Project Purpose

The purpose of this project is to provide a comprehensive visualization of gene expression data in the human brain's frontal cortex tissues. Through the Shiny app, we aim to assist scientists in identifying trends, outliers, and correlations within the dataset. The visualizations serve as a tool for researchers to gain insights into the demographics of the samples, explore gene expression correlations, analyze differences between male and female samples, and investigate potential age-related patterns.

The Shiny app is designed for researchers, scientists, and data analysts working in the field of genomics and neuroscience. It caters to users who seek an interactive platform for exploring and analyzing gene expression data in the context of demographic information.

This app has support 4 main goals:

(1) Demographic Exploration: Understand the demographics of the samples, including age, gender, and Hardy Scale classification.
(2) Correlation Analysis: Investigate correlations between gene expressions to identify potential patterns and relationships.
(3) Gender Comparison: Explore differences between male and female samples based on gene expression levels.
(4) Age Segregation: Analyze whether there is any segregation between samples of different age groups.

Data Description

Data Collection

The data was downloaded from GTex Portal. The gene expression data was collected from the frontal cortex tissues of the human brain. The sample ID was encoded and linked with demographic information such as age, gender, and Hardy Scale classification.

Questions

What are the demographics of the samples?
What are the correlations between gene expressions?
Is there a difference between male and female samples?
Is there any segregation between samples of different age groups?

Insights

The exploration of the gene expression data yields valuable insights into various facets of the sample population. The demographic information encapsulates a comprehensive overview, providing a holistic understanding of the characteristics of the individuals represented in the dataset. From the demographic graph, it is easily noticeable that the largest population group in the dataset is male from 60 to 69-year-old with fast death of natural causes.

Delving into gene expression correlations uncovers potential relationships between genes. As the genes are sorted based on their locations in the human's genomes, we can view that location closeness does affect the expression correlation. Also, as the visualization allow users to zoom in, we can see the special, outliers case in the region.

Additionally, the visualizations effectively highlight discernible patterns related to gender and age, enabling researchers to glean meaningful observations and further refine their investigations into the nuanced factors influencing gene expression in this context. In both graphs, there is significant overlapping areas between the ellipses, suggesting no segregation between the groups.

Reproducibility Process

To ensure reproducibility, a structured process was followed:

Data Cleaning: The raw data underwent cleaning and transformation steps in "datacleaning.rmd".

Exploratory Data Analysis: Initial explorations were conducted to understand the characteristics of the dataset, and preliminary plots are attached to the end of "datacleaning.rmd".

Shiny App Development and Visualization Techniques: The Shiny app was developed using R. Visualization techniques such as PCA, heatmaps, and scatterplots, were employed with the support of related packages. The code is in "geneviz.r". All the code has documentation and distinctives variable names for easy understanding

Design Decisions

Demographic Summary Tab

What –

Demographic information is encoded using a balloon plot, showcasing the distribution of age and Hardy Scale classification across genders.

Why –

Enables users to quickly grasp the demographic composition of the sample population.

How –

Balloon Plot: Utilizes ggplot2 to create a visually appealing and informative representation of demographic data, enhancing interpretability.

Expression Correlation Tab

What –

Visual Representation: Heatmaply_cor function is employed to create an interactive heatmap representing gene expression correlations.

Why -

Correlation Analysis: Offers an intuitive way to identify patterns and relationships between gene expressions.

How -

Heatmaply_cor Function: Leverages the heatmaply_cor function with specific parameters to enhance clarity and interactivity in the heatmap.

Gender Groups Tab

What:

Visual Representation: A 3D scatterplot is generated to display gene expressions among male and female samples, incorporating confidence ellipses.

Why:

Gender Comparison: Facilitates the comparison of gene expressions between genders in a three-dimensional space.

How:

Scatter3D Plot: Utilizes the canvasXpress library to create an interactive 3D scatterplot, allowing users to explore gene expressions in a visually engaging manner.

Age Groups Tab

What: Principal Component Analysis (PCA) is applied to reduce dimensionality and visualize sample segregation based on age.

Why: Allows users to identify any segregation patterns between samples of different age groups.

How: Employs the fviz_pca_ind function from the ggpubr package to generate PCA plots with overlapping confidence ellipses, aiding in the interpretation of age-related patterns.

Interactive Elements

What:

Interactive Controls: Users can interact with the visualizations by hiding/unhiding specific groups, zooming in on selected regions, and hovering over data points for detailed information.

Why:

Interactivity enhances user engagement and facilitates a more in-depth exploration of the dataset.

How:

Leverages the Shiny Plotly and canvasXress framework to implement reactive elements and controls, allowing users to dynamically interact with the visualizations.

Improvement Wishlist

- One thing I have not been able to implement is representing the genes in a more simplified way.
- Secondly, I hope to have more control over the R built-in function to better personalized the graph.

References

Broad Institute of MIT and Harvard (2021). Download Open Access Datasets. https://www.gtexportal.org/home/downloads/adult-gtex/overview

Appendix

Appendix 1: Time/Tasks Log Chart