Statistical Analysis of Single Cell Data

James O'Reilly
Student Number: r0773125

Problem definition: Perform data preprocessing, dimensionality reduction, clustering, latent variable modelling, and factor analysis to understand and disentangle sources of variation in single cell transcriptome data.

1 Motivation

I chose to focus on statistical analysis of single cell data for a number of reasons. Firstly, and most importantly, I find studying cellular populations fascinating and understanding sources of cellular heterogeneity is an active area of research. Secondly, for my thesis next term I will be focusing on sources of transcriptional heterogeneity in cancer. I will use factor analysis to decompose deterministic and stochastic sources of transcriptional heterogeneity in lung adenocarcinomas. Understanding how clustering and factorial latent variable models can be used to determine sources of variability will be useful for this. Lastly, I wanted to understand and present the theory behind factorial single cell latent variable modelling, as I find it interesting.

2 Data, Software, and Packages

This assignment uses a publicly available single-cell transcriptome sequencing dataset provided by 10X-Genomics [1]. The dataset contains single-cell gene expression data from approximately 3000 peripheral blood mononuclear cells (PBMC) from a healthy human donor. Pre-processing, dimensionality reduction, and clustering were performed using the Seurat package [2, 3]. Single cell latent variable modelling was performed using the slalom package [4]. Gene sets used in the factor analysis were obtained from the molecular signatures database MSigDB and the REACTOME pathway database [5, 6].

3 Introduction and Background

The advent of single-cell sequencing technologies has enabled the study of cellular populations at a resolution which was previously not possible [7]. One characteristic of cellular populations which is of particular interest to researchers is cellular heterogeneity. Understanding the sources of heterogeneity in a cellular population is relevant for diseases such as cancer, where intra-tumoural heterogeneity is one of the major mechanisms through which cancers form resistance to treatment [8]. Cellular heterogeneity is usually quantified at the level of the transcriptome using single-cell RNA-sequencing (scRNA-seq). Such transcriptional heterogeneity can arise due to both technical and biological factors, which makes decomposing sources of variation in cellular populations difficult [9, 10]. Importantly, such technical and biological factors can act upon the same genes, meaning that they need to be modelled jointly to fully understand heterogeneity in scRNA-seq data.

Unobserved factors that cause variation in scRNA-seq data include cell-cycle state and the number of detected genes in a cell. These factors often dominate or mask other biological sources of variation. A number of tools have recently been developed to account for such factors [11, 2]. However, these methods *independently* fit individual processes (such as cell cycle) and do not test for the presence of additional unannotated biological factors or confounding sources of variation. The factorial single-cell latent variable modelling (f-scLVM) approach used in this coursework *jointly* infers factors that capture multiple sources of variation, including:

1. Variation in expression due to pre-annotated sets of genes
2. Variation due to other sparse factors (relating to putatively meaningful biological effects not captured by variation in pre-annotated gene sets)
3. Confounding (dense) factors which affect expression across the majority of genes

For a given factor, if the factor explains variation in the data, then it is assumed the expression levels of all genes assigned to that factor co-vary in a reliable manner. This assumption is key to the model as it allows the activity of each factor to be inferred from the data. The annotated biological factors are created using pre-annotated gene sets from public gene set databases (MSigDB and REACTOME), which assign biologically related genes to the same factor. For the unannotated factors which model potentially meaningful biological effects, a generic sparsity is assumed so that these factors are responsible for the variation in a small number of genes. Lastly, dense factors are used to capture the confounding effects which effect the expression of large numbers of genes. After defining the factors, the model then infers which factors explain variability in the given dataset. The relevance of each factor is then inferred by calculating the expected variance in expression levels across cells using genes assigned to the factor.

3.1 The Factorial Single-Cell Latent Variable Model in Detail

In effect, f-scLVM decomposes the gene-cell matrix into a sum of contributions of C measured covariates, A annotated factors, and H unannotated factors. The gene expression matrix $\mathbf{Y}$ with cell rows N and gene columns M is then given by

$$\mathrm{Y} = \underbrace{\sum_{\mathrm{c}=1}^{\mathrm{C}} \mathbf{u}_{\mathrm{c}} \mathbf{V}_{\mathrm{c}}^{\boldsymbol{T}}}_{\text{cell covariates}} + \underbrace{\sum_{\mathrm{a}=1}^{\mathrm{A}} \mathbf{p}_{\mathrm{a}} \mathbf{R}_{\mathrm{a}}^{\boldsymbol{T}}}_{\text{annotated factors}} + \underbrace{\sum_{\mathrm{h}=1}^{\mathrm{H}} \mathbf{s}_{\mathrm{h}} \mathbf{Q}_{\mathrm{h}}^{\boldsymbol{T}}}_{\text{unannotated factors}} + \boldsymbol{\Psi} \tag{1}$$

The vector μ_c represents the known cell covariates. $\mathbf{p}_a$ and $\mathbf{s}_h$ correspond to cell states for the annotated and unannotated factors, respectively. $\mathbf{V}_c$, $\mathbf{R}_a$, and $\mathbf{Q}_h$ are the regulatory weights for a given factor (cell covariates, annotated factors, unannotated factors) on all genes. The matrix $\boldsymbol{\Psi}$ represents the residual noise, which is modelled in this assignment as a log-normal noise model. This means our expression matrix consists of log count values which we assume are modelled by independently and identically distributed heteroscedastic residuals.

Without going into too much detail, the f-scLVM method first specifies priors for the annotated and unannotated factors, and then iteratively updates the posterior estimate for these factors using variational Bayes. The fitted f-scLVM model can then used to determine factor relevance and to visualise the posterior distribution over inferred factors.

4 Data Preprocessing

The data is stored in an $M \times N$ matrix where M is the number of genes and N is the number of cells. For each cell we have the expression of each gene. It should be noted that this matrix is rather sparse, as in most cases the expression for a given gene in a given cell is zero. This sparsity can lead to misleading results when using some standard distance metrics, so it's good to be mindful of this. Before normalising and scaling the expression data, low quality cells are removed from the dataset. Cells with a high proportion of mitochondrial DNA are likely dying and therefore removed, along with cells which exhibit abherrantly high or low gene counts. The data was normalised using a global-scaling normalisation method that first normalises the gene expression measurements for each cell by the total expression, before multiplying this by a scale factor, and log-transforming the result.

For downstream analysis of single-cell expression data, it is best to only take a subset of genes which exhibit high cell-to-cell variation. Low levels of cell-to-cell variation are more likely the result of technical noise, and could negatively impact the analysis [9]. As selecting genes based solely on log-normalised single cell variance does not account for the mean-variance relationship, a variance stabilising transformation is first applied to correct for this. The variance of these standardised expression values are then computed for each gene, giving a measure of variance after controlling for mean expression. The genes with the highest standardised variance are then selected. For this analysis the top 2,000 most variable genes were used as input for downstream analysis. Lastly, it is necessary to scale the data before performing dimensionality reduction, as highly-expressed genes would otherwise dominate. The expression for each gene is scaled so that the mean expression across cells is zero and the variance across cells is one.

5 Analysis

5.1 Dimensionality Reduction

Principle component analysis was performed on the pre-processed data, using only the most variable genes as input. To analyse the distribution of variance in this dataset, the genes with the highest loadings for each principal component can be visualised (see Figure 1). We can see that the first PC has a large negative loading for the MALAT1 gene, and positive loadings otherwise. The second PC is split more evenly, with a number of genes with positive and negative loadings. These genes can be used in downstream analyses to cluster the cells and determine cell types. A scatter plot for these two principal components was then plotted, highlighting some structure present in the cellular population (see Figure 1). A number of heatmaps were then plotted, focusing on each principal component (see Figures 2). Within these heatmaps, both cells and genes are sorted according to their principle component loadings. This plot is particularly useful as it clearly shows the sources of heterogeneity in the dataset.

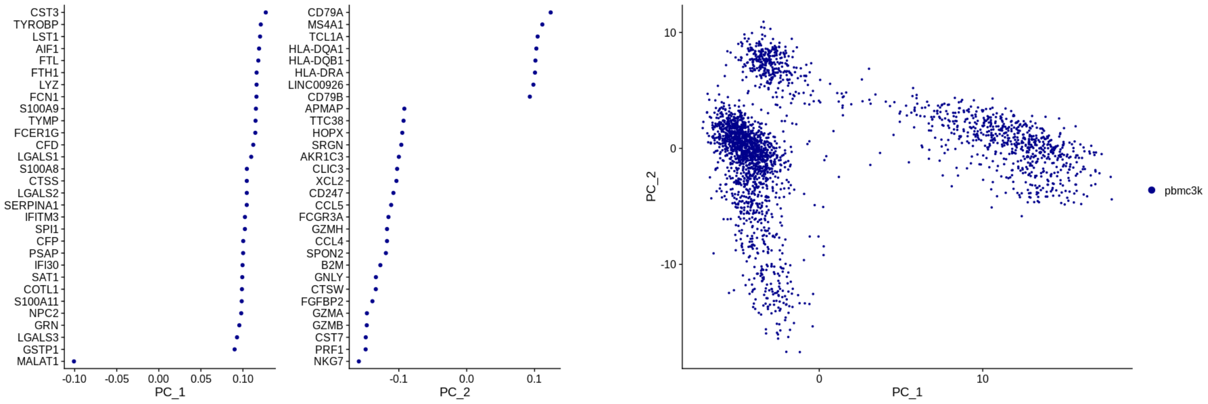

Figure 1: (left) Top loadings for the first two principle components. (right) Scatter plot for the first two principal components. There is already some population structure present with these two PCs, with some clusters of cells.

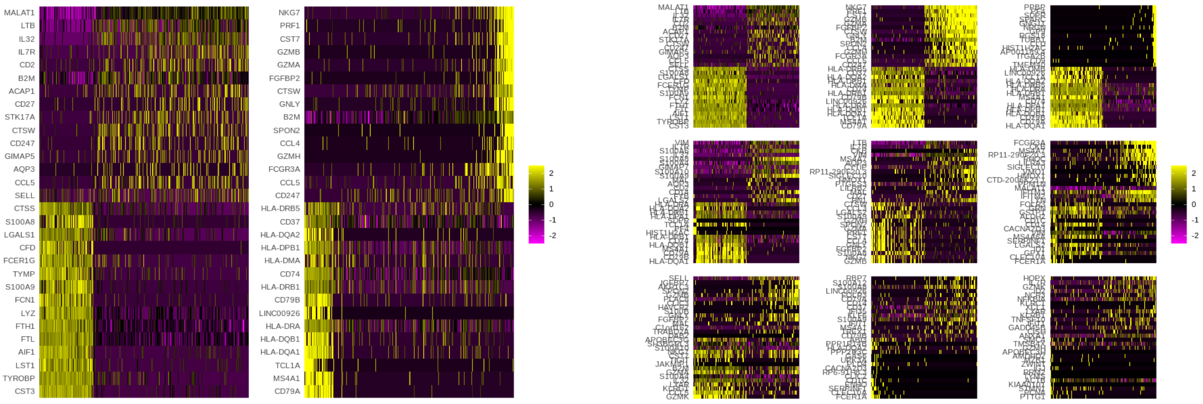

Figure 2: (left) Heatmaps for the first two principal components and the first 500 cells. The rows contain genes and the columns contain cells. The colour gradient represents log-fold expression change. (right) Heatmaps for the first 9 principal components, we can clearly see a drop-off in the variability in gene expression across cells after the first few principal components.

For clustering, the cells are clustered based on their PCA scores. Using only the top PCs effectively give a low-dimensional representation which mitigates the technical noise associated with gene expression data. We therefore need to decide how many principal components we will use for clustering and other downstream analyses. One way of doing this is to estimate the PCs with the most variance from the heatmaps given in Figure 2. A more statistically robust method for determining the number of PCs is to use either an Elbow plot or the Jackstraw procedure. The elbow plot is a simple heuristic which ranks the PCs based on the percentage of variance explained. We can then observe how many PCs are needed to capture the majority of the variance in this dataset. Figure 3 indicates that a cutoff at PC9-10 or PC15 would be good choices, depending on the resolution required for downstream analyses.

The Jackstraw method is statistically robust but also computationally intensive. A null distibution of feature scores is generated by iteratively permuting a random subset of the data before running principle components analysis. The statistical significance of each principle component can then be calculated against this null distribution. Seurat provides a visualisation tool which compares the distribution of p-values for each principal component (see Figure 3). Significant PCs are those which a have a large number of genes with low p-values, which is given by a solid curve above the dashed line, where the dashed line represents the uniform distribution for p-values. Figure 3 indicates that the significance of PCs decreases after PC11.

5.2 Clustering

Seurat uses a graph-based clustering approach. A detailed description of this method is given in Levine et al [12]. Briefly, a K-nearest neighbours (KNN) graph is first generated in PCA space using a Euclidean distance

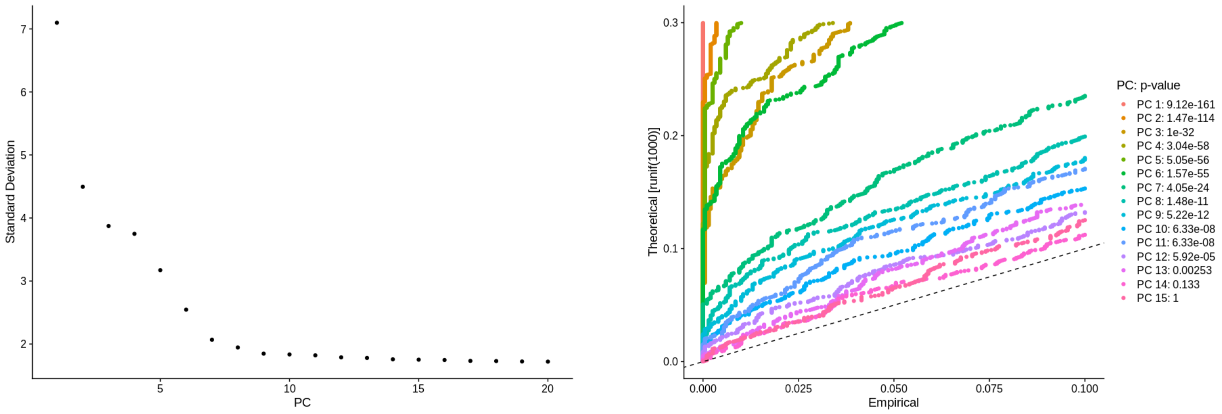

Figure 3: (left) Elbow plot for the top 20 principal components. (right) Jackstraw plot comparing the distribution of p-values for each principal component agains the uniform distribution (dashed line).

metric, for the given number of significant PCs previously determined. For any given two cells, the weight of the edge between those cells in the graph is assigned based on the Jaccard similarity between their neighbourhoods. The cells are then clustered by maximising the modularity of the this KNN graph, where modularity is defined as the strength of division of a network into clusters or communities. This is implemented in Seurat using modularity optimisation algorithms such as the Louvain algorithm or SLM [13, 14].

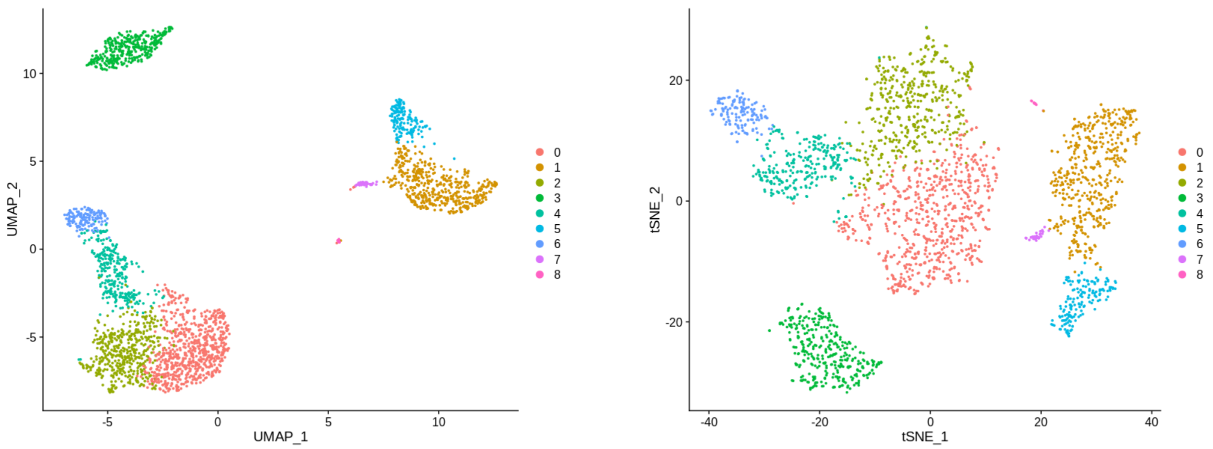

Figure 4: (left) UMAP. (right) t-SNE.

5.3 Non-Linear Dimensionality Reduction

Each cell was then labelled with a cluster, although these clusters do not yet have any identity. We would like to visualise these clusters, however as they exist in high-dimensional space, this is not usually possible. To visualise the clusters, we can use non-linear dimensionality reduction techniques such as UMAP or t-SNE. Both methods are known to preserve the local-structure of the data well (within-cluster distance), however UMAP preserves global structure (between cluster distances) better than t-SNE. A beautifully simple and well written artcle explaining the intuition and mathematics behind this is given here. We can use the sample principal components as input to the UMAP and t-SNE. The results are given in Figure 4. To complete this part of the analysis, we can analyse differentially expressed genes in each cluster (called cluster biomarkers) and assign each cluster to a cell-type based on known markers. Cell-type annotation is primarily bioinformatics question and not as relevant to the course, so I won't discuss it here, in order to save space.

5.4 Factorial Single Cell Latent Variable modelling

The slalom R package was used to train a factorial single cell latent variable model on the scaled, normalised log-counts of the most variable genes in the PBMC dataset. The MSigDB hallmark gene set was used for this analysis [15]. It contains 50 gene sets which represent specific well-defined biological states or processes. The relevance of each factors included in the model can then be plotted, identifying important pathways or processes that contribute to cellular variability. We can choose whether or not to include unannotated (hidden) factors. Figure 5 shows the most relevent annotated and unannotated factors.

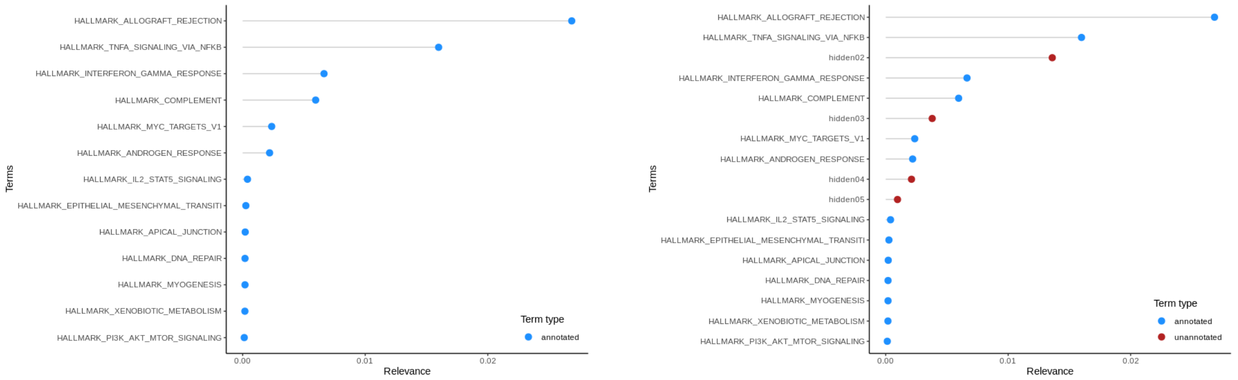

Figure 5: (left) Relevance of annotated factors. (right) Relevence of annotated and hidden factors.

The loadings of specific features (genes) on these factors can then be viewed and compared with knowledge in literature and gene databases to attempt to understand the underlying source of this variation (see Figure 6). For example, when working with cancer data, it is useful to determine how epithelial-mesenchymal transition (EMT) contributes to cellular variability, as EMT is one of the main drivers of metastasis. Viewing the loadings of the most relevant dense unannotated factor showed that ribosomal protein genes (RSP) had the highest loading, indicating that the variability in ribosomal genes was not captured by the MSigDB hallmark gene set and a more robust gene set such as the REACTOME pathways geneset might be more approriate.

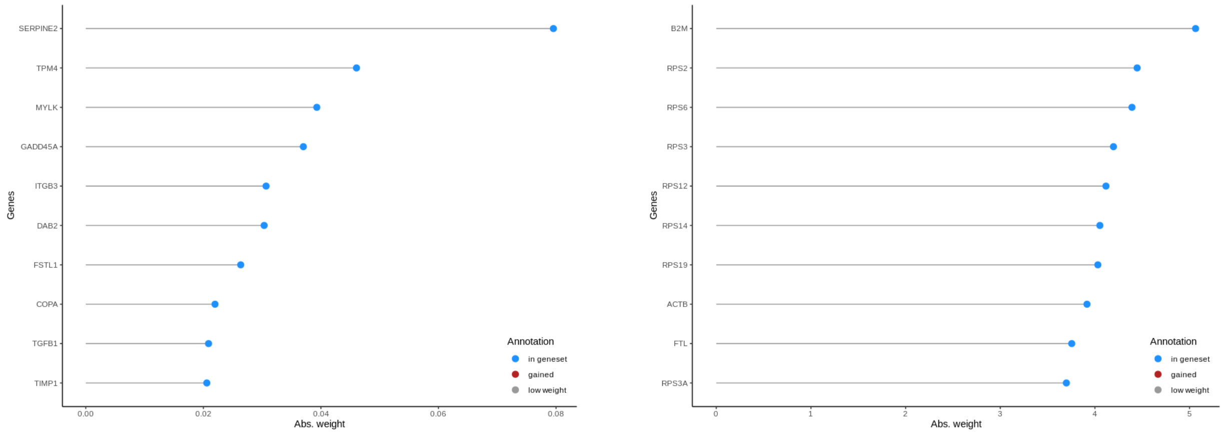

Figure 6: (left) Gene loadings for annotated EMT factor. (right) Gene loadings for the most relevant unannotated factor.

6 Discussion

With this assignment I aimed to demonstrate how data preprocessing, dimensionality reduction, clustering, and factor analysis can be applied to single cell datasets. I had hoped that the factor analysis would reveal that genes related to cell cycle contributed to the cellular variance, but surprisingly this was not the case. In the end, I couldn't make meaningful inferences from the final results, but I hope to have shown how this analysis could be used to generate meaningful insight into sources variance in single cell data.

Note: I had originally intended to do this analysis modelled on a Seurate vignette with a 9K mouse brain dataset, provided by 10X-genomics. However, after using this dataset for the dimensionality reduction and clustering part of this assignment, I found that my computer couldn't train the f-scLVM model for a dataset this large. In the end I had to use a smaller 3K PBMC dataset used in the Seurat vignette, and so my dimensionality reduction and clustering results are the same as those given in the vignette. I understand if I lose marks because part of my code is the same as code presented online, this was never my intention.

References

[1] 10X Genomics public datasets. https://support.10xgenomics.com/single-cell-gene-expression/datasets. Accessed: 2021-01-21.

[2] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[3] Seurat public vignettes. https://satijalab.org/seurat/vignettes.html. Accessed: 2021-01-23.

[4] Florian Buettner, Naruemon Pratanwanich, Davis J McCarthy, John C Marioni, and Oliver Stegle. f-sclvm: scalable and versatile factor analysis for single-cell rna-seq. *Genome biology*, 18(1):212, 2017.

[5] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[6] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2018.

[7] Xiaoning Tang, Yongmei Huang, Jinli Lei, Hui Luo, and Xiao Zhu. The single-cell sequencing: new developments and medical applications. *Cell & Bioscience*, 9(1):53, 2019.

[8] Nicholas A Saunders, Fiona Simpson, Erik W Thompson, Michelle M Hill, Liliana Endo-Munoz, Graham Leggatt, Rodney F Minchin, and Alexander Guminski. Role of intratumoural heterogeneity in cancer drug resistance: molecular and clinical perspectives. *EMBO molecular medicine*, 4(8):675–684, 2012.

[9] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*, 10(11):1093–1095, 2013.

[10] Stephen Smith and Ramon Grima. Single-cell variability in multicellular organisms. *Nature communications*, 9(1):1–8, 2018.

[11] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):1–13, 2015.

[12] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

[13] Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[14] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[15] Molecular Signalling Database. http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp. Accessed: 2021-01-23.

A Pre-processing, Clustering and Dimensionality Reduction

```r
# loading libraries
library(dplyr)
library(Seurat)
library(patchwork)

# Load the data
pbmc.data <- Read10X(data.dir = "/home/james/Documents/leuven/year-2/AMSA/Single-Cell-Transcriptome-Analysis/data/10X-genomics/PBMC_3K")

# Initialize the Seurat object
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k", min.cells = 3, min.features = 200)

# Quality Control and cell selection: filtering low-quality cells
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt < 5)

# Normalisation
pbmc <- NormalizeData(pbmc)

# Identification of highly-variable genes
pbmc <- FindVariableFeatures(pbmc, selection.method = "vst", nfeatures = 2000)

# Data Scaling
all.genes.pbmc <- rownames(pbmc)
pbmc <- ScaleData(pbmc, features = all.genes.pbmc)

# Dimensionality Reduction
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))

# Examining the PCA results
VizDimLoadings(pbmc, dims = 1:2, reduction = "pca", combine = TRUE, col = "darkblue")
DimPlot(pbmc, reduction = "pca", label = FALSE, cols = c("darkblue"))

DimHeatmap(pbmc, dims = 1:2, cells = NULL, balanced = TRUE, fast = FALSE, slot = "scale.data")
DimHeatmap(pbmc, dims = 1:9, cells = 500, balanced = TRUE, fast = FALSE, slot = "scale.data")

pbmc <- JackStraw(pbmc, num.replicate = 100)
pbmc <- ScoreJackStraw(pbmc, dims = 1:20)
JackStrawPlot(pbmc, dims = 1:15)

ElbowPlot(pbmc)

# Clustering
pbmc <- FindNeighbors(pbmc, dims = 1:10)
pbmc <- FindClusters(pbmc, resolution = 0.5)
head(Idents(pbmc), 5)

pbmc <- RunUMAP(pbmc, dims = 1:10)
DimPlot(pbmc, reduction = "umap")

pbmc <- RunTSNE(pbmc, dims = 1:10)
DimPlot(pbmc, reduction = "tsne")

# Save Seurat object
saveRDS(pbmc, file = "/home/james/Documents/leuven/year-2/AMSA/Single-Cell-Transcriptome-Analysis/output/pbmc.rds")
```

B Factorial Single Cell Latent Variable Modelling with slalom

```r
library(slalom)

# Convert Seurat object to SingleCellExperiment object
pbmc.rds <- readRDS(file = "/home/james/Documents/leuven/year-2/AMSA/Single-Cell-Transcriptome-Analysis/output/seurat_object.rds")
pbmc.sce <- as.SingleCellExperiment(pbmc.rds)

# Format data correctly for slalom
logcounts.matrix <- as.matrix(SingleCellExperiment::logcounts(pbmc.sce))
pbmc <- SingleCellExperiment::SingleCellExperiment(
  assays = list(logcounts = logcounts.matrix)
)

# Load geneset
gmtfile <- "/home/james/Documents/leuven/year-2/AMSA/Single-Cell-Transcriptome-Analysis/data/genesets/MSigDB_hallmark.gmt"
genesets <- GSEABase::getGmt(gmtfile)

# Creating the model with a set number of hidden factors and minimum number of genes per gene set
model <- newSlalomModel(pbmc, genesets, n_hidden = 5, min_genes = 10)

# Initialising the model and set seed for reproducible analysis
model <- initSlalom(model, seed = 99)

# Training the model
trained.model <- trainSlalom(model, minIterations = 400, nIterations = 2000, shuffle = TRUE, seed = 99)

# View most relevant terms and their respective gene set sizes
topTerms(model)
plotRelevance(model, mad_filter = 0.1, unannotated_dense = TRUE, unannotated_sparse = FALSE)

# View most relevant annotated and unannotated factors
plotTerms(model, mad_filter = 0.1, unannotated_dense = FALSE, unannotated_sparse = FALSE)
plotTerms(model, mad_filter = 0.1, unannotated_dense = TRUE, unannotated_sparse = FALSE)

# View loadings for specific factors
plotLoadings(model, "hidden01")
plotLoadings(model, "HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION")
```