SINGLE CELL TRANSCRIPTOME ANALYSIS REVEALS SIMILARITIES AND DIFFERENCES IN GENE EXPRESSION OF ADULT AND EMBRYONIC NEURAL STEM CELLS

By

NIRALI PATEL

A thesis submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Biomedical Engineering

Written under the direction of

Li Cai, Ph.D.

And approved by

New Brunswick, New Jersey

May, 2018

ABSTRACT OF THE THESIS

SINGLE CELL TRANSCRIPTOME ANALYSIS REVEALS SIMILARITIES AND DIFFERENCES IN GENE EXPRESSION OF ADULT AND EMBRYONIC NEURAL STEM CELLS

by **NIRALI PATEL**

Thesis Director: Li Cai, Ph.D.

Adult and embryonic stem cells both harbor advantages and disadvantages for use in cell therapy. Embryonic stem cells are pluripotent and have a greater potential for differentiation than adult stem cells. Adult stem cells, although considered less plastic, have less risk of immune rejection. However, specific differences between adult and embryonic stem cells are not clear. Using single cell transcriptome analysis, adult and embryonic neural stem cells (NSCs) were examined for differences in gene expression patterns, the top 60 highest expressed genes, and cell homogeneity. When examining the top 60 expressed genes, only 19 genes were similar in terms of expression levels amongst adult and embryonic NSCs, indicating more differences in the genes that were highest expressed than similarities. In both adult and embryonic NSCs, genes encoding for cell growth and differentiation, neurogenesis, and tumor suppression were present, however, genes for each function were expressed at different levels within the two cells types. Within adult NSCs, Meg3 was highly expressed for tumor suppression, however in embryonic NSCs, the Sparc gene was highly expressed for the same function. In terms of genes coding for neurogenesis, adult NSCs expressed Ptprs and Gpm6a while embryonic NSCs highly expressed Npm1, Tubb3, Enc1, and Sox11. Differences in genes coding for the same

function such as differentiation can potentially lead to differences in differentiation efficiency, or the time it takes for cells to differentiate. By clustering cells into different groups, differences in gene expression patterns were observed. Embryonic NSCS failed to cluster with adult NSCs in two out of three studies implying differences in gene expression patterns across the two cell types. Upregulated genes in adult NSCs were downregulated in embryonic NSCs, e.g., Kctd16, Kcnh 3, Kcnh1 and Rab26 were amongst genes that were upregulated for most adult NSCs, however they were downregulated across all embryonic NSCs. Kcnh1, in particular, is a gene specific to the brain and regulates myoblast differentiation, neurotransmitter release, and neuronal excitability. Overexpression of this gene may be important in embryogenesis or lead to cancer cell formation, however, this gene was found to be downregulated amongst all embryonic NSCs, while being upregulated in many adult NSCs. It is possible that this gene is not upregulated in embryonic NSCs as they are considered more pluripotent than adult stem cells. Genes that were distinct to only adult NSCs function in ionotropic glutamate receptor signaling pathway, excitatory postsynaptic potential, central nervous system development, and associative learning such as: memory, cognition, and behavior. The highest regulated genes in embryonic stem cells function, instead, in telomerase holoenzyme complex assembly and the regulation of cell size. Genes that were common to both adult and embryonic NSCs were mainly involved in protein folding and cell cycle regulation. Together, single cell transcriptome analysis reveals that differences in gene expression patterns amongst adult and embryonic NSCs are evident and these molecular differences are the basis for the different properties of the two types of NSCs.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1: THE IMPORTANCE OF STEM CELLS IN MEDICINE:

Stem cells are powerful tools in medicine and have the potential to revolutionize regenerative medicine. (Bongso and Richards 2004) Their ability to develop into various different cell types and their plasticity within the body has led to groundbreaking discoveries such as functioning as repair systems to old or damaged cells and regenerating tissues within the body. Stem cells can differentiate to reproduce cells identical to themselves or into cells with specific functions such as neurons or cardiac cells. Their ability to take different paths of differentiation is what makes them such a powerful tool in medicine.

Since stem cells are able to regenerate tissues and even entire organs, scientists and medical professionals are examining their abilities to treat diseases such as heart disease, Parkinson’s, and diabetes. Research is being done to inspect properties of cells, such as their gene expression, in order to understand what distinguishes one cell type from another. Such understanding has allowed researchers to develop models of biological systems and screen the effectiveness of drugs on any defects that may be present. Although under study for many years, much work remains to effectively use stem cells without side effects.

There are two types of stem cells under detailed investigation: adult stem cells embryonic stem cells. Alongside targeted therapy, where stem cells, such as embryonic stem cells, are cultured ex vivo and administered to the area of injury or damage, some are already present in the body and can be signaled to become specialized cells for tissues or organs. For example, there is a population of quiescent neural stem cells in the brain, which upon injury, can be activated to specialize into immature and mature neurons. (Shin, Berg et al. 2015) These cells will be examined in greater detail throughout this study.

1.2: ADULT STEM CELLS

Adult stem cells are already present in the body and represent the population of undifferentiated cells that are found within specific tissues and organs. Adult stem cell can be activated to differentiate into the same cell type as those present in the tissue or organ in which they reside.

1.2.1: Adult Stem Cells - Background

The purpose of adult stem cells in the body is to replace dead or damaged cells to regenerate tissues and organs. The exact origin of adult stem cells is still under study, however, they are known to be present in different tissues and organs, such as the brain and heart. Adult stem cells are under investigation for transplants, such as bone marrow transplant. Within the brain, adult neural stem cells can differentiate into neurons, astrocytes and oligodendrocytes. This study focused on neural stem cells.

Adult stem cells have the ability to remain in the inactive, or quiescent, stage where they do not divide until stimulated. The quiescent cells are signaled by the body, perhaps through disease or injury in the tissue or organ, to be activated and begin to divide and replenish the old or damaged cells. For example, during traumatic brain injury, quiescent neural stem cells can be activated to divide and replace the damaged cells.

1.2.2: Adult Neural Stem Cells Examined in This Study

To study gene expression in adult cells, adult neural stem cells from a transgenic mouse line consisting of a population of various precursor cells were examined. (Shin, Berg et al. 2015, Habib, Li et al. 2016) Precursor cells were present in different developmental stages consisting of: quiescent neural stem cells, induced progenitor cells, and immature neurons. (Shin, Berg et al. 2015) The accession numbers for the studies were *GSE71485* and *GSE84371*. The gene expression for each cell was analyzed and compared against the embryonic neural stem cell dataset.

1.3: EMBRYONIC NEURAL STEM CELLS

Embryonic neural stem cells represent the population of stem cells that are derived from embryos. Generally, eggs undergo in vitro fertilization and stem cells are derived from resulting embryos. (Bongso and Richards 2004)

1.3.1: Embryonic Neural Stem Cells - Background

Embryonic neural stem cells are harvested from cells in an embryo in the preimplantation stage into a culture medium in a cell culture dish where they continuously divide and disperse across the dish surface. If cells combine to form an embryoid body, they have the ability to spontaneously differentiate. Embryoid bodies can form muscle cells, nerve cells, and a diverse set of other cell types as well. Cells can be modified by altering the gene expression to differentiate the cell into a more specific cell type. When genes are altered, these cells have the potential to replenish cells that have been damaged in a specific tissue or organ.

1.3.2: Embryonic Neural Stem Cells Examined in This Study

The embryonic neural stem cell datasets in this study consisted of cells present in two stages: neural progenitor cell and immature neurons. The accession numbers comprising the embryonic neural stem cell dataset were *GSE94579* and *GSE30765*. (Ayoub, Oh et al. 2011, Chen, Friedman et al. 2017) The gene expression in these cells was examined for comparison against the adult neural stem cell dataset to mark key differences in gene expression that can lead to variability in the efficacy within the cell types.

1.4: Key Similarities and Differences in the Properties of Adult and Embryonic Stem Cells

There are both advantages and disadvantages in the usage of adult and embryonic neural stem cells, and it is important to study gene expression to highlight some key similarities and differences between the two cell types. For example, embryonic stem cells are able to be cultured with relative ease compared to adult stem cells. Adult stem cells are not numerous within mature tissues and thus isolating and growing them is a challenge. Large numbers of cells are needed for successful cell based therapies which represents a disadvantage in adult neural stem cells. Embryonic neural stem cells on the other hand, are derived from embryos and transplanted into another organism. This process can lead to immune rejection, whereas adult neural stem cells are housed within the same organism thus avoiding the possibility of immune rejection. Since adult neural stem cells have a much lower rate of immune rejection, it is important to study the differences between adult and embryonic neural stem cells in order to identify causes of variability in cell processes between embryonic stem cells for implementation to adult stem cells.

Gene expression is a key determinant in cell structure and function. Different varieties of cells express different genes depending on their ultimate purpose. Different gene expression patterns lead to different behaviors and thus, the aim of this study is to identify whether there is an underlying difference in gene expression pattern that is causing advantageous of disadvantageous properties in adult and embryonic neural stem cells.

In particular, the embryonic and adult neural stem cells datasets were chosen because although gene expression has been analyzed in adult neural stem cells and embryonic neural stem cells, a study comparing the gene expression between the two cell types was not conducted. This study addressed specific differences in gene expressions based on the chosen datasets.

CHAPTER 2: EXPERIMENTAL/COMPUTATION PROCEDURES

2.1: SINGLE CELL TRANSCRIPTOME ANALYSIS

Similarities and differences in gene expression were analyzed in this study through single cell RNA sequencing. Single cell RNA sequencing has the ability to sequence thousands of individual cells for gene expression analysis. (Zhu, Qing et al. 2017) This method has become an important tool in analyzing gene expression profiles in individual cells. Such an analysis was important to this study as it allowed for determination of homogeneity or heterogeneity of cells within populations.

Single cell RNA sequencing measures the distribution of expression levels for each gene across a population of cells. It allows molecular classification of individual cells into subpopulations, e.g., cells at different developmental stages. It is especially suitable for the study of new biological questions in which cell-specific changes in transcriptome are important, e.g., transcriptome changes in embryonic and adult NSCs. While bulk RNA sequencing measures the average expression level for each gene across a large population of cells. However, it is insufficient for studying heterogeneous tissues, e.g., NSCs at different developmental stages.

2.2: PREPROCESSING AND QUALITY CONTROL OF CELLS

All cellular data was obtained from the NBCI database. Both adult and embryonic neural stem cells were obtained from the organism, *Mus musculus*. Initial cellular data was downloaded in .sra format with pair end reads. Reads were split using fastq-dump and hisat2 was utilized to align the sequence against the mouse genome. Following hisat2, the HTSeq package within Python was applied to analyze the sequencing data for outputting gene expression per cell. Each file was output in .csv format and were assigned and merged into a larger .csv file using a file merge algorithm in Python.

Within R, the Scater package and the SingleCellExperiment class was used for analysis of gene expression data. A phenotype file was generated based on cell name and cell type for input into the SingleCellExperiment class. After all data was loaded into a proper format in the SingleCellExperiment class, quality check was performed. This ensured that any low quality reads, such as those that did not capture enough RNA, were eliminated from further analysis. In Scater, a histogram of the number of expressed features was created for each group of cells analyzed. Three groups of cells were analyzed: a combination of adult and embryonic neural stem cells, adult neural stem cells, and embryonic neural stem cells. Based on the visualization of data, low quality cells were eliminated, such as those that expressed 0 genes. Data for assessing the quality of the datasets is visualized in a histogram containing information on the number of expressed genes present across each cell. As most cells were high quality, threshold values to eliminate any low quality reads were chosen to be a log-transformed number of 3 mean absolute deviations below the median log expression size. The number of spike-in and mitochondrial genes were also identified and removed as such genes are not important and distract from the study.

Following the removal of low quality cells, genes of low expression across most cells were also removed. Low expressed genes were filtered out based on a threshold value of 1. This value was verified by creating a histogram that plotted the log-means distribution across all genes.

CHAPTER 3: ASSESSING HOMOGENEITY OF ADULT NEURAL STEM CELLS AND EMBRYONIC NEURAL STEM CELLS

3.1: ANALYSIS OF THE AMOUNT OF GENES EXPRESSED ACROSS CELLS

As the purpose of this study was to analyze similarities and differences in gene expression across adult neural stem cells and embryonic neural stem cells, it was important to assess whether the amount of genes expressed across all cells was similar or different within both datasets. This study was important in determining if either adult or embryonic neural stem cells expressed a more diverse set of genes.

3.1.1: METHODS

The Scater package within R was used to determine diversity in gene types across adult and embryonic neural stem cells. A phenotype spreadsheet in Excel with the cell name and cell type was created for each dataset and another spreadsheet for gene count data was output by Python and R. Both were saved to the SingleCellExperiment class within Scater. The histogram feature within R was used to plot the total amounts of genes expressed against the number of cells.

3.1.2: RESULTS

The number of expressed genes against the number of cells expressing those genes for each dataset was plotted. Three datasets were examined in this study which consisted of adult neural stem cells, embryonic neural stem cells, and a combination of adult and embryonic neural stem cells that have been pooled together. The histograms below (**Figure 1a-1c)** showed that there are more cells in the adult dataset that expressed a lower amount of genes than there are in the embryonic neural stem cells dataset. In the embryonic neural stem cell data, most peaks are present towards the right of the graph, indicating that most cells are expressing large amounts of genes. The combined dataset takes into account both adult and embryonic neural stem cell counts of all genes and contains a significant amount of cells expressing and average number of genes and a high number of genes, and few cells expressing less than 1000 genes.

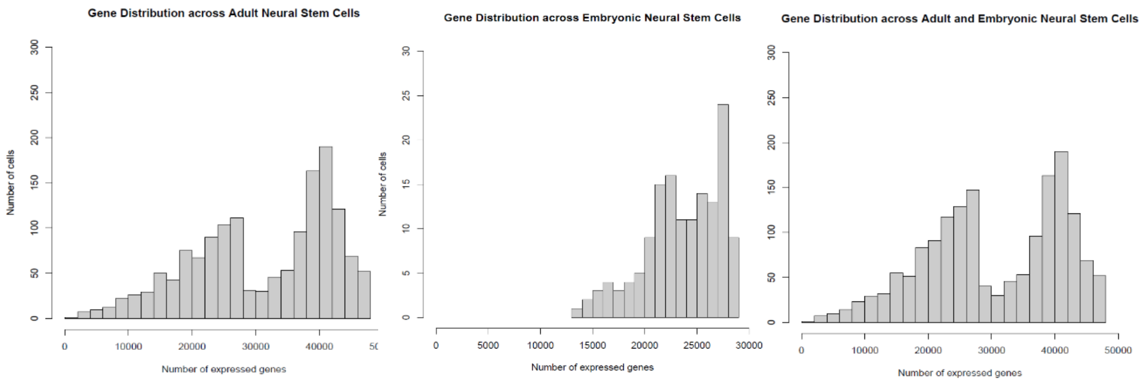


Figure 1a: Histogram plotting the number of expressed genes in adult neural stem cells across the total number of cells.

Figure 1b: Histogram plotting the number of expressed genes in embryonic neural stem cells across the total number of cells.

Figure 1c: Histogram plotting the number of expressed genes in adult and embryonic neural stem cells across the total number of cells.

3.2: ANALYSIS OF GENE EXPRESSION VALUES ACROSS ALL GENES

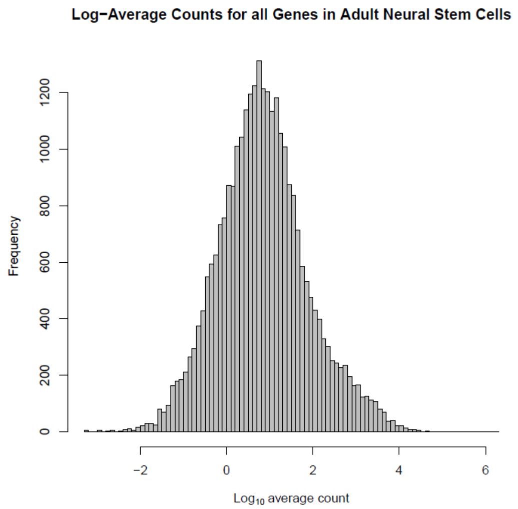
Gene expression values across all genes were evaluated to determine whether there were similarities or differences within adult and embryonic neural stem cells. The purpose of this test is to evaluate whether certain? genes are more strongly expressed in one dataset over another.

3.2.1: METHODS

Using the SingleCellExperiment class within Scater in R, data was extracted from the adult neural stem cell countfile, the embryonic neural stem cell countfile, and the countfile that pooled both adult and embryonic neural stem cells together. (McCarthy, Campbell et al. 2017) The histogram feature within R was utilized to plot the intensity of gene expression values against the number of genes expressed.

3.2.2: RESULTS

Histograms depicting the log average counts of all cells is depicted below for each dataset. Based on the results of the three histograms, it was evident that embryonic neural stem cells contained more genes that are lowly expressed than the adult dataset. The rectangular region to the left of the embryonic neural stem cell histogram is indicative of this as there is a higher frequency of genes with counts under a log average of 0. Adult neural stem cells, as shown by **figure 2a**, contained a high amount of moderately expressed genes with low abundances of lowly and highly expressed genes. On the other hand, embryonic stem cells contain an abundance of highly expressed genes, an average amount of moderately expressed genes, and a low to moderate amount of lowly expressed genes as shown by **figure 2b**. When combining the two datasets, shown in **figure 2c,** there are two peaks present: one is at moderate frequency and one is at high frequency.



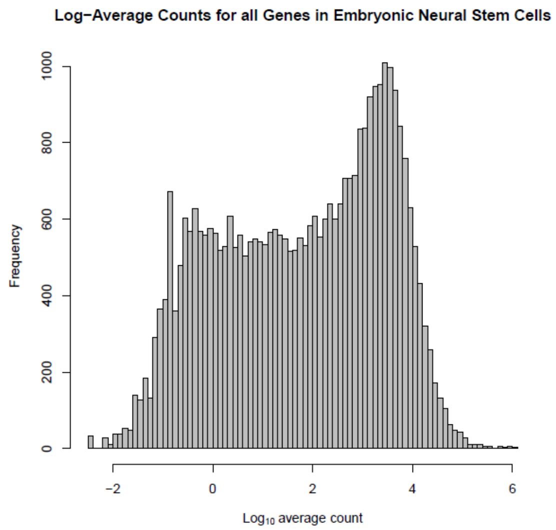


Figure 2a (left): Histogram plotting the log averaged counts of across all genes in the adult neural stem cell dataset.

Figure 2b (right): Histogram plotting the log averaged counts of across all genes in the embryonic neural stem cell dataset.

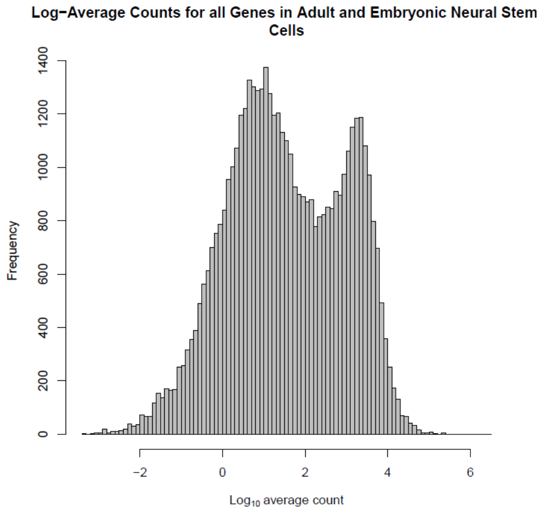


Figure 2c: Histogram plotting the log averaged counts of across all genes in the adult and embryonic neural stem cell dataset.

3.3: COMPUTING SIZE FACTORS TO ASSESS HOMOGENEITY OF CELLS

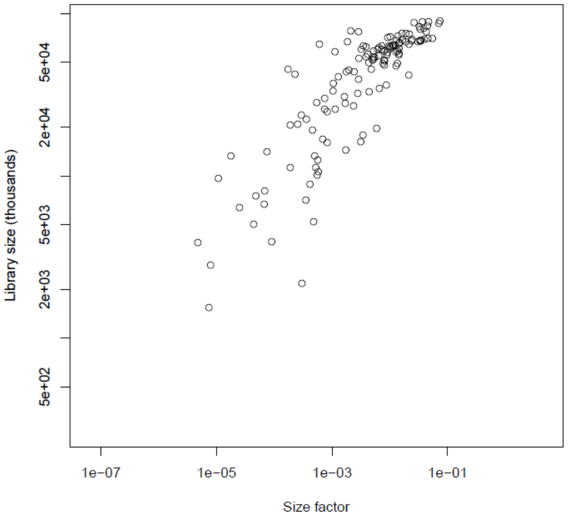
Computing size factors was important in assessing homogeneity of cell types as this method accurately groups clusters together to determine if they belong to a specific cell type. (McCarthy, Campbell et al. 2017)

3.3.1: METHODS:

A graph was generated which plotted the size factor determined by deconvolution against the library sizes for the adult neural stem cell dataset, embryonic neural stem cell dataset, and the combined adult and embryonic neural stem cell dataset. To compute size factors, cell specific biases were normalized and deconvolution to cluster similar cells together based on differential gene expression was performed using the computeSumFactors function. The final step was to calculate size factors by scaling them to compare cells that were part of different clusters.

3.3.2: RESULTS:

Figure 3a and **figure 3b** below represents the size factor scaling of clusters based on differentially expressed genes for adult neural stem cells and embryonic neural stem cells respectively. **Figure 3c** is representative of size factor scaling of clusters within the dataset that included both adult neural stem cells and embryonic neural stem cells. There is a slightly linear trend exhibited in **figure 3a** with a cluster of cells in the top region of the graph. **Figure 3b** displays a strongly linear trend and **figure 3c** shows a strongly linear trend in the bottom region with a slightly linear, more clustered trend above it.



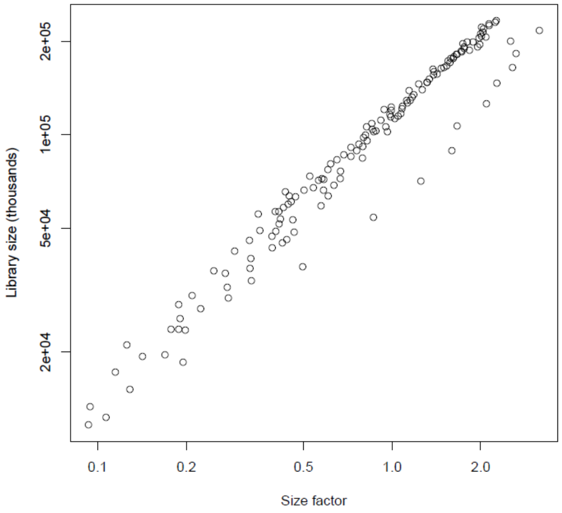


Figure 3a (left): Plot of size factors of clusters against library size for adult neural stem cells.

Figure 3b (right): Plot of size factors of clusters against library size for embryonic neural stem cells.

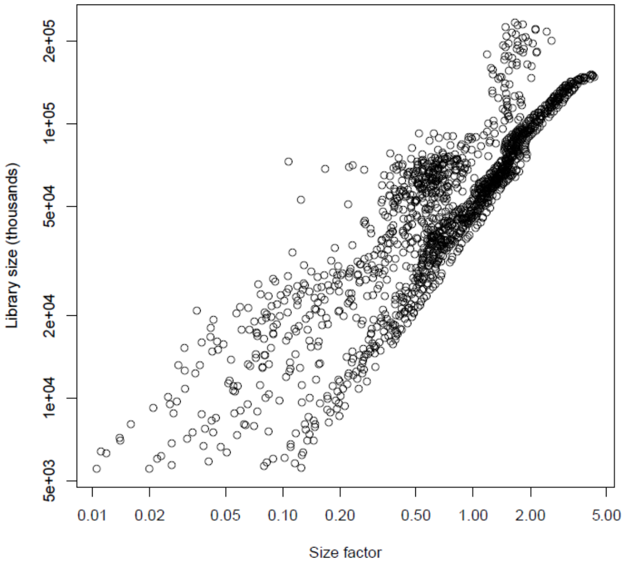


Figure 3c: Plot of size factors of clusters against library size for adult and embryonic neural stem cells.

3.4: DISCUSSION

3.4.1: Analysis of the Amount of Genes Expressed Across Cells

The 3 histograms showed that there were more cells in the adult dataset that expressed a lower amount of genes than there were in the embryonic neural stem cells dataset. In the embryonic neural stem cell dataset, most peaks are present towards the right of the graph indicating that most cells expressed large amounts of genes. The combined dataset took into account both adult and embryonic stem cell counts of all genes and contained a significant amount of cells expressing a moderate number of genes and a high number of genes. Very few cells expressed less than 1000 genes. When analyzing the combined dataset histogram, it was evident that adult neural stem cells and embryonic neural stem cells occupied their own regions of the graph as they retained the patterns they exhibited when they were plotted solely on their own. The inhomogeneous pattern of **figure 2c** indicated that there were differences in the amounts of genes cells expressed between adult and embryonic neural stem cells.

3.4.2: Gene Expression Values across Genes

Based on these histograms, it was evident that in the adult dataset, there was a high frequency of genes that were moderately expressed, whereas in the embryonic dataset, there was a more diverse distribution. In the embryonic dataset, a fair amount of genes were moderately expressed and a large amount of genes were highly expressed. The nature of gene expression between adult and embryonic neural stem cells was therefore quite different as most genes in the adult dataset were moderately expressed whereas in the embryonic stem cell dataset, there was an abundance of highly expressed genes. The last graph which combined both adult and embryonic genes, had two distinct peaks. The peak of genes with a moderate amount of gene expression represented the adult dataset and the peak toward more highly expressed genes represented the adult dataset. This was important because based on this histogram, adult and embryonic datasets were grouped separately based on the gene expression values indicating an inhomogeneous cell population.

3.4.3: Computing Size Factors to Assess Homogeneity

As shown by **figure 4a**, there was a slightly linear trend with a dense cluster of cells toward the top right of the graph. The cluster of cells and moderate linearity implied that all the cells were of the same population. The scatter of cells present around the main cluster can be attributed to cells being in a different, perhaps very early, stage of differentiation. Due to this, there may be some differential expression between cells in the adult population. In the embryonic population of cells, as shown by **figure 4b**, there was a strongly linear trend suggesting a very homogeneous population with little to no differential expression between cells. Once the adult and embryonic cells were pooled together, as depicted by **figure 4c**, there appeared to be two linear trends present implying that there were multiple populations of cells. Based on a comparison of **figure 4c** against **figure 4a** and **figure 4b**, the adult population of neural stem cells remained distinct from the embryonic population of neural stem cells. The data suggested that there was differential expression between the cells in each dataset. Since both the adult stem cells and embryonic stem cells retained the shape of their trend, the graph implied that a homogenous population of cells did not exist.

CHAPTER 4: EXAMINING SIMILARITIES AND DIFFERENCES BETWEEN GENES EXPRESSED IN ADULT AND EMBRYONIC NEURAL STEM CELLS

Expression of different types of genes within cells plays a big role in determining their final function. Different genes are highly expressed or differentially expressed for different types of cells. Throughout the following analyses, adult neural stem cells and embryonic neural stem cells are analyzed for gene expression similarities and differences.

CHAPTER 4.1: EXAMINATION OF SIMILARITIES AND DIFFERENCES IN THE TOP 60 GENES OF ADULT NEURAL STEM CELLS AND EMBRYONIC NEURAL STEM CELLS

To determine differences in gene expression, a graph of the top 60 highest expressed genes was created for both adult and embryonic neural stem cell datasets.

4.1.2: Results

The most expressed genes were analyzed by utilizing the plotQC function in the Scater package in R. The type was set to highest expression in order to extract the 60 genes that displayed the highest count values. A plot was created for the 60 highest expressed genes in the adult neural stem cell dataset and the embryonic neural stem cell dataset. For visualization of the number of genes that were similar or different between the two datasets, a Venn diagram was created. Finally, a table was created to display exactly which genes were similar or different between the two datasets.

Upon plotting the top genes expressed in each dataset, there is a noticeable overlap in the highest expressed genes within both the adult and embryonic datasets. Presented below are the top genes that are present in the adult and embryonic datasets. The most highly expressed gene in the adult dataset was Malat1 while the most highly expressed gene in the embryonic dataset expressing Ubb. There is a 31.6% overlap amongst the 60 highest expressed genes in the adult dataset and the embryonic dataset.

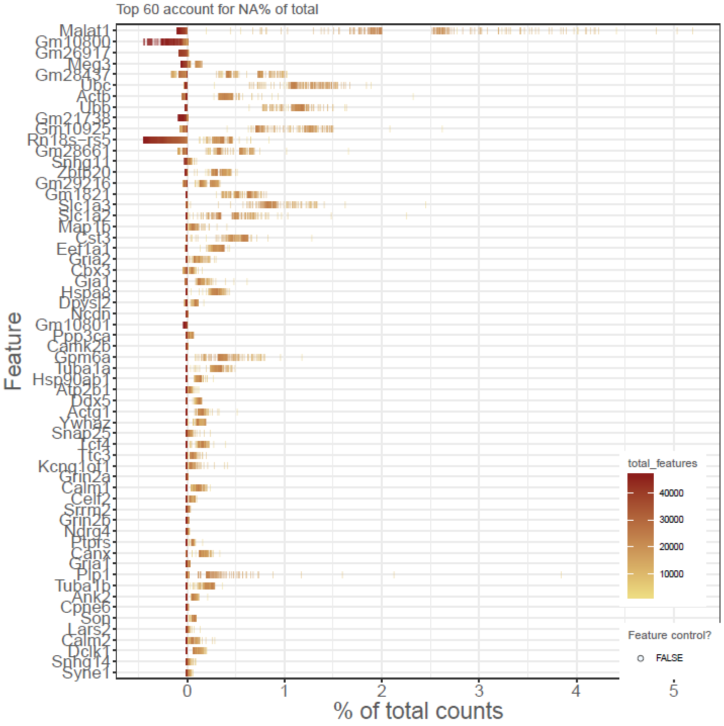


Figure 4a: Plot of the top 60 highest expressed genes in adult neural stem cells.

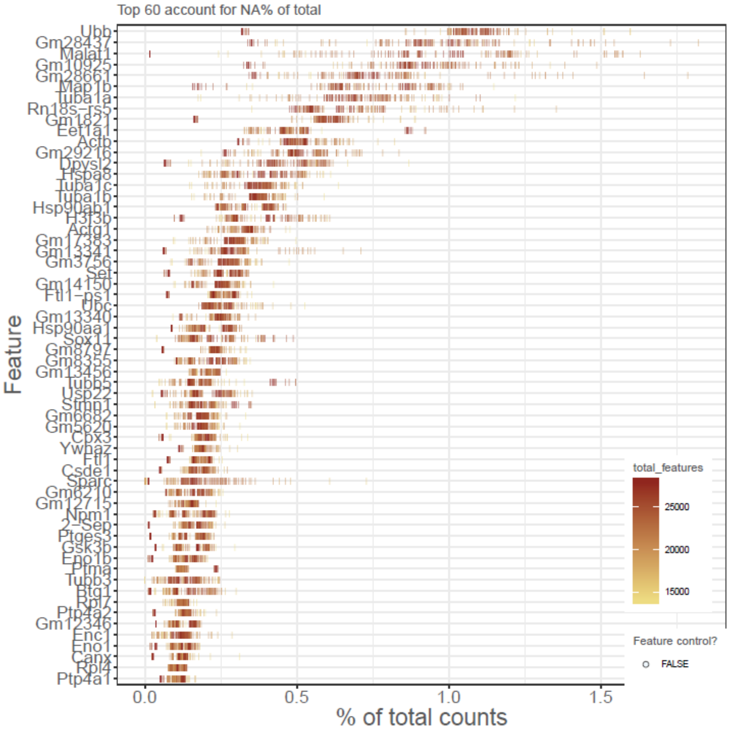


Figure 4b: Plot of the top 60 highest expressed genes in embryonic neural stem cells.

The Venn diagram below represents the number of genes that are highly expressed amongst multiple groups and genes that are distinct to a particular group. A total of 19 out of the 60 highest expressed genes were common between the adult neural stem cell dataset and the embryonic neural stem cell dataset. There were a total of 40 genes that were distinct within the adult neural stem cell dataset and a total of 39 genes that were distinct in the embryonic neural stem cell dataset.

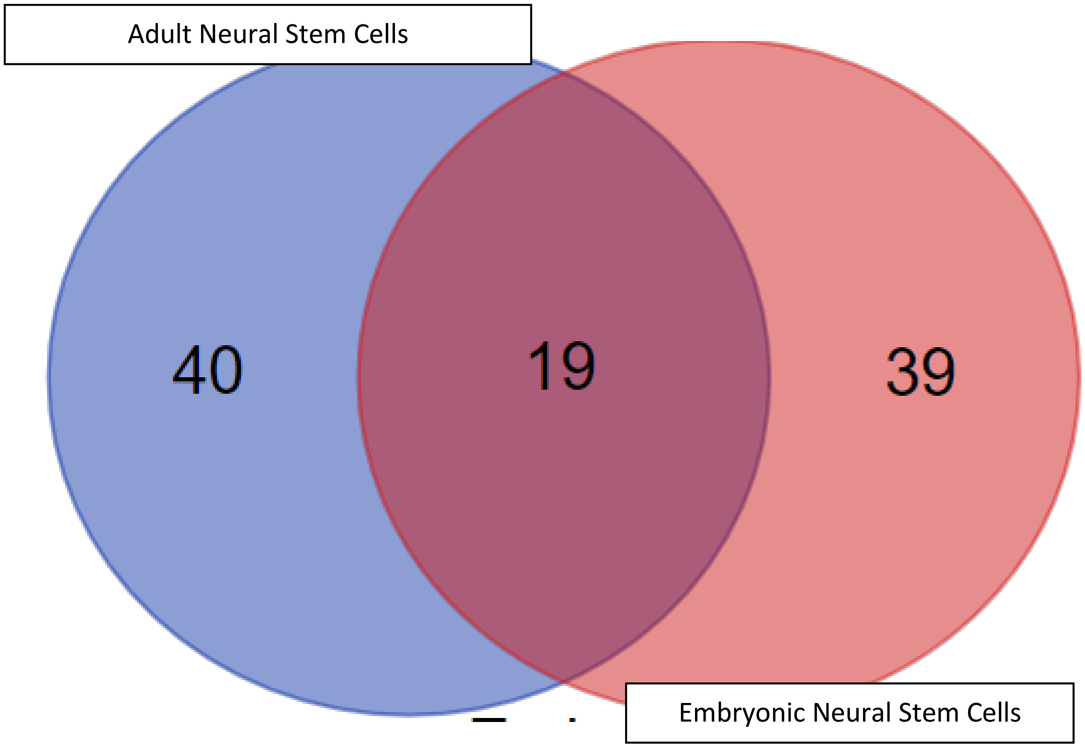
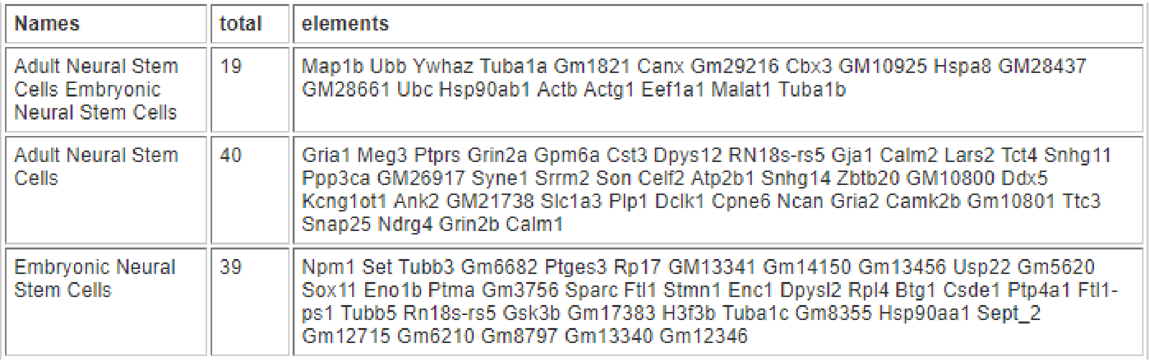


Figure 4c: Venn diagram depicting the number of similar and different genes within the top 60 highest expressed genes for adult and embryonic neural stem cells.

The table below depicts exactly which genes are common and which genes are different within adult neural stem cells and embryonic neural stem cells.



Names	total	elements
Adult Neural Stem Cells Embryonic Neural Stem Cells	19	Map1b Ubb Ywhaz Tuba1a Gm1821 Canx Gm29216 Cbx3 GM10925 Hspa8 GM28437 GM28661 Ubc Hsp90ab1 Actb Actg1 Eef1a1 Malat1 Tuba1b
Adult Neural Stem Cells	40	Gria1 Meg3 Ptprs Grin2a Gpm6a Cst3 Dpys12 RN18s-rs5 Gja1 Calm2 Lars2 Tct4 Snhg11 Ppp3ca GM26917 Syne1 Srrm2 Son Celf2 Atp2b1 Snhg14 Zbtb20 GM10800 Ddx5 Kcng1ot1 Ank2 GM21738 Slc1a3 Plp1 Dclk1 Cpne6 Ncan Gria2 Camk2b Gm10801 Ttc3 Snap25 Ndrg4 Grin2b Calm1
Embryonic Neural Stem Cells	39	Npm1 Set Tubb3 Gm6682 Ptges3 Rp17 GM13341 Gm14150 Gm13456 Usp22 Gm5620 Sox11 Eno1b Ptma Gm3756 Sparc Ftl1 Stmn1 Enc1 Dpysl2 Rpl4 Btg1 Csde1 Ptp4a1 Ftl1-ps1 Tubb5 Rn18s-rs5 Gsk3b Gm17383 H3f3b Tuba1c Gm8355 Hsp90aa1 Sept_2 Gm12715 Gm6210 Gm8797 Gm13340 Gm12346

Table 1: Table depicting similar and different genes in adult neural stem cells and embryonic neural stem cells.

CHAPTER 4.2: PCA CLUSTERING OF NEURAL STEM CELLS FOR CELL TYPE ASSORTMENT

PCA clustering is a method present in the Scater package in R to plot cells in clusters based on similarities in log-expression values of genes. This can be useful in evaluating different populations or subpopulations within cells.

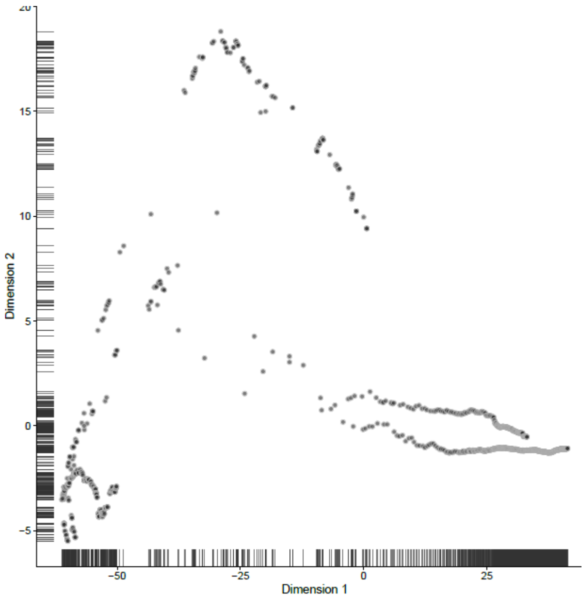
4.2.1: Methods

In order to perform PCA clustering, deconvolution for normalization must first be performed for dimensionality reduction. The purpose of normalization is to ensure that any cell-specific bias is eliminated. This is conducted by assuming that there is no differential expression between cells. As a measure of how much the counts should be scaled per library, size factors are calculated. This was done using a deconvolution method where counts from many different cells were pooled to higher the number of counts for a more accurate estimate of size factor.(McCarthy, Campbell et al. 2017) A graph was generated to plot size factors against library sizes. To ensure that normalization was successful, size factors for spike-ins were calculated as these values should be uniform since each cell contains the same RNA spike composition. Using the count data generated after applying size factors, values of normalized log expression were calculated per cell. A log transformation was utilized to stabilize variance across highly abundant genes.

Following normalization of count data, dimensionality reduction was applied to depict similarities and differences between cell types. A PCA, or principal component analysis, plot was created to visualize this based on the top 500 most variably expressed genes. The cells were arranged based on highly correlated genes. A plot graph was outputted using the plotReducedDim function within Scater in R. (McCarthy, Campbell et al. 2017)

4.2.2: Results

The PCA plot presented in **figure 5a** below shows the similarities between only adult neural stem cells. **Figure 5B** indicates two possible groupings of embryonic neural stem cells, however, a wide scattering of cells still exists. **Figure 5c** shows that embryonic neural stem cells resided very close to each other while adult neural stem cells were more scattered yet still remained fairly close to each other.



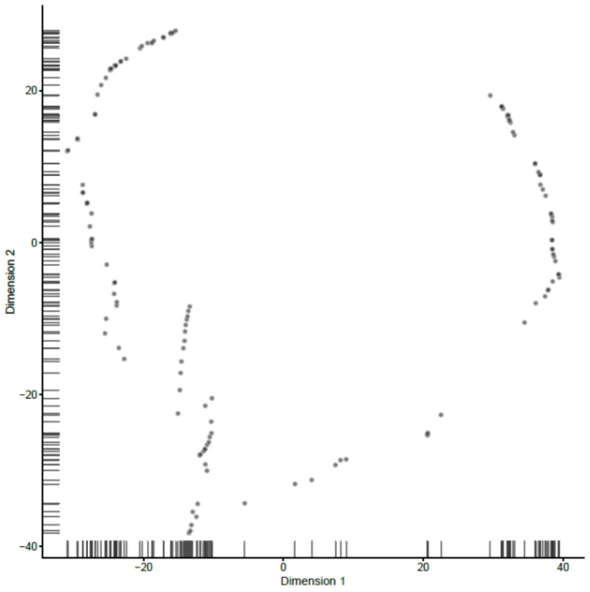


Figure 5a (left): PCA plot showing dispersion of adult neural stem cells in reduced dimension.

Figure 5b (right): PCA plot showing dispersion of embryonic neural stem cells in reduced dimension.

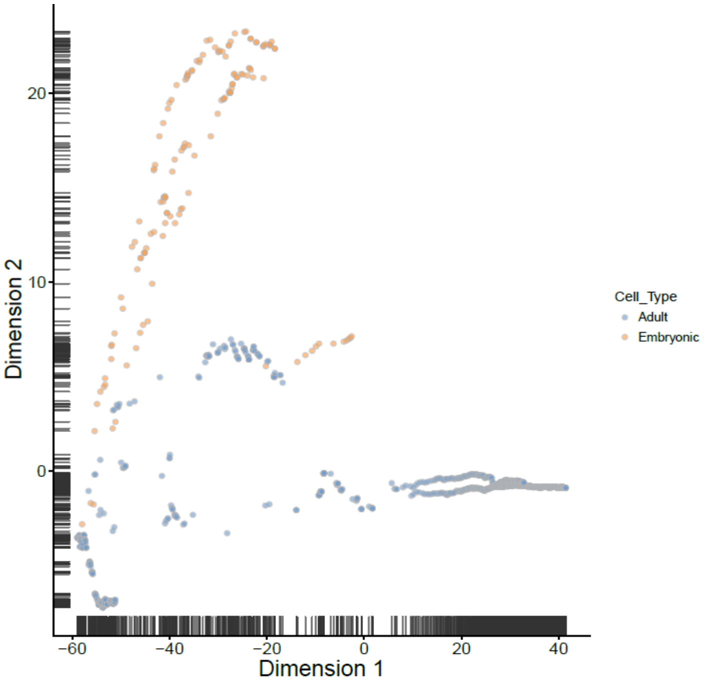


Figure 5c: PCA plot showing dispersion of embryonic neural stem cells in reduced dimension.

CHAPTER 4.3: T-SNE CLUSTERING OF NEURAL STEM CELLS FOR CELL TYPE ASSORTMENT

4.3.1: Methods

Along with a PCA plot to visualize the similarities and differences between the datasets, a t-SNE, or a t-stochastic neighbor embedding method was used. This test is known to require more computational effort when grouping cells and is also more accurate as it can readily detect non-linear relationships. (McCarthy, Campbell et al.

2017) This method was run about 5 times to confirm that the results were reproducible each time. The plotTSNE function with a seed of 100 was chosen. The perplexity parameter was set to 5, 10, and 20 to see whether the distribution of cells would be altered.

4.3.2: Results

In the t-SNE plot of adult cells shown in **figure 6a**, approximately 3 groups were formed based on correlation between gene expressions, however, in comparison to the PCA plot, the three groups were not as distinct. Each perplexity value resulted in the same placement of cells. In the t-SNE plot of embryonic neural stem cells presented in **figure 6b** below, there were 3 distinct groups of embryonic neural stem cells present. The splitting into different groups can be attributed to the two different stages that embryonic neural stem cells were present in: proliferating neural progenitors and immature neurons. (Chen, Friedman et al. 2017)

When examining the t-SNE plot of the combined dataset with both adult neural stem cells and embryonic neural stem cells present, three adult neural stem cell clusters remained, however, embryonic neural stem cells are all clustered together. The cluster of embryonic cells in the t-SNE plot is very distinct from the clusters of adult neural stem cells indicating a high level of variation in gene expression between the adult and embryonic neural stem cell dataset. As is evident by the t-SNE plot, the embryonic stem cells were grouped into one cluster instead of the three that were present before.

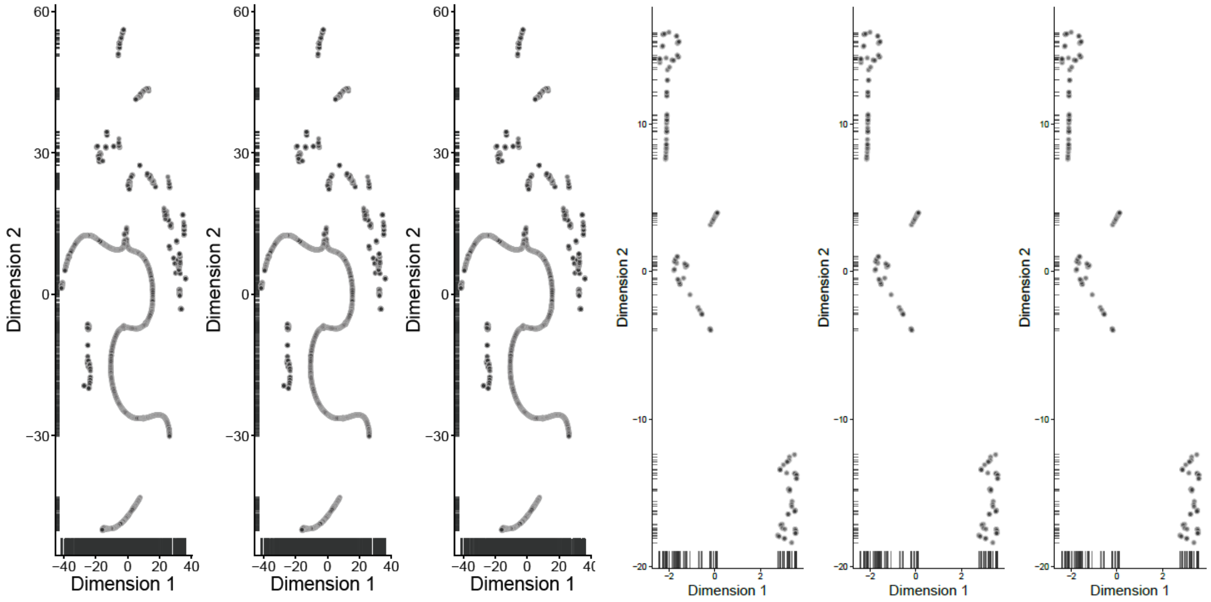


Figure 6a (left): t-SNE plot depicting scatter of cells in reduced dimension for adult neural stem cells with different perplexity values.

Figure 6b (left): t-SNE plot depicting scatter of cells in reduced dimension for embryonic neural stem cells with different perplexity values.

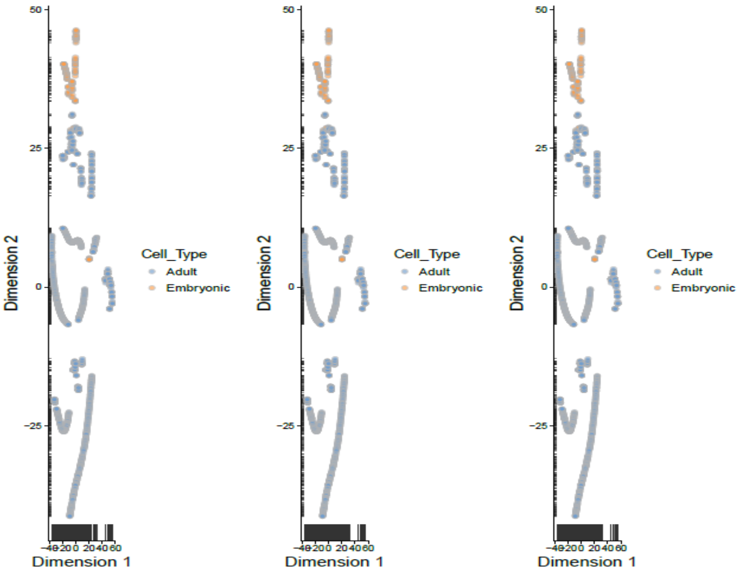


Figure 6c: t-SNE plot depicting scatter of cells in reduced dimension for adult and embryonic neural stem cells with different perplexity values.

CHAPTER 4.4: HIERARCHICAL CLUSTERING ANALYSIS SPLITS CELLS INTO GROUPS BASED OF DIFFERENCES IN DIFFERENTIAL GENE EXPRESSION

4.4.1: Methods

Following PCA and t-SNE analysis, hierarchical clustering, with the Dynamic Tree cut method in particular, was implemented to detect important gene clusters in each of the three datasets. The Dynamic Tree cut algorithm is an iterative process that started by viewing the entire dataset as a singular cluster or small amount of large clusters. Then, the algorithm deconstructed the clusters based on gene expression patterns until stability was achieved within a single smaller cluster. The top 500 differentially expressed genes were accounted for when constructing the clusters as genes that were not differentially expressed would alter the data as they would be too highly correlated between cells. Each cluster in this case contains genes of similar differential expression and avoids clustering based on genes that are constantly highly expressed as this could lead to false correlation values between genes. Over splitting was avoided by joining very small clusters to their nearest neighbors. After the algorithm grouped cells together, the distribution of cells from adult and embryonic datasets was examined and information was output into a table.

After cells were grouped into clusters, the width of each cluster was analyzed to check if clusters were stable. Clusters with silhouette widths of values near 1 indicated that over-clustering did not occur and clusters were distinct enough to stand on their own. If cluster widths were closer to zero, over-clustering had occurred and cells within clusters had the potential to be part of another cluster.

4.4.2: Results

Based on the Dynamic Tree Cut method for hierarchical clustering, there were three different clusters that formed from the adult and embryonic combined dataset. The first cluster only contained adult neural stem cells, the second one contained 379 adult neural stem cells and 61 embryonic neural stem cells, and the third cluster contained 114 adult neural stem cells and 89 embryonic neural stem cells. Based on this clustering method, there were no clusters that contained only embryonic neural stem cells.

Clusters	Adult	Embryonic
1	885	0
2	379	61
3	114	89

Table 2a: Table depicting number of clusters and amounts of each cell type per cluster using the dynamic tree cut method for adult and embryonic neural stem cells.

Upon examining the dataset for clusters within just the adult neural stem cells, there were three clusters present. Clusters one and three contained a similar amount of cells in comparison to the adult and embryonic combined dataset, however cluster 2 in the adult neural stem cell dataset contained significantly more cells than were present in the combined dataset. This can be attributed to cells within cluster 2 having a higher amount of similarities within differentially expressed genes to certain embryonic neural stem cells than other adult neural stem cells.

Clusters	Adult
1	877
2	459
3	161

Table 2b: Table depicting number of clusters and amounts of adult neural stem cells per cluster using the dynamic tree cut method.

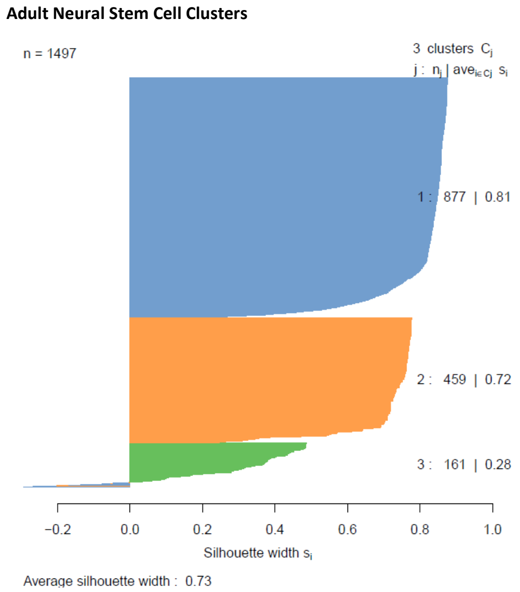
As evident by **table 1b**, embryonic neural stem cells were split into 5 clusters with a variable amount of genes in each cluster. When paired with adult neural stem cells, embryonic neural stem cells were only split into two distinct clusters. There were fewer clusters of embryonic neural stem cells present due to high similarities in differential gene expression between clusters 2 and 3 in adult neural stem cells. Thus certain embryonic neural stem cells have a higher correlation in differential gene expression to adult neural stem cells than other embryonic neural stem cells. This can be attributed to both groups of cells containing subgroups of intermediate progenitor cells and immature neurons.

Clusters	Embryonic
1	42
2	34
3	25
4	22
5	21

Table 2c: Table depicting number of clusters and amounts of embryoinc neural stem cells per cluster using the dynamic tree cut method.

The clusters above were tested for silhouette width to indicate whether clusters were well defined. As shown by **figure 6a** below, the average silhouette width for adult clusters was 0.73. The width of clusters 1, 2, and 3 were 0.81, 0.72, and 0.28 respectively. The third cluster had a very small width indicating the potential for over-clustering. This can be attributed to cells within this cluster having similar differentially expressed genes to cells within other clusters, however cells in other clusters may not exhibit differential expression in these genes. When examining the embryonic neural stem cell dataset, the average silhouette width was 0.46, indicating the presence of over-clustering. Two clusters were present with silhouette widths of 0.58 and 0.56 representing the most distinct clusters within the embryonic neural stem cell dataset. The remaining 3 clusters all had silhouette widths that were 0.50 and below indicating that these clusters had similar patterns in differential gene expression with slight variances.

Figure 6c depicts silhouette widths of the adult and embryonic combined dataset. The average silhouette width was 0.69 indicating an overall stability of clusters. Cluster 1, which contained only adult neural stem cells was the most stable with a silhouette width of 0.83. This was followed by cluster 2 which contained a silhouette width of 0.63. This cluster contained 379 adult neural stem cells and 61 embryonic neural stem cells. The third cluster, containing 114 adult neural stem cells and 89 embryonic neural stem cells, had a much lower value of 0.22 indicating a possibility of over-clustering.



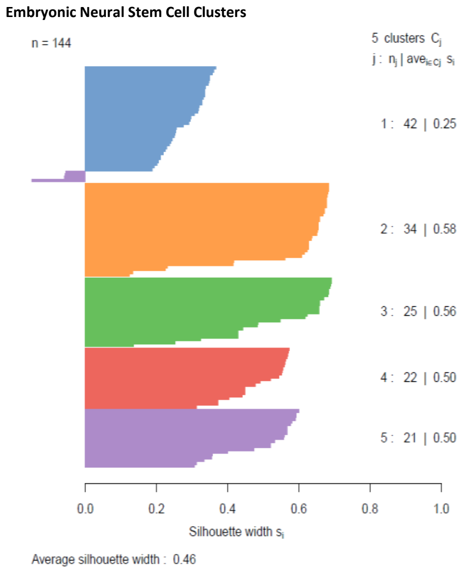


Figure 6a (left): Graph depicting silhouette widths for adult neural stem cell clusters. The numbers to the right depict the number of cells per cluster.

Figure 6b (right): Graph depicting silhouette widths for embryonic neural stem cell clusters. The numbers to the right depict the number of cells per cluster.

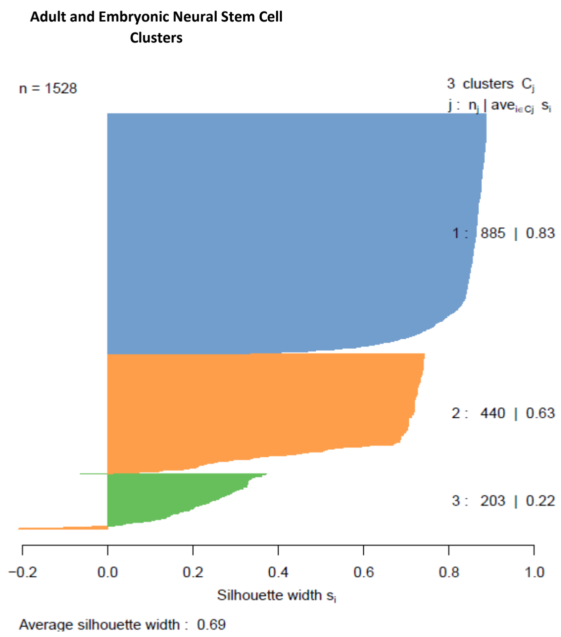


Figure 6c: Graph depicting silhouette widths for adult and embryonic neural stem cell clusters. The numbers to the right depict the number of cells per cluster.

CHAPTER 4.5: COMPARING DIFFERENTIALLY EXPRESSED GENES BETWEEN CLUSTERS FOR INSIGHT ON SIMILARITIES AND DIFFERENCES ACROSS CLUSTERS

4.5.1: Methods

Differential Expression of the top 50 genes was analyzed to determine marker genes that distinguished clusters from each other. Values for differential expression were log-scaled from -2 to 2 with positive values (red) indicating upregulation of genes, negative values (blue) indicating downregulation of genes, and a value of 0 (yellow) indicating that there was no change in expression. Each cluster and cell type was represented by a different color. A legend was plotted to the right of the heatmap.

4.5.2: Results

When the top 50 differentially expressed genes were plotted, as shown by **figure 7** below, it was evident that there is a clear distinction between genes that are upregulated

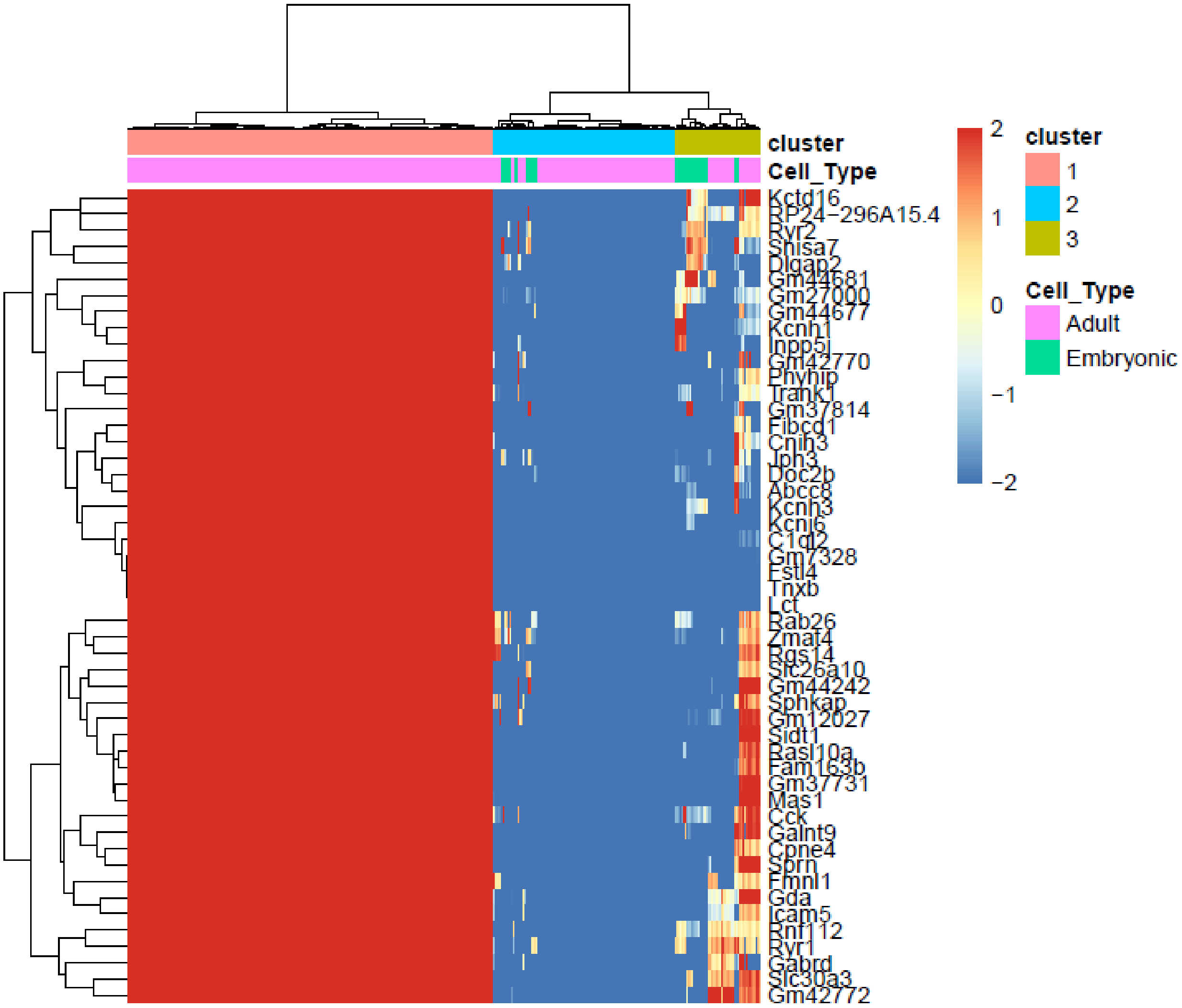


Figure 7: Heatmap showing the top 50 differentially expressed genes in adult and embryonic neural stem cell clusters

CHAPTER 4.6: DISCUSSION

4.6.1: Examination of Similarities and Differences in the top 60 Genes of Adult Neural Stem Cells and Embryonic Neural Stem Cells

Based on the data provided by the Venn diagram of the 60 highest expressed genes within each dataset, it was evident that there were more genes that are different within the two datasets than were similar. This signified that adult and embryonic neural stem cells potentially contained more differences than similarities in gene expression.

4.6.2: PCA Clustering of Neural Stem Cells for Cell Type Assortment

In **figure 5a**, there were 3 separate regions where cells congregated which can be attributed to different stages in differentiation. The three known cell groups present within the adult stem cells were: quiescent neural stem cells, pre-active neural stem cells, active neural stem cells and early intermediate progenitor cells. (Shin, Berg et al. 2015) The number of groups correlated with the number of cell subtypes and thus, as the PCA plot suggested, there was a distinct gene expression profile within each cell group.

In **figure 5b**, there existed a scattering within the embryonic neural stem cell population. There were two regions in the graph with large numbers of cells indicating two subpopulations within the embryonic neural stem cell population. This can be attributed to the embryonic neural stem cell dataset containing both immature neurons and intermediate progenitor cells.

As depicted by **figure 5c**, there were distinct regions where adult neural stem cells exist and embryonic neural stem cells resided on the PCA plot. However, there were two regions where adult and embryonic neural stem cells were very close to each other on the plot. There existed distinct regions where adult and embryonic stem cells separate, indicating high levels of gene expression differences between both cell types. Although highly scattered when graphed alone, embryonic stem cells group more closely to each other when combined with adult neural stem cells indicating a higher correlation between gene pair expression amongst themselves than adult neural stem cells. Such clustering indicated that there were gene expression differences present between embryonic and adult neural stem cells.

4.6.3: t-SNE Clustering of Neural Stem Cells for Cell Type Assortment

Based on **figure 6a**, the clustering of adult neural stem cells into 3 different regions correlated with cells being present as quiescent neural stem cells, progenitor cells, and immature neurons. Based on there being three different clusters, there were differentially expressed genes present in all three clusters. In **figure 6b**, embryonic neural stem cells were divided into three different clusters. This can be attributed once again to cells being present in different stages. When cells were pooled into one dataset, three clusters were present for adult neural stem cells as was the case in the t-SNE plot for solely adult neural stem cells. However, unlike the t-SNE plot for solely embryonic neural stem cells, embryonic neural stem cells were clustered into one region in **figure 6c**. This indicated that embryonic neural stem cells were more similar to each other based on differential gene expression than they were to adult neural stem cells. Embryonic neural stem cells clustered closely next to one group of adult neural stem cells, indicating that embryonic cells had more similarities to one cluster of adult neural stem cells than the others.

4.6.4: Hierarchical Clustering Analysis Splits Cells into Groups Based of Differences in Differential Gene Expression

The two clusters in the adult neural stem cell dataset that contained silhouette widths that were close to 1 implied that both clusters are stable to exist by themselves. This implied that there were a significant amount of differentially expressed genes between these two clusters. The final cluster in the adult dataset had a much smaller value indicating a possibility of over-clustering. The embryonic neural stem cell dataset was split into 5 clusters, however all silhouette widths were under 0.60 indicating that there are not many differences in differentially expressed genes amongst all cells in the embryonic neural stem cell dataset and over-clustering might have occurred. The two most distinct clusters were likely representative of neural progenitor cells and immature neurons. Due to the lower silhouette widths throughout the embryonic neural stem cells dataset, a more homogenous population of cells is implied compared to the adult neural stem cells.

Within the adult and embryonic combined dataset, splitting of cells into three different clusters indicated the presence of differences in differentially expressed genes. Cluster 1 and cluster 2 were stable with silhouette widths close to 1, indicating the presence of an inhomogeneous population of neural stem cells. A lower silhouette width within cluster 3 can be attributed to some similarities in differential gene expression in cells of these clusters with cells of other clusters. Due to the presence of both embryonic and adult neural stem cells in clusters 2 and 3, it is implied that there are similarities in differentially expressed genes between adult and embryonic neural stem cells within these two clusters. A plausible explanation for the grouping of adult neural stem cells and embryonic neural stem cells into the same clusters is that both groups of cells contain neural progenitor cells and immature neurons. There might be similar patterns in differential gene expression for cells in these stages which caused for the overlap between the adult and embryonic neural stem cell dataset in cluster 2 and cluster 3.

In order to look more closely at the differential expression pattern in the three clusters, a heatmap was generated containing the top 50 differentially expressed genes. Evidently, cluster 1, which contained only adult neural stem cells, contained all upregulated genes. This pattern of gene expression was very distinct from cluster 2 and 3 indicating differences within the adult neural stem cell population itself as well as the embryonic neural stem cell population. Cluster two was distinct from cluster 1 as it contained cells that were downregulated for the top 50 differentially expressed genes. Since this cluster contained both embryonic and adult neural stem cells, there was indication of some similarities in differential gene expression to certain adult neural stem cells. This can once again be attributed to both adult neural stem cell and embryonic neural stem cells containing neural progenitor cells and immature neurons. The third cluster of cells, which again contained both adult neural stem cells and embryonic neural stem cells contained cells that were both upregulated and downregulated for the top 50 differentially expressed genes. This cluster contained similarities in gene expression pattern to both cluster 1 and cluster 2.

When examining embryonic neural stem cells and adult neural stem cells within cluster 3, there are differences present in differential gene expression patterns. For example, for the first 10 genes (Kctd16-Kcnh1), adult neural stem cells vary between upregulation and downregulation with most genes being downregulated, however, embryonic neural stem cells mainly do not exhibit any change or are upregulated for the same genes. Although part of the same cluster, adult and embryonic neural stem cells within cluster three do not contain the same patterns of expression, once again indicating differences in gene expression between adult and embryonic neural stem cells.

CHAPTER 5: CONCLUSION AND FUTURE WORK

CHAPTER 5.1: CONCLUSIONS

Based on the analysis of gene expression patterns in adult and embryonic neural stem cells, there was a clear difference in gene expression between the two datasets. Due to differences in highest expressed genes and differences in differentially expressed genes, adult and embryonic neural stem cells may have different properties. This information can provide a mechanism, for example, to increase the efficiency of adult neural stem cell differentiation. Gene expression patterns in adult stem cells can be potentially altered to be more similar to embryonic neural stem cells to make them as potent and versatile as embryonic neural stem cells. This is advantageous as adult neural stem cells are not controversial since no embryos are being killed in the process. Adult neural stem cells also do not undergo immune rejection as often as embryonic neural stem cells, and thus it would be advantageous if their gene expression patterns could be altered to adopt the advantages of embryonic neural stem cells.

This study can be translated to the human genome. Although gene composition may vary slightly, the mouse genome allows for accurate comparisons against the human genome.

CHAPTER 5.2: FUTURE WORK

In the future, this study can be furthered to pinpoint specific proliferation genes. This is important as one of the main issues with adult neural stem cells is their inability to form a sufficient amount of cells for effective therapies. After determining which genes in embryonic neural stem cells are attributed to proliferation and differentiation efficiency, Monocle, a package within R, can be used to perform pseudotime analysis to track the upregulation and downregulation of these genes at specific timepoints, Using this information, gene expression data in adult neural stem cells can be altered to fit the pattern of that in embryonic neural stem cells and proliferation efficiency can be evaluated.

References:

Ayoub, A. E., et al. (2011). "Transcriptional programs in transient embryonic zones of the cerebral cortex defined by high-resolution mRNA sequencing." Proceedings of the National Academy of Sciences **108**(36): 14950.

> Characterizing the genetic programs that specify development and evolution of the cerebral cortex is a central challenge in neuroscience. Stem cells in the transient embryonic ventricular and subventricular zones generate neurons that migrate across the intermediate zone to the overlying cortical plate, where they differentiate and form the neocortex. It is clear that not one but a multitude of molecular pathways are necessary to progress through each cellular milestone, yet the underlying transcriptional programs remain unknown. Here, we apply differential transcriptome analysis on microscopically isolated cell populations, to define five transcriptional programs that represent each transient embryonic zone and the progression between these zones. The five transcriptional programs contain largely uncharacterized genes in addition to transcripts necessary for stem cell maintenance, neurogenesis, migration, and differentiation. Additionally, we found intergenic transcriptionally active regions that possibly encode unique zone-specific transcripts. Finally, we present a high-resolution transcriptome map of transient zones in the embryonic mouse forebrain.

Bongso, A. and M. Richards (2004). "History and perspective of stem cell research." Best Practice & Research Clinical Obstetrics & Gynaecology **18**(6): 827-842.

Chen, Y.-J. J., et al. (2017). "Single-cell RNA sequencing identifies distinct mouse medial ganglionic eminence cell types." Scientific Reports **7**: 45656.

Habib, N., et al. (2016). "Div-Seq: Single nucleus RNA-Seq reveals dynamics of rare adult newborn neurons." Science (New York, N.Y.) **353**(6302): 925-928.

> Single cell RNA-Seq provides rich information about cell types and states. However, it is difficult to capture rare dynamic processes, such as adult neurogenesis, because isolation of rare neurons from adult tissue is challenging and markers for each phase are limited. Here, we develop Div-Seq, which combines scalable single nucleus RNA-Seq (sNuc-Seq) with pulse labeling of proliferating cells by EdU to profile individual dividing cells. sNuc-Seq and Div-Seq can sensitively identify closely related hippocampal cell types and track transcriptional dynamics of newborn neurons within the adult hippocampal neurogenic niche, respectively. We also apply Div-Seq to identify and profile rare newborn GABAergic neurons in the adult spinal cord, a non-canonical neurogenic region. sNuc-Seq and Div-Seq open the way for unbiased analysis of diverse complex tissues.

McCarthy, D. J., et al. (2017). "Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R." Bioinformatics **33**(8): 1179-1186.

> Motivation: Single-cell RNA sequencing (scRNA-seq) is increasingly used to study gene expression at the level of individual cells. However, preparing raw sequence data for further analysis is not a straightforward process. Biases, artifacts and other sources of unwanted variation are present in the data, requiring substantial time and effort to be spent on pre-processing, quality control (QC) and normalization. Results: We have developed the R/Bioconductor package scater to facilitate rigorous pre-processing, quality control, normalization and visualization of scRNA-seq data. The package provides a convenient, flexible workflow to process raw sequencing reads into a high-quality expression dataset ready for downstream analysis. scater provides a rich suite of plotting tools for single-cell data and a flexible data structure that is compatible with existing tools and can be used as infrastructure for future software development. Availability and Implementation: The open-source code, along with installation instructions, vignettes and case studies, is available through Bioconductor at http://bioconductor.org/packages/scater. Contact: davis@ebi.ac.uk Supplementary information: are available at Bioinformatics online.

Shin, J., et al. (2015). "Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis." Cell Stem Cell **17**(3): 360-372.

Zhu, S., et al. (2017). "Advances in single-cell RNA sequencing and its applications in cancer research." Oncotarget **8**(32): 53763-53779.

> Unlike population-level approaches, single-cell RNA sequencing enables transcriptomic analysis of an individual cell. Through the combination of high-throughput sequencing and bioinformatic tools, single-cell RNA-seq can detect more than 10,000 transcripts in one cell to distinguish cell subsets and dynamic cellular changes. After several years' development, single-cell RNA-seq can now achieve massively parallel, full-length mRNA sequencing as well as in situ sequencing and even has potential for multi-omic detection. One appealing area of single-cell RNA-seq is cancer research, and it is regarded as a promising way to enhance prognosis and provide more precise target therapy by identifying druggable subclones. Indeed, progresses have been made regarding solid tumor analysis to reveal intratumoral heterogeneity, correlations between signaling pathways, stemness, drug resistance, and tumor architecture shaping the microenvironment. Furthermore, through investigation into circulating tumor cells, many genes have been shown to promote a propensity toward stemness and the epithelial-mesenchymal transition, to enhance anchoring and adhesion, and to be involved in mechanisms of anoikis resistance and drug resistance. This review focuses on advances and progresses of single-cell RNA-seq with regard to the following aspects: 1. Methodologies of single-cell RNA-seq 2. Single-cell isolation techniques 3. Single-cell RNA-seq in solid tumor research 4. Single-cell RNA-seq in circulating tumor cell research 5. Perspectives