**Searching for differentially expressed genes in seven cell types of Mus Musculus.**

Alexey Efanov

Data Science, Bellevue University

DSC 680: Applied Data Science

Instructor: Professor Catherine Williams

February 1, 2021

**Business Problem**

Next generation sequencing is a powerful tool to analyze gene expression. Differential gene expression analysis allows us to select genes, whose expression is critical for tissues differentiation, for different disease development, including cancer. That makes differential gene expression analysis a highly important tool to uncover mechanisms of different diseases [1].

**Background/History**

Information about a cell or an organism development is located in chromosomes. A chromosome is a single large polymer molecule consisting of 4 monomers (deoxyribonucleotides) denoted as A, T, G, C. A stands for deoxyadenosine, T stands for deoxythymidine, C stands for deoxycytosine, and G stands for deoxyguanosine. Nucleotides are building blocks of chromosomes. Each chromosome consists of two strands. Each strand consists of a linear sequence of nucleotides. Two strands are linked by hydrogen bonds. A in one strand pairs with T in another (complimentary strand), and C in one strand pairs with G in the complimentary strand as shown in Figure 1.

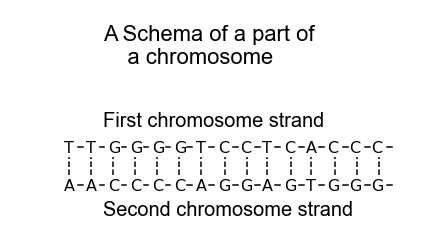
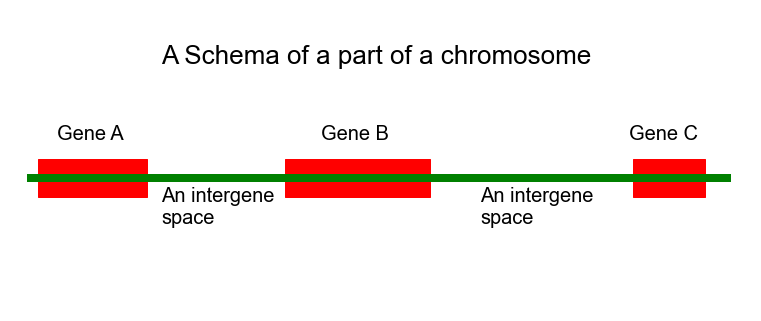


Figure 1. Solid lines depict intra-strand covalent bonds. Dotted lines depict inter-strand hydrogen bonds

In living organisms, nearly of processes and features, with a few exceptions, are carried out by proteins. Examples of such processes and features-weight, height, color of eyes, color of hairs, digestion of food, wound healing, and the whole metabolism. Metabolism is a collection of all biochemical reactions in a body. A protein is a polymer macromolecule consisting of a linear sequence of 20 amino acids. Information about the structure of a protein is stored in a gene. A gene is a segment of a chromosome as shown in Figure2.

 Figure 2

An important question is how information stored in a gene is transformed into a protein. The process of transformation information stored in a gene consists of two separate stages. The first stage is called transcription. Transcription is a process of synthesis of a messenger RNA copy of a gene (mRNA). During transcription only one strand (coding strand) of a gene is copied (Figure 3). An RNA is a single strand molecule consisting of 4 monomers denoted by letters A, U, G, C- ribonucleotides. Let’s assume that a part of a gene coding strand consists of 5 nucleotides - AGCTA. In this case, corresponding part of mRNA is AGCUA. In mRNA U corresponds to T in DNA. Ribonucleotides (building blocks of RNA) are pretty similar to deoxyribonucleotides (building blocks of DNA).

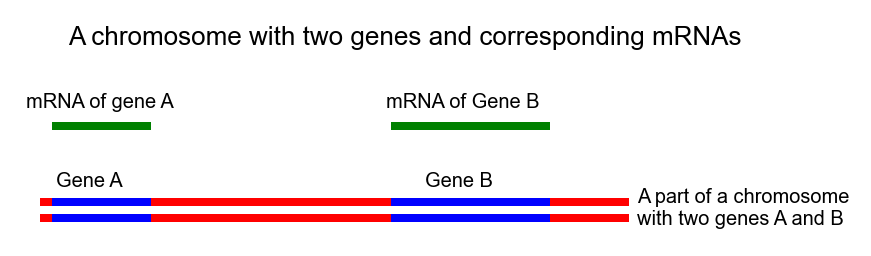


Figure 3

A mRNA is converted to a protein in a process called translation. Two processes -transcription and translation-are called gene expression. In this project only one side of gene expression-transcription will be investigated. The number of mRNA molecules of a gene can be varied substantially in different conditions. For example, tumor suppressor genes (genes opposed to cancer development) are downregulated in (the number of mRNAs molecules are reduced) in tumors, while oncogenes (tumor promoting genes) are upregulated (the number of mRNAs molecules are increased) in tumors. Generally, the more mRNA molecules of a gene persist in a cell, the more corresponding protein is synthesized. That’s why information about the number of mRNA molecules is highly important.

Next generation sequencing is a technique that allows us to compare the number mRNAs molecules for a given gene in different conditions. These conditions include- a normal tissue vs a cancer tissue, a neuron vs glia and so on. A cell transcriptome is a collection of all mRNAs molecules in a cell. Nowadays, transcriptome of a single cell can be investigated. How Next Generation Sequencing (NGS) works? It works by producing an enormous (millions) number of reads of a transcriptome (collection of all mRNAs in a cell). A read is a copy of a short portion (75-200 nucleotides in length) of an mRNA, Figure 4. The number of reads of a mRNA is directly proportional to the number of the mRNA molecules. Given the number of reads of an mRNA between conditions, relative amount of the mRNA molecules can be determined.

## 

Figure 4

Next, sequences of nucleotides of all reads are determined, and reads are quantified for each individual gene. An example of a sequence consisting of 15 nucleotides: AGGGCTTCACTTGCA. A typical read consists of 75-200 nucleotides. A tabular form containing genes as features and conditions (cells) as rows is called count matrix. Values of count matrix are reads for the corresponding gene and condition. Example is shown in Table 1.

|  |  |  |
| --- | --- | --- |
| Cells | Gene A | Gene B |
| oligodendrocyte | 22 | 4 |
| neuron | 2 | 4 |
| astrocyte | 100 | 5 |

Table 1. Count matrix

In Table 1, the count matrix for two genes A, and B, and for 3 cells-oligodendrocyte, neuron, and astrocyte is shown. It is seen from that table that gene A is up regulated (100 reads) in an astrocyte in comparison with an oligodendrocyte (22 reads) and a neuron (2 reads). The lowest expression of the gene A (2 reads) is observed in a neuron. The gene B is approximately expressed in the same level between conditions (cells). In an oligodendrocyte and a neuron 4 reads were detected, and 5 reads were detected in an astrocyte for gene B.

The goal of the current project is to investigate transcriptome of different mice single cells and to uncover gene expression signature of different cells. Gene expression signature of a cell is a set of genes whose expression (transcription) are crucial for that cell formation [1].

**Dataset**

For the current project a dataset from Tabula Muris was used [2]. Tabula Muris is a collection of single-cell transcriptome data from Mus musculus (House mouse), containing nearly 100000 cells from 20 organs and tissues. The initial dataset contains a count matrix. Seven types of cells were used for this project are Bergman glial cell, astrocyte, brain pericyte, endothelial cell, neuron, oligodendrocyte, oligodendrocyte precursor cell. Totally, 3401 cells (3401 rows) were analyzed, and 23433 genes (23433 columns) were detected.

**Methods**

1. For the data analysis the scanpy [3] library will be used. This library contains several important functions for differential gene expression analysis. Differential gene expression analysis is an analysis aiming to determine which genes are expressed (transcribed) differently between conditions. The heart of the scanpy based analysis is the annotation object. An annotation object -Ann.Data-contains 4 different things:
2. Ann.Data.X stores the count matrix as an numpy array
3. Ann.Data.obs stores the metadata as a dataframe
4. Ann.var stores the variable metadata as a dataframe
5. Ann.Data.uns stores any additional data generated during the analysis
6. Principle component analysis (PCA) will be used to reduce dimensionality of the data set. Using PCA for the dataset analysis is mandatory because it has 23433 features.
7. Uniform Manifold Approximation and Projection (UMAP) will be used to reduce dimensionality of the dataset on top of PCA
8. K-means will be used to cluster cells based on gene expression profiling on top of UMAP.
9. The Louvain method will be used to cluster cells based on count matrix data on top of UMAP.

**Analysis**

Workflow of the analysis

1. Downloading data and metadata
2. EDA of the downloaded data
3. Generation of an annotated object
4. Visualization of highly expressed genes
5. Calculation of quality control metrics
6. Visualization of some quality control metrics
7. Filtering
8. Post-filtering QC
9. Logarithmizing and scaling data
10. PCA (dimensionality reduction)
11. UMAP (dimensionality reduction) on top of PCA
12. K means on top of UMAP
13. Louvaine on top of MAP
14. Visualizing genes ranked for their importance for a cell type (dendrogram)
15. Printing genes ranked for their importance for a cell type in a tabular form.

Initial datasets for this research were downloaded locally. The first dataset contains a count matrix. The count matrix contains 3401 rows (each row corresponds to a mouse cell) and 23433 columns (each column corresponds to an individual gene). Metadata for the count matrix contains 3401 rows and 5 columns. The first column describes a class of a cell, for example astrocyte, oligodendrocyte, endothelial cell etc. The second column describes a subtissue of a cell origin. For example, from Figure 1, showing the first 5 rows of the metadata dataframe, we can figure out that an astrocyte from the first row was isolated from striatum. The third column describes mouse sex. The fourth column describes mouth.id, and the fifth column describes plate barcode.

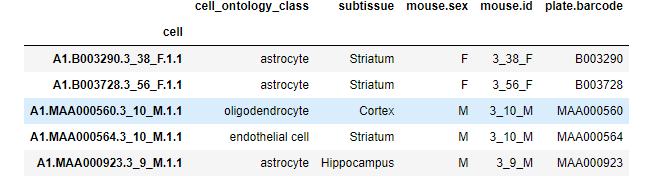


Figure 1. 5 rows of the metadata dataframe

Next, the top 20 highest expressed genes (20 top genes with the highest read count) were visualized Figure 2.

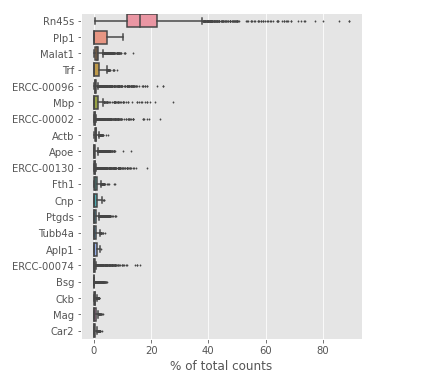


Figure 2. 20 top genes with the highest read count

To analyze the quality of the experiment, quality control metrics were generated. There are two dataframes containing control metrics. One dataframe contains control metrics for cells. Another dataframe contains control metrics for genes. Using the quality control metrics for cells a histogram of the total read counts per cell and detected genes per cell were generated, Figure 3, Figure 4

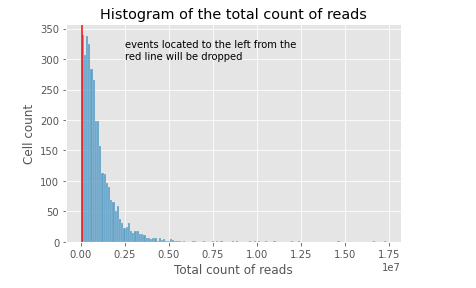


Figure 3. Histogram of the total counts of reads per cell

Cells having less than 60 thousand reads are considered as damaged and were filtered out (red line marks 60 thousand reads).

Chart, histogram

Description automatically generated

Figure 4. Histogram of the total counts of detected genes per cell

In the Figure 4, a small peak on the left side of the histogram represents cells where less than 900 genes were detected, and such cell (events) should be filtered out (red line marks 900 genes). Figure 5 shows a scatter plot of the total number of reads per cell vs total number of genes per cell. Two distributions on the top and on the right from graph are mini versions of graphs shown on Figure 3, and 4 respectively. It is clearly seen from the graph that the total number of reads per cell is proportional to the total number of detected genes per cell.

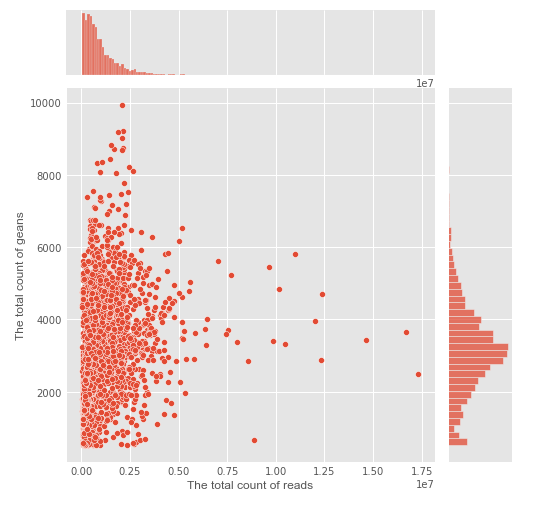


Figure 5. Scatter plot of the total number of reads per cell vs total number of genes per cell

During the experiment the spike-ins synthetic RNAs were added for quality control. Cells containing greater than 11percent of spike-ins RNA among total mRNAs will be removed from consideration as damaged. Also, genes that were found in 3 or less cells, and have read count 10 or less were filtered out as unimportant for further consideration. Quality control check of the counts post-filtering are shown in Figure 6 and 7. Figure 6 shows the violin plot of the total number of genes and the total number of read count per cell. Figure 7 shows a scatter plot of the total number of reads vs the total number of genes per cell.

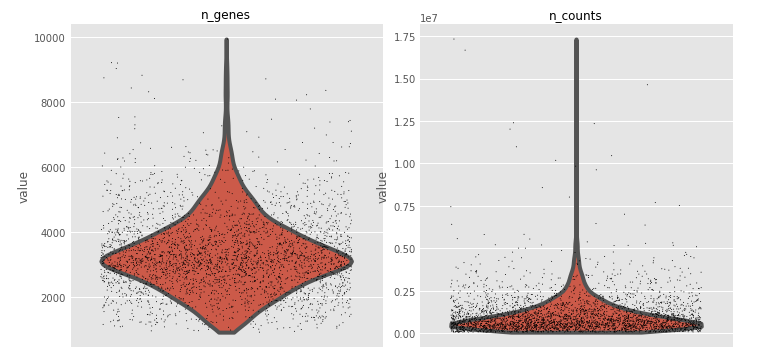


Figure 6. The violin plot of the total number of genes and the total number of read count per cell

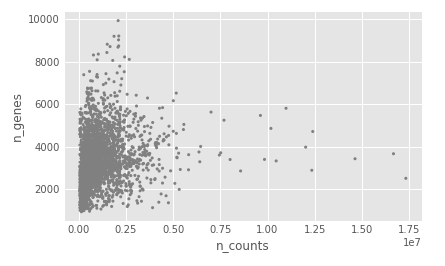


Figure 7. Scatter plot of the total number of reads vs the total number of genes per cell

After filtering cells and genes, data was log transformed and scaled and PCA was performed. The variance ratio plot (Figure 8) shows explained variance ration by each component.

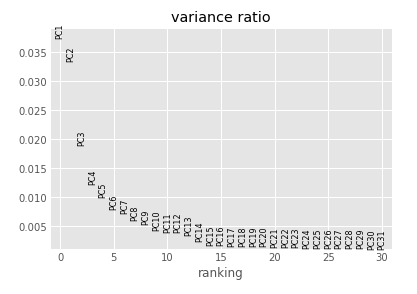


Figure 8. Explained variance ratio by each component

Figure 9 shows the most impactful genes on the first three PCA components.

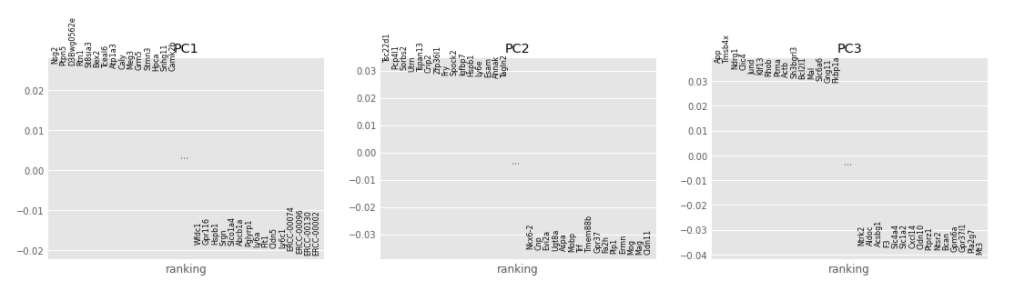
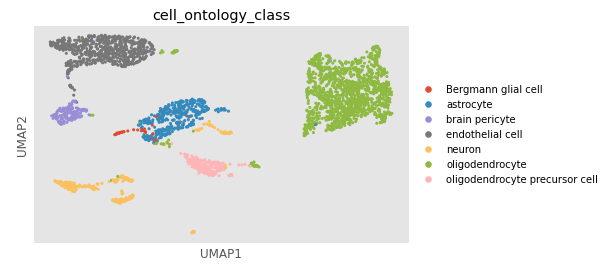


Figure 9. The most impactful genes on the corresponding PCA component

Next, Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction was performed on top of PCA. UMAP is based on the neighbor graph. The number of UMAP components is 2. All 7 types of cells used in the current research are separated well in UMAP coordinates Figure 10.

 Figure 10. Graph in UMAP coordinates of 7 different types of mouse cells

Next, K-Means algorithm was used to cluster all 3401 cells (7 different cell types) based on gene expression. K-Means was used on top of UMAP. Results of clustering are shown in Figure 11.

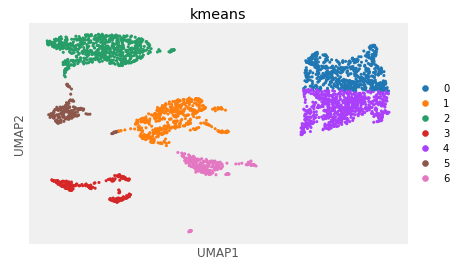


Figure 11. K-Means clustering of cells using UMAP components

Adjusted rand score for K-Means clustering is 0.65. This value shows that K-Means clustering was not effective. The same conclusion can be made by comparison Figure 10 and 11. Since K-Means algorithm was not effective enough, the Louvain clustering algorithm on top of UMAP was employed. Results are shown in Figure 12. This algorithm identified 6 clusters, but adjusted rand score is 0.91. It means that the Louvain clustering algorithm is much more effective than K-Means in clustering gene expression data.

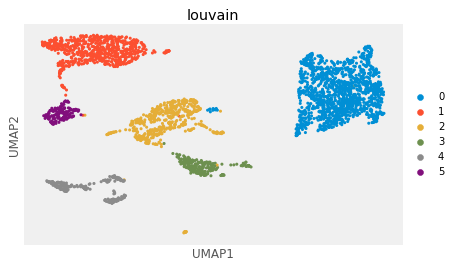


Figure 12. Louvain clustering algorithm based on UMAP 2 main components

Next, 5 of the most important genes for each type of cell were calculated and results are shown in Table 2. From the table, it is clearly seen that all 7 types of cells have a unique gene expression signature. The importance of genes listed in Table 2, should be confirmed by other experiments, for example by real-time PCR.

**Table

Description automatically generated**

Table 2

Finally, the hierarchical clustering analysis was performed and a dendrogram was generated (Figure 13). This dendrogram shows the most important (differentially expressed) genes per a group that differ from the rest of the groups. Results are the same as Table 2 shows and the only difference is that the dendrogram shows 10 most important (differentially expressed) genes per a group while Table 2 shows only 5.

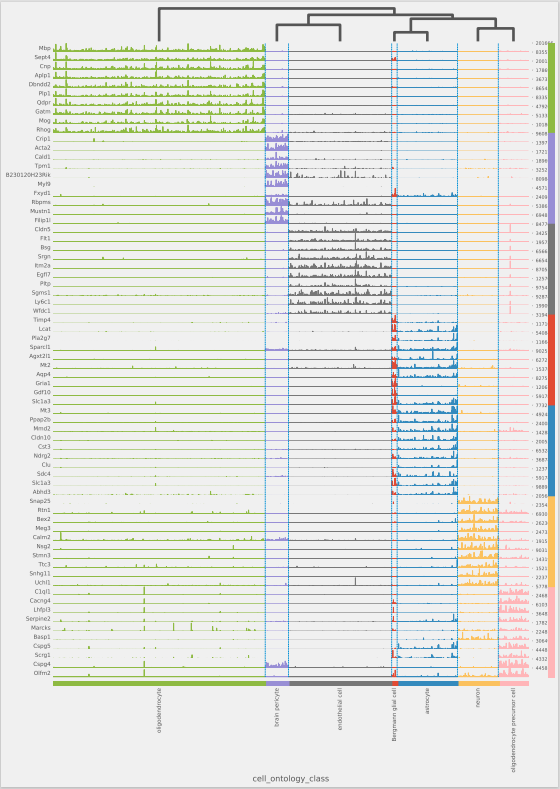
****

Figure 13. dendrogram showing differentially expressed genes for each cell type

**Conclusion**

A pipeline to analyze differentially expressed genes in 7 different mouse types of cells was set up. This pipeline effectively uncovers differentially expressed genes (the most important) in different types of cells. It was established that the Louvain clustering algorithm is much more effective in comparison with K-means in clustering data containing single cell RNAseq data.

**Assumptions**

All cells were isolated in the same conditions and the level of mRNA degradation during the procedure is the same across all cells.

**Limitations**

Limitations of the current study come from the fact that mice for this research were a similar age and a few stains of mice were investigated. Age range of mice and the number of analyzed strains should be broadened in future research.

**Challenges**

Potential issues in the project can come from the data normalization process. Simple linear scaling (CPM) was used in this project. The CMP-count per million-converts individual gene reads to count per million reads for the whole cell.

**Future Uses/Applications**

This project is aiming to uncover gene expression signatures of different types of mouse cells. Unveiling such signatures opens a broad avenue in future research and applications. Such research and applications include but are not limited to selecting genes that are responsible for senescence, cancer, heart diseases, for different organisms’ development.

**Recommendations**

Methods that were used to analyze gene expression signature of several types of mouse cells can be used to analyze gene expression signature of human cells. It will help to uncover genes involved in different diseases development including cancer and heart diseases.

**Implementation plan**

Given that the pipeline for the gene expression analysis has been established, NGS sequencing data for different diseased cells and their healthy counterparts are needed.

**Ethical Assessment**

This project is aiming to analyze differentially expressed genes in 7 different cell types of Mus musculus. The dataset does not contain any sensitive patient information and the findings of the project can be used only in a scientific field. Results of the research cannot be used to any harmful actions.

**References**

1. [Gogol-Döring](https://pubmed.ncbi.nlm.nih.gov/?term=Gogol-D%C3%B6ring+A&cauthor_id=22130885), A., & [Chen](https://pubmed.ncbi.nlm.nih.gov/?term=Chen+W&cauthor_id=22130885), W. An overview of the analysis of next generation sequencing data. (2012). Methods Mol Biol.,802,249-57.
2. https://www.czbiohub.org/tabula-muris/
3. Wolf A., Angerer, P., & Theis F. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biology, 6,19(1):15.