ecosystem}. The dataset has 33694 genes and 90653 cells which will be pre-processed and filtered for analysis.

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

```
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
```

Figure 1.4: *Memory used in sparse and normal matrices*

1.3.1 Challenges with Datasets:

The scRNA data is recorded at a single cell resolution, thus the size of the count matrices may vary from The matrices used to store biological information are large and working with these datasets is computationally exhaustive. Therefore it is important to have 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. (?reference)

1.3.2 Sparse Matrices

The matrices obtained from biological data are usually big and have many zeros. Sparse matrices 5 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 [R] to understand the advantages of sparse matrices compared to dense matrices.

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

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 zer

object.size(m2)
# 5744 bytes
# size slightly increases as matrix saves the space by representi
```

Figure 1.5: *Storage space for Sparse Matrices increase as they store the non-zero elements*

If we add a single non-zero observation to the dataset we can clearly see that space in memory occupied by the sparse matrix will increase while the space occupied by the normal matrix will remain unchanged.

Both the datasets in this project have been stored as sparse matrix for the ease of calculations.

1.4 RNA-Sequencing Workflow

A typical RNA-Seq analysis consists of five steps. There are variety of softwares and environments that can be used in 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 [1] **??** describes the workflow commonly used for analysing RNA-Seq Data.

[1]This image has been downloaded from https://www.nature.com/articles/s41467-018-06318-7[@macparland2018single]

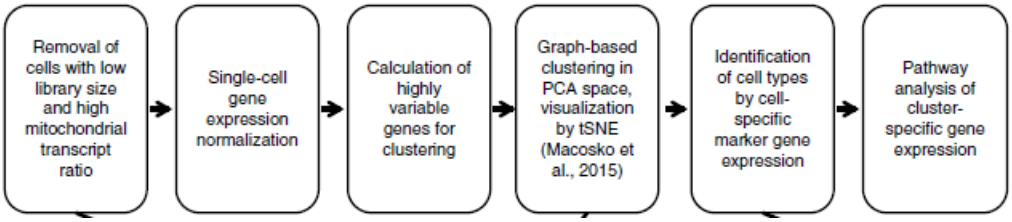

Figure 1.6: *RNA-Seq Workflow*

The raw counts in the dataset are filtered to remove low quality and dying cells. The resultant matrix is them normalized to make data comparable and to 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 (Jolliffe, 2011), t-SNE (Maaten and Hinton, 2008). Then clustering is done on the results of PCA and the resultant clustered are labelled (refer chapter 4). The differentially expressed genes also known are marker genes are then located (refer chapter 4).

Chapter 2

Outline of Bioconductor Ecosystem

Bioconductor is a collection of packages in R mainly used for genomics study and analysis.

2.1 Seurat object

The analysis of the two datasets is conducted with the help of Seurat Package version 3.0. Seurat is a toolkit for quality control, analysis, and exploration of single cell RNA sequencing data. 'Seurat' aims to enable users to identify and interpret sources of heterogeneity from single cell transcriptomic measurements, and to integrate diverse types of single cell data. Seurat is developed and maintained by the Satija lab (Stuart et al., 2018).

The Seurat package supports improved methods of normalization and removes any variations that occur due to technical faults10. 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.

Seurat package stores all information of dataset and the analysis results as a Seurat object.A Seurat object is created with the raw counts which contains various slots which will store not only the raw input data but also results from various computations. The function to create a Seurat object is:

CreateSeuratObject(counts, project = “SeuratProject”, assay = “RNA”)

Here, the counts refer the unnormalized data such as raw counts and the project sets name for the Seurat object and assay gives name of assay corresponding to input data, in our case RNA.

A Seurat object named pbmc and dall were created for liver and bonemarrow datasets respectivey. using package Seurat.

```{r}

pbmc

#To view slot names in the seurat object
slotNames(pbmc)


```

```
An object of class Seurat
20007 features across 8444 samples within 1 assay
Active assay: RNA (20007 features)
 2 dimensional reductions calculated: pca, tsne
 [1] "assays"       "meta.data"    "active.assay" "active.ident" "graphs"
 [6] "neighbors"    "reductions"   "project.name" "misc"         "version"
[11] "commands"     "tools"
```

Chapter 3

Literature Review

Chapter 4

METHODOLOGY

Analysis of scRNA datasets follows a similar pipeline for all datasets however the computational functions used may vary from one package to another. This section introduces the experimental design, describes pre-processing workflow to obtain suitable data for analysis and explains the various methods used for analysing the scRNA-seq datasets.

4.1 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 (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 removed from the dataset **??**.

4.1.1 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 normalised **??** to get normalised counts.(refer MacParland et al., 2018).

4.1.2 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 [1] (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. The plot **??** has been created to visualize QC metrics and feature-feature relationships to find the optimal cut-off value for filtering of cells.

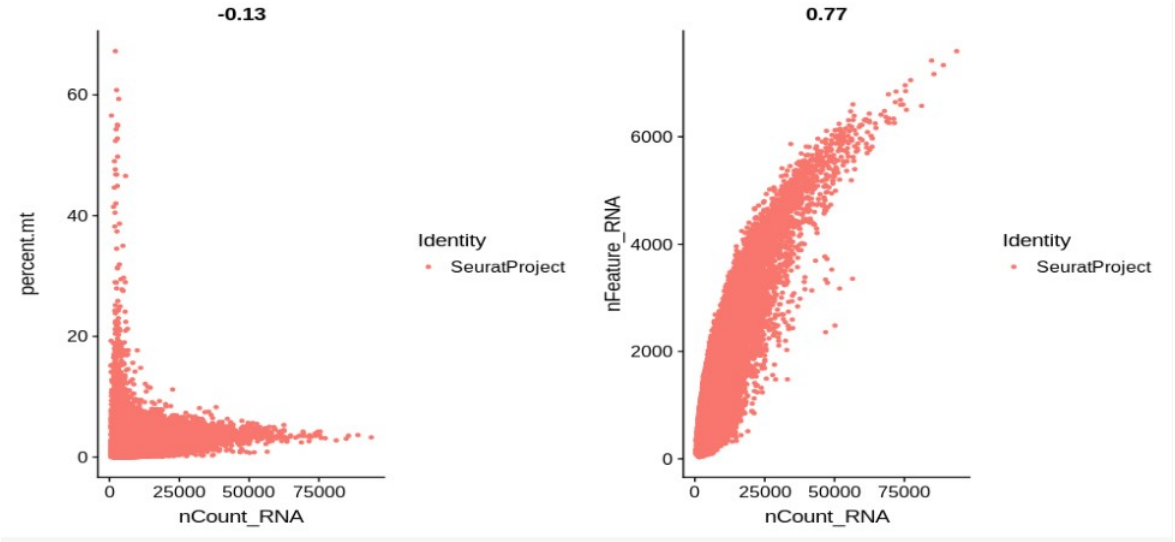

Figure 4.1: *Feature Scatter Plot*

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

Figure **??** represents 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. 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. Refer section **??**.

4.2 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() [2] in R(Team et al., 2013).

4.3 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.

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

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 [3] 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 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., 2019).By default this function selects 2000 genes but this number can be adjusted by the use of argument "nfeatures"[??].The results have been discussed in section **??**

4.4 Dimentionality Reduction

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 are more than the number of cases.Various Dimensionality Reduction techniques like Principle Component Analysis (PCA) (**wold1987principal**), t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008) and Uniform Manifold Approximation and Projection (UMAP) (Maaten and Hinton, 2008) were used to reduce the dimensions and project the data in lower dimensions.This section describes the methodology used for performing dimensionality reduction.

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

4.4.1 Principle Component Analysis

PCA is a linear feature extraction un-supervised learning technique widely for data with a high number of features (**wold1987principal**). 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. (**wold1987principal**).

The data selected containing the HVG is scaled prior to running PCA using ScaleData[4] 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 [5] function in Seurat v3 and the results are stored in the reductions slot of Seurat Object.Results of PCA are discussed in section **??**. The optimal number of Principal Components (PCs) are picked using the JackStraw[6] and Elbow\footnote{https://www.rdocumentation.org/packages/GMD/versions/0.3.3/topics/elbow 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 (**chung2018jackstraw**).

The JackStraw Plot function [7] was used to visualize JackStraw results. Figure **??** demonstrates 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.

[4] https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/ScaleData
[5] https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/RunPCA
[6] https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/JackStraw
[7] https://www.rdocumentation.org/packages/Seurat/versions/3.0.2/topics/JackStrawPlot

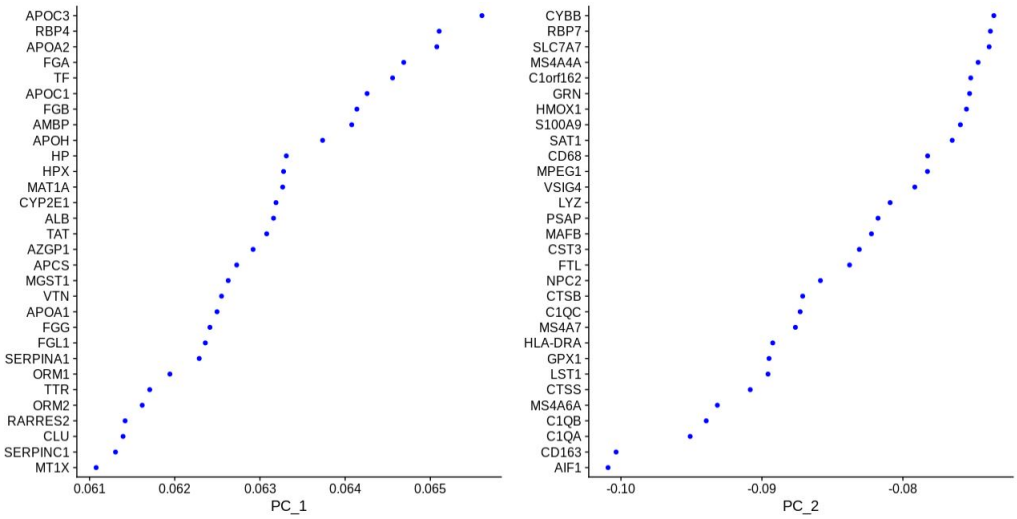

Figure 4.2: *JackStaw Plot*

4.4.2 t-SNE

4.5 Clustering of cells and visualization

4.6 Differential Expression Analysis

4.7 Gene-coexpression analysis

Chapter 5

RESULTS (a)

Liver Dataset:

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.Gene co-expression analysis was performed to find modules of senescent associated genes.

5.1 Highly Variable Genes

The FindVariableFeatures **??** function facilitated the selection of 7000 HVG from normalized data.The selected HVG were used for PCA.

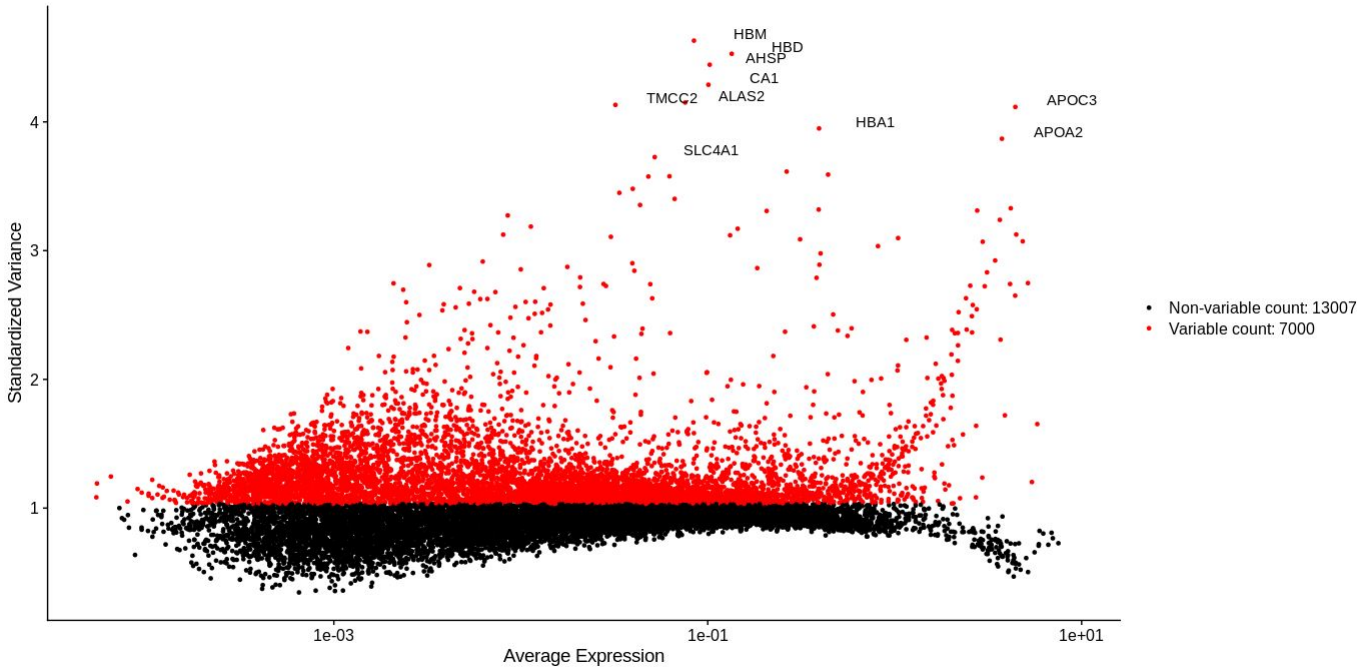

Figure 5.1: *Variable Genes Plot*

Figure **??** labels the top 10 HVG and highlights HVG and the genes which are non-variable.In the plot, X-axis function is the mean expression level, and for Y-axis it is the log(Variance/mean).

Note:All mean/variance calculations are not performed in log-space, but the results are reported in log-space.[1]

5.2 Principal Component Analysis:

After computation of HVG, the data was scaled The principal was performed on variable genes and 50 principal components (PC) were obtained.The first two principal components accounted for maximum variability in the data.

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

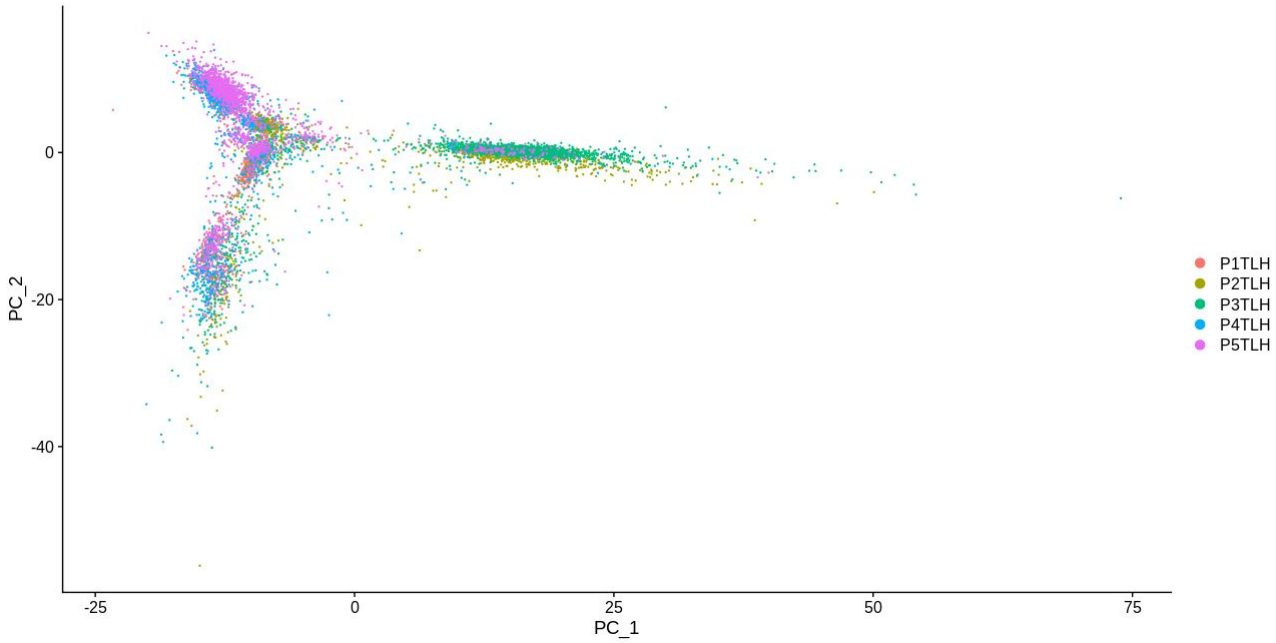

Figure 5.2: *Principal Component Plot of five samples*

Figure **??** plots the first two principal components of the samples.VizDimLoadings[2] function in Seurat v3 can be used to visualize top genes associated with principal components.

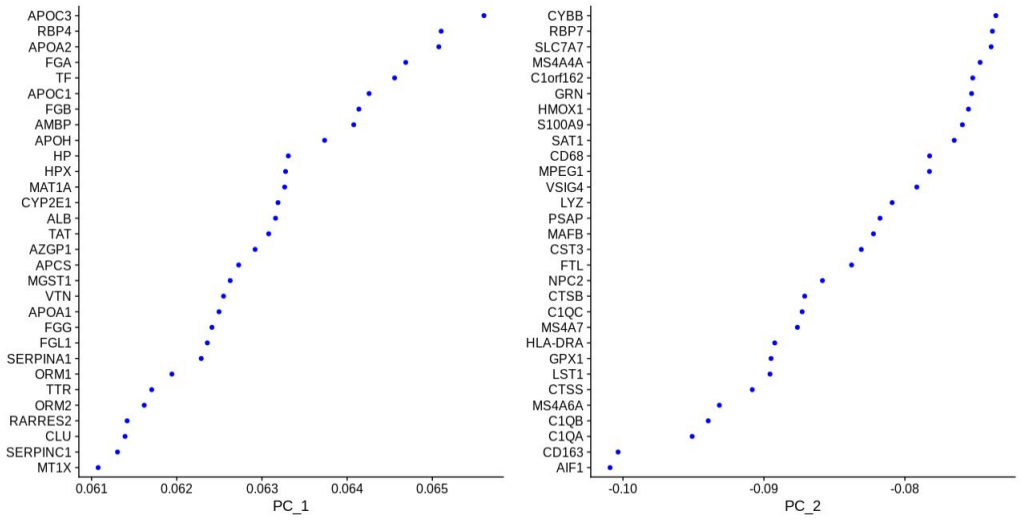

Figure 5.3: *Loadings of PC1 and PC2*

Figure **??** visualizes 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.

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

Chapter 6

RESULTS (b)

Bonemarrow 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.

Appendix A

Additional stuff

You might put some computer output here, or maybe additional tables.

Note that `\appendix` must appear before your first appendix although this is not required for other appendices.

Appendix B

Linear Mixed Model

The extension to the linear models is the linear mixed models. A general form for the linear mixed model is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \tag{B.1}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of observations on the trait (or response) of interest, $\boldsymbol{\tau}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{u}$ is a $q \times 1$ vector of random effects, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are associated (known) design matrices specifying the factors and covariates (explanatory variables) with corresponding fixed and random effects vector respectively, and $\boldsymbol{e}$ is the $n \times 1$ vector of errors. We assume that rank of $\boldsymbol{X}$ is $p_0 < p$ (i.e. non-full rank).

To complete the specification, we assume that the joint distribution of $(\boldsymbol{u}, \boldsymbol{e})$ is

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix} \right)$$

where $\boldsymbol{G}(\boldsymbol{\kappa}_G)$ and $\boldsymbol{R}(\boldsymbol{\kappa}_R)$ are covariance matrices which depend on vectors $\boldsymbol{\kappa}_G$ and $\boldsymbol{\kappa}_R$ respectively; and θ is the global scale parameter. We let $\boldsymbol{\kappa} = (\boldsymbol{\kappa}_G{}^\top, \boldsymbol{\kappa}_R{}^\top)^\top$ denote the complete vectors of variance parameters.

It follows that $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\tau}, \boldsymbol{V})$ where $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top + \boldsymbol{R}$ where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\kappa})$.

Bibliography

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.

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.

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.

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.

Jolliffe, I (2011). *Principal component analysis*. Springer.

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

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.

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.

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.

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.

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

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.

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, WMM III, M Stoeckius, P Smibert, and R Satija (2018). Comprehensive integration of single cell data. *bioRxiv*.

Stuart, T, A Butler, P Hoffman, C Hafemeister, E Papalexi, WM Mauck III, Y Hao, M Stoeckius, P Smibert, and R Satija (2019). Comprehensive Integration of Single-Cell Data. *Cell*.

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

Yip, SH, PC Sham, and J Wang (2018). Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform*, bby011.

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.