MATH 7343

Applied Statistics

PROJECT REPORT

Group B

RNA SEQUENCING

STATISTICAL DATA ANALYSIS

By

ABHILASHA JAIN

AILIN DOLSON-FAZIO

MUKUND KUDLUGI

NIKHIL ANAND

PRATHAMESH THITE

SPRING 2023

NORTHEASTERN UNIVERSITY, BOSTON

Introduction

RNA-Sequencing is a topic in Functional Genomics that uses next-generation sequencing to analyze the transcriptome of a particular cell. Transcriptome is a biology term used to describe the types of mRNA used to create proteins. Next-generation sequencing, or high throughput sequencing, is a piece of technology that most efficiently sequences DNA and RNA.

RNA-seq is a powerful technique for analyzing gene expression and separation or classification of different samples. The advantage of RNA-seq over other methods, such as microarray, is its ability to detect novel transcripts, gene fusions, single nucleotide variants, indels and any other new changes found in RNA. RNA-seq data is high dimensional, and it requires statistical analysis to identify differentially expressed genes, transcripts, and other biological features. The aim of this project is to develop a statistical analysis pipeline to identify differentially expressed genes and transcripts from RNA-seq data.

The goal of the project is to perform a statistical analysis of RNA sequencing data and to identify potential biomarkers, pathways, and gene signatures associated with aging, dementia, and traumatic brain injury (TBI) using various statistical techniques such as differential expression analysis, functional enrichment analysis, correlation analysis, clustering analysis, and diagnostic performance analysis using R. Ultimately, the results of this project can provide insights into the underlying biological mechanisms that are differentially regulated between different conditions and potentially lead to the development of new diagnostic and therapeutic strategies.

The dataset used is from a long-term open-source study on aging, dementia, and Traumatic Brain Injury (TBI) called “Adult Changes in Thought” (ACT). The dataset was curated by the University of Washington, Kaiser Permanente Health Research Institute, and the Allen Institute for Brain Science.

Methods and Methodology

RNA-Sequencing data for this project is sourced from publicly available databases and quality of the data was assessed using FastQC. The expression data was normalized using the trimmed mean of M values (TMM) method, and differential gene expression analysis was performed using DESeq2 package. Pathway enrichment analysis and Gene Ontology analysis was conducted using the clusterProfiler package along with genome wide annotation for human database.

The chosen dataset was from a long-term Aging, Dementia, and Traumatic Brain Injury (TBI) study (Zorzetto, 2022). It contains detailed neuropathologic, molecular, and transcriptomic characteristics of brains suffering from aging, dementia, and injury. The dataset analyzes documents various gene expressions and biomarkers dependent on sex, cortex, grey matter, white matter, inflammation and whether the patient has dementia, traumatic brain injury or both.

Differential Expression Analysis

A. Overview

Packages such as DESeq2 and edgeR can be used to perform differential expression analysis, which provide statistical methods for identifying genes that show significant changes in expression levels between different conditions or groups. The analysis typically involves estimating fold changes and p-values, which represent the magnitude of expression and statistical significance of the changes respectively.

B. Model

The number of reads in sample *j* that are assigned to gene *i* can be modeled by a negative binomial (NB) distribution.

$$K_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2)$$

The negative binomial (NB) distribution is frequently employed for count data modeling in the presence of overdispersion, using two parameters: the mean μ_{ij} and the variance σ_{ij}^2. The read counts K_{ij} are represented as non-negative integers.

In real-world scenarios, the parameters μ_{ij} and σ_{ij}^2 are unknown and require estimation from the available data. However, due to the limited number of replicates, additional assumptions and modeling are necessary to obtain reliable estimates.

Firstly, the mean parameter μ_{ij} - which represents the anticipated value of gene *i's* observed counts in sample j - is a function of two factors: a per-gene value $q_{i,\rho(j)}$that depends on the experimental condition $\rho(j)$ of sample j , and a size factor s_j.

$$\mu_{ij} = q_{i,\rho(j)} S_j$$

$q_{i,\rho(j)}$is proportional to the expectation value of the true (but unknown) concentration of fragments from gene i under condition $\rho(j)$. The size factor represents the coverage, or sampling depth, of library j , and we will use the term common scale for quantities, such as $q_{i,\rho(j)}$, that are adjusted for coverage by dividing by s_j.

Second, the variance σ_{ij}^2 is the sum of a shot noise term and a raw variance term,

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_{i,\rho(j)}$$

Third, we assume that the per-gene raw variance parameter $v_{i,\rho}$ is a smooth function of $q_{i,\rho}$,

$$y_{i,\rho(j)} = y_\rho(q_{i,\rho(j)})$$

This assumption is necessary as the limited number of replicates makes it challenging to obtain a reliable estimate of the variance for gene *i* solely based on its available data. By making this assumption, we can combine the data from genes that exhibit similar expression levels, thereby improving our ability to estimate the variance.

C. Fitting the model

Let us now elaborate on how the model can be applied to the given data. The dataset consists of an n × m table that contains the counts, k_{ij}, where $i = 1,...,$ n represents the genes, and $j = 1,...,$ m represents the samples. The model incorporates three parameter sets:

i) A set of m size factors, s_j. The expected count values for all samples in j are proportional to s_j.

ii) For each experimental condition ρ, a set of n expression strength parameters, $q_{i\rho}$. These parameters represent the anticipated abundance of fragments from gene i under condition ρ, such that the expected count values for gene i are proportional to $q_{i\rho}$.

iii) The smooth functions $v_\rho : \mathbb{R}+ \rightarrow \mathbb{R}+$; for each condition ρ, v_ρ models the dependence of the raw variance $v_{i\rho}$ on the expected mean $q_{i\rho}$.

Size factors (s_j) are implemented to enable comparability between counts from various samples, which may have undergone sequencing at varying depths. Consequently, the expected count ratios $(EK_{ij})/(EK_{ij'})$ for the same gene i in different samples j and j' should correspond to the size ratio $s_j/s_{j'}$ if gene i is not differentially expressed or samples j and j' are replicates. While the total number of reads, $\Sigma_i k_{ij}$, may appear to be a sensible choice for s_j and a measure of sequencing depth, empirical observations suggest otherwise. This is because a few highly and differentially expressed genes can exert a strong influence on the total read count, thereby leading to a ratio of total read counts that does not accurately estimate the expected count ratio.

Hence, to estimate the size factors, we take the median of the ratios of observed counts. Generalizing the procedure just outlined to the case of more than two samples, we use:

$$\widehat{s_j} = \frac{median_i \cdot k_{ij}}{(\Pi_{v=1} m \cdot k_{iv})^{\frac{1}{m}}}$$

The pseudo-reference sample, obtained by calculating the geometric mean across all samples, is represented by the denominator of the given expression. This allows for the estimation of each size factor ($\hat{s}_j$) by computing the median of the ratios of the counts from the j-th sample to those of the pseudo-reference.

To estimate $q_{i\rho}$, we use the average of the counts from the samples j corresponding to condition ρ, transformed to the common scale:

$$\widehat{q_{ip}} = \frac{1}{m_p} \sum_{\rho(j)=\rho}^{j} \frac{k_{ij}}{\widehat{s_j}}$$

where m_ρ is the number of replicates of condition ρ and the sum runs over these replicates. the functions v_ρ, we first calculate sample variances on the common scale

$$w_{ip} = \frac{1}{(m_\rho - 1)} \sum_{\rho(j)=\rho}^{j} \left(\frac{k_{ij}}{\widehat{s_j} - \widehat{q_{i\rho}}} \right)^2$$

and define

$$z_{ip} = \frac{\widehat{q_{i\rho}}}{(m_\rho - 1)} \sum_{\rho(j)=\rho}^{j} \left(\frac{1}{\widehat{s_j}} \right)$$

In cases where there are few replicates, m_ρ - as is commonly encountered in practical applications - the values of $w_{i\rho}$ can be exceedingly inconsistent, resulting in an inadequate variance estimator for statistical inference if $w_{i\rho}$ - $z_{i\rho}$ were used. Instead, we utilize local regression on the graph $(\widehat{q_{i\rho}}, w_{i\rho})$ to derive a smooth function $w_\rho(q)$, with

$$\widehat{y_\rho}(\widehat{\rho_{ip}}) = w_\rho(\widehat{q_{i\rho}}) - z_{i\rho}$$

as our estimate for the raw variance.

To prevent estimation biases in the local regression, it is essential to exercise caution. The residuals $w_{i\rho} - w(\hat{q}_{i\rho})$ are skewed because $w_{i\rho}$ is a sum of squared random variables. To address this issue, we employ a generalized linear model of the gamma family for the local regression.

D. Testing

Assume we have m_A and m_B replicate samples for biological conditions A and B, respectively. Our aim is to evaluate the evidence in the data for differential expression of each gene i between the two conditions by testing the null hypothesis $q_{iA} = q_{iB}$. We use the total counts in each condition as a test statistic for this purpose.

$$K_{iA} = \sum_{j:\rho(j)=A}^{A} K_{ij}, K_{iB} = \sum_{j:\rho(j)=B}^{B} K_{ij}$$

and their overall sum $K_{iS} = K_{iA} + K_{iB}$. We can compute the probabilities of $K_{iA} = a$ and $K_{iB} = b$ for any pair of numbers a and b under the null hypothesis, using the error model described in the previous section. These probabilities are denoted by $p(a, b)$. To calculate the P value for an observed count pair (k_{iA}, k_{iB}), we sum up all the probabilities that are less than or equal to $p(k_{iA}, k_{iB})$, given that the overall sum is k_{iS}, given that the overall sum is k_{iS} :

$$p_{i=\sum_{p(a,b)\le p(k_iA,k_iB)}^{a+b=k_iS}} \frac{p(a,b)}{\sum_{a+b=k_iS} p(a,b)}$$

The values of the variables a and b range from 0 to k_{iS} in the aforementioned sums. This method is similar to the approach taken in other conditioned tests, such as Fisher's exact test.

To compute *p(a, b)* under the null hypothesis, we assume that counts from different samples are independent, then calculate the probability of the event $K_{iA} = a$ and $K_{iB} = b$. To do this, we need to compute the probability of the random variable K_{iA}=*sum of* m_A...

We approximate the distribution of NB-distributed random variables by a NB distribution. We obtain the parameters of this distribution from those of the K_{ij}. We compute the pooled mean estimate from the counts of both conditions before obtaining these parameters.

$$\widehat{q_{\iota 0}} = \sum_{\rho(j)\in\{A,B\}}^{j} \frac{k_{ij}}{s_j}$$

which accounts for the fact that the null hypothesis stipulates that $q_{iA} = q_{iB}$. The summed mean and variance for condition A are

$$\widehat{\mu_{\iota A}} = \sum_{j\in A} sj\widehat{q_{\iota 0}}$$

$$\sigma_{iA}^2 = \sum_{j\in A} \quad \widehat{s_j}\widehat{q_{\iota 0}} + \hat{s}j2v^{A(\widehat{q_{\iota 0}})}$$

Functional Enrichment Analysis

A. Overview

Identification of differentially expressed genes is followed by functional enrichment analysis to gain insights into the biological functions and pathways that may be affected. Gene Ontology (GO) terms and pathways that are overrepresented among the differentially expressed genes can provide clues about the biological processes or pathways that may be dysregulated in the conditions of interest. Tools such as Gene Set Enrichment Analysis (GSEA), clusterProfiler are available in R.

B. Procedure

i) The algorithm ranks the genes based on log2 fold-change(LFC).
ii) The upregulated (LFC > 1) and downregulated (LFC < -1) genes are extracted.
iii) The significant genes with an absolute log-fold change greater than 1 and a p-value less than 0.05 are filtered and obtained to visualize the data.
iv) KEGG pathway enrichment analysis is performed using the "enrichKEGG" function in R.

Overall GSEA's statistical and mathematical approach allows researchers to identify pathways that are significantly enriched or depleted in each condition, and interpret the biological processes involved in that condition.

Gene Ontology and Pathway Analysis

A. Overview

Functional Enrichment analysis is then followed by Gene Ontology (GO) analysis which uses statistical methods to identify overrepresented biological processes, cellular components, and molecular functions in a set of genes of interest compared to a reference set. Further analysis can be performed using DAVID bioinformatics resources or other similar tools to perform gene ontology and pathway analysis. The mathematical representation of GO analysis involves the use of probability distributions, such as the hypergeometric and binomial distributions, to calculate the statistical significance of enrichment. This additional analysis can help in interpreting the results in the context of existing biological knowledge and literature and provide further insights into the potential functional implications of the gene expression changes observed in the study.

B. Model

Assuming a set of genes of interest with size n, and a reference set with size N. Let k represent the number of genes in the set of interest that are annotated to a specific GO term (biological processes), and K represent the total number of genes annotated to that term in the reference set. The probability of observing k or more genes in the set of interest annotated to the GO term by chance can be calculated using the hypergeometric distribution:

$$P(k\ or\ more) = 1 - \sum_{i=0}^{k-1} K_{C_i} \cdot \frac{N - K_{c_{n-i}}}{N_{C_n}}$$

where *Ci* represents the number of ways of choosing i genes from K genes annotated to the GO term, and ${}^{N}C_{n}$ represents the total number of possible ways of selecting n genes from *N* genes.

The p-value is calculated as the probability of observing k or more genes annotated to the GO term (biological processes), given the null hypothesis that the genes in the set of interest are randomly selected from the reference set.

After calculating the p-value, we calculate the q-value. The q-value represents the false discovery rate (FDR) corrected p-value, which accounts for the multiple hypothesis testing that is inherent in GO analysis. The q-value for each GO term is calculated using Benjamini-Hochberg procedure. Benjamini-Hochberg procedure can be explained by:

i) Arrange all the obtained p-values in ascending order.
ii) If there are 'm' number of genes provided in the differentially expressed dataset and 'i' ranges from 1 to m, then the critical value *'k'* for a threshold value of 0.05 can be calculated by:

$$k = max(i),\ such\ that\ P(i) \leq \frac{1}{m} \cdot 0.05$$

In other words, k is the largest value of *i*, such that the corresponding value *p(i)* is less than or equal to the adjusted threshold *1/m * 0.05*.

1. The q-values for each p-value *p(i)* is calculated for every *'i'* less than k using the formula:

$$q(i) = \min\left\{p(j) \cdot \frac{m}{j}\right\} \textit{ for all } j=1,2,\ldots m$$

In other words, *q(i)* is the minimum FDR at which *p(i)* is deemed significant based on all tests with smaller or equal p-values.

C. Procedure:

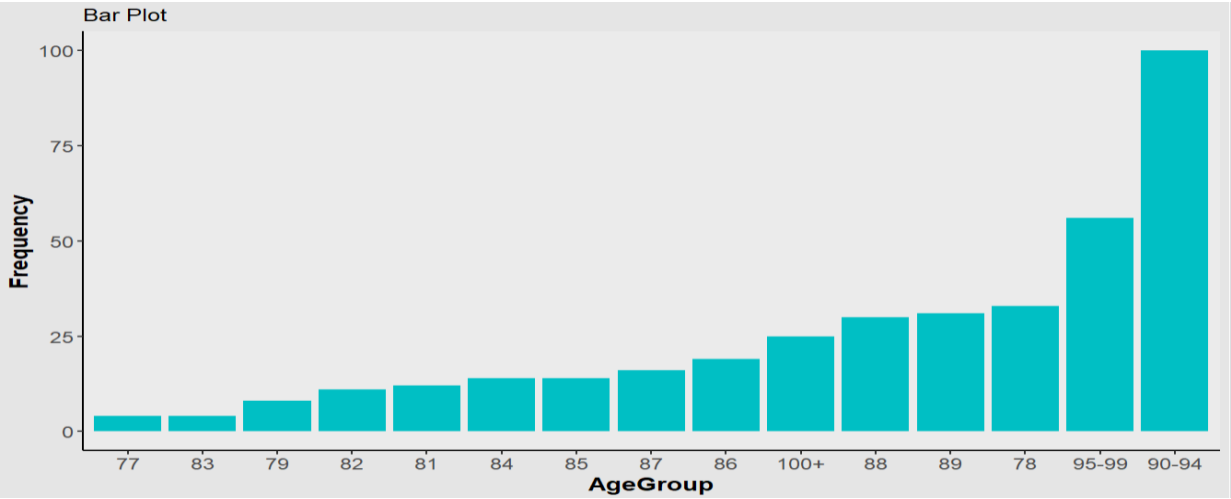
Gene Ontology Enrichment Analysis was performed using enrichGO function from clusterProfiler package in R.

i) A set of differentially expressed genes obtained from Differential Expression analysis is used for further analysis.
ii) Genes were annotated with "biological processes" Gene Ontology terms in human genes database. This step assigns functional annotations to each gene in the list using the available GO annotation files.
iii) Enrichment analysis: EnrichGO first calculates the hypergeometric test statistic for each GO term in the database, using the differentially expressed gene set provided and the reference set of genes from the human genes database which gives p-values for each GO term. The q-values are calculated using Benjamini-Hochberg procedure. Based on the given threshold of 0.05 for both p-value and q-value, a list of enriched genes was obtained.
iv) Visualization: Finally, the enriched set of genes are plotted on a bar chart or dot charts.

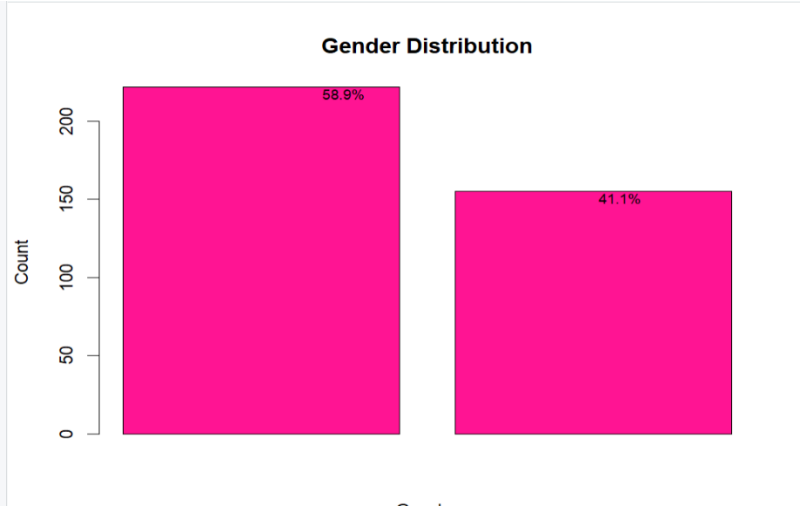
Visualization and Interpretation of Results

To better understand and communicate the findings of the analysis, the results are visualized using various plots such as volcano plots, heatmaps, and gene expression plots. These visualizations can provide a graphical representation of the gene expression changes and facilitate the interpretation of the results. Finally, the results can be summarized in the form of figures, tables, and text, and interpreted in the context of biological knowledge and literature. This interpretation can provide insights into the potential biological mechanisms.

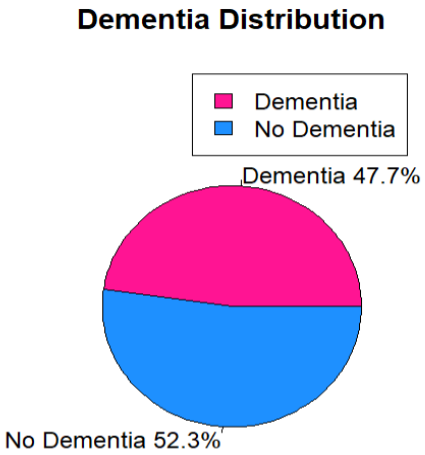
Aging Bar Graph:



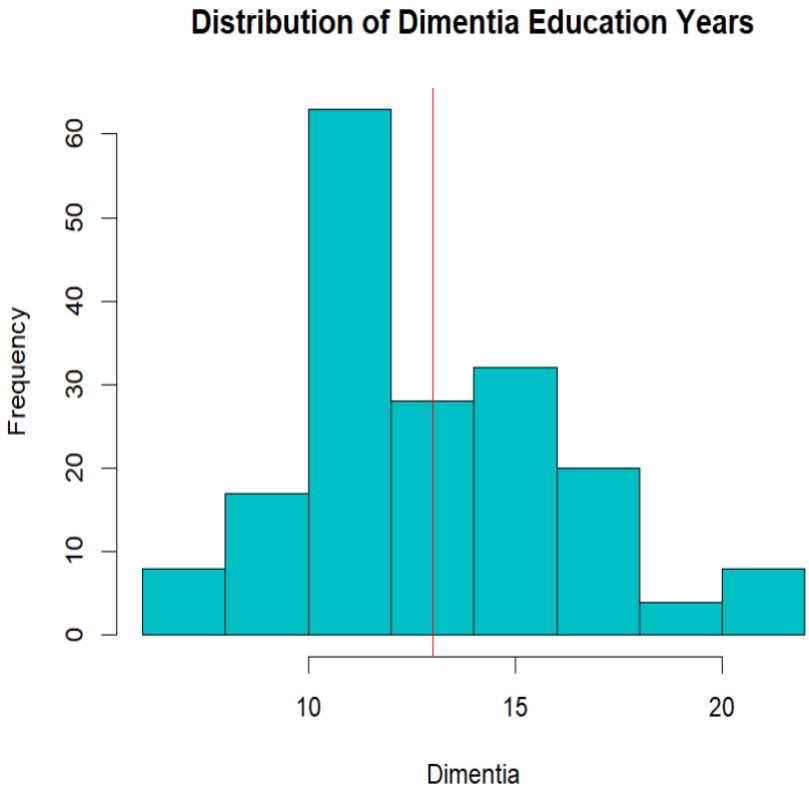
Gender Distribution:



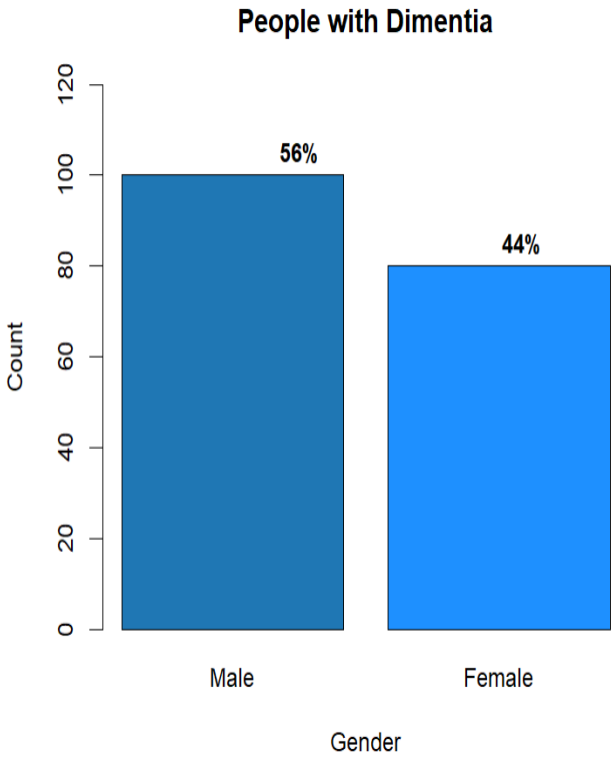
Dementia Distribution:



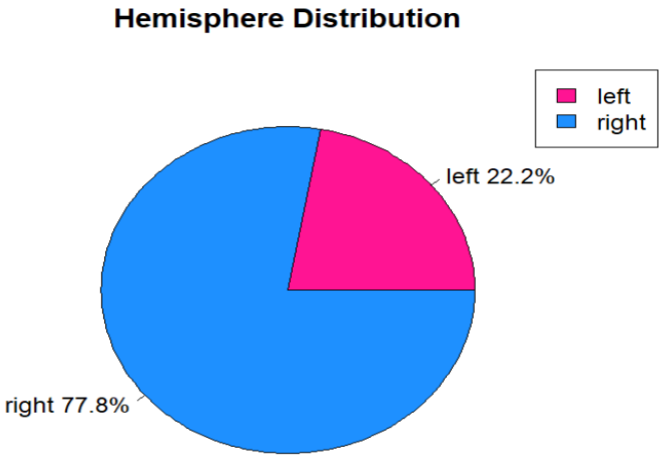
Distribution of Dementia Education Years:



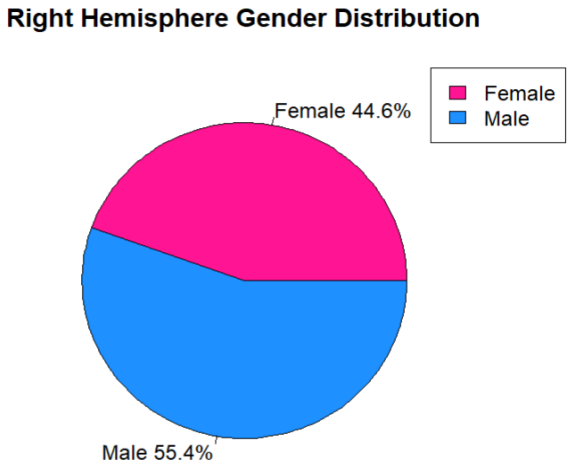
People With Dementia:



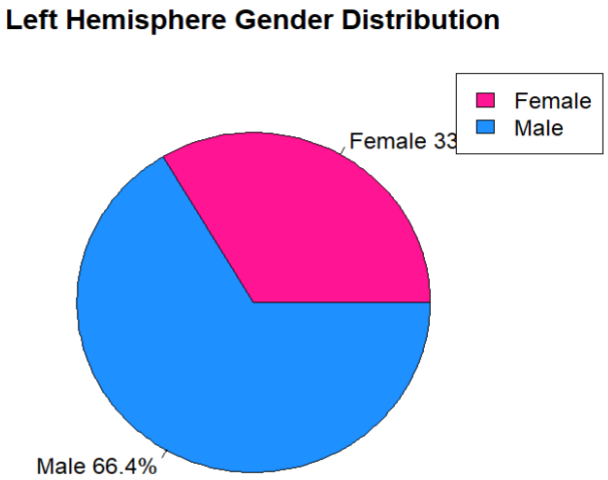
Hemisphere Distribution:



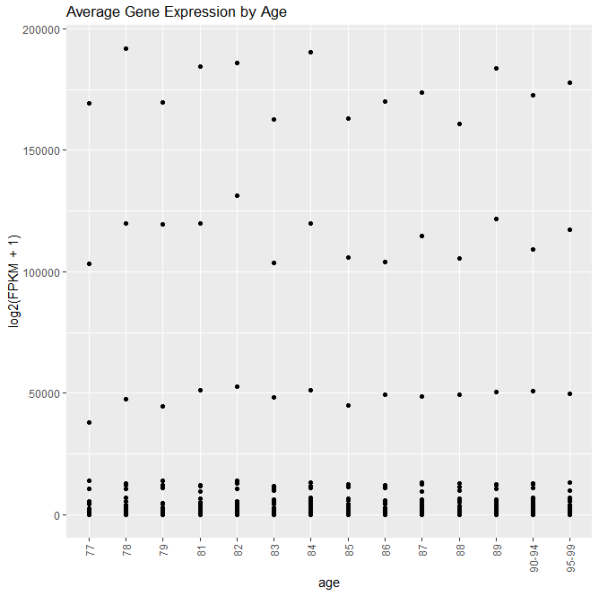
Right Hemisphere Gender Distribution:



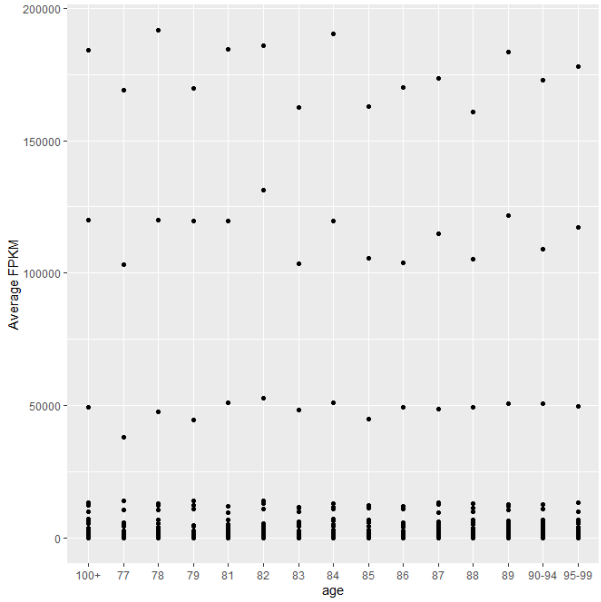
Left Hemisphere Gender Distribution:



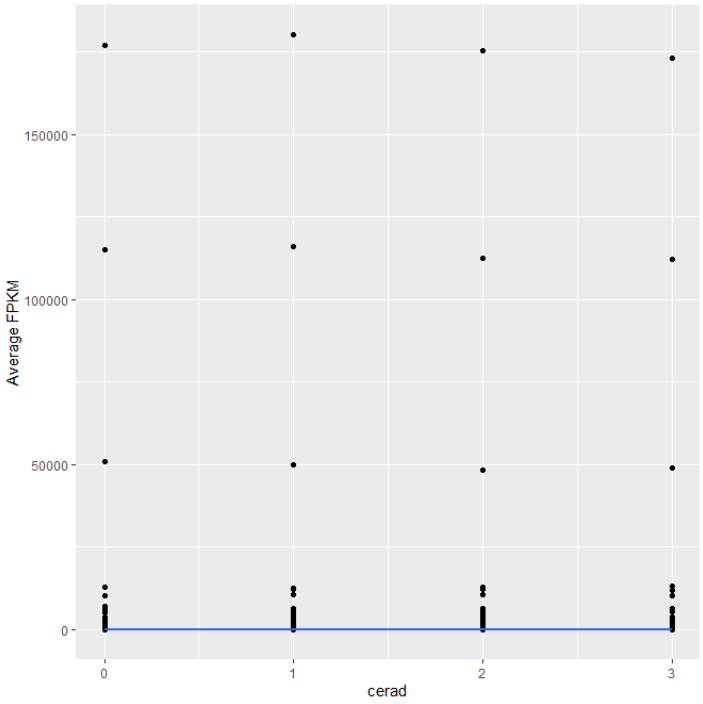
Average Gene Expression by Age:



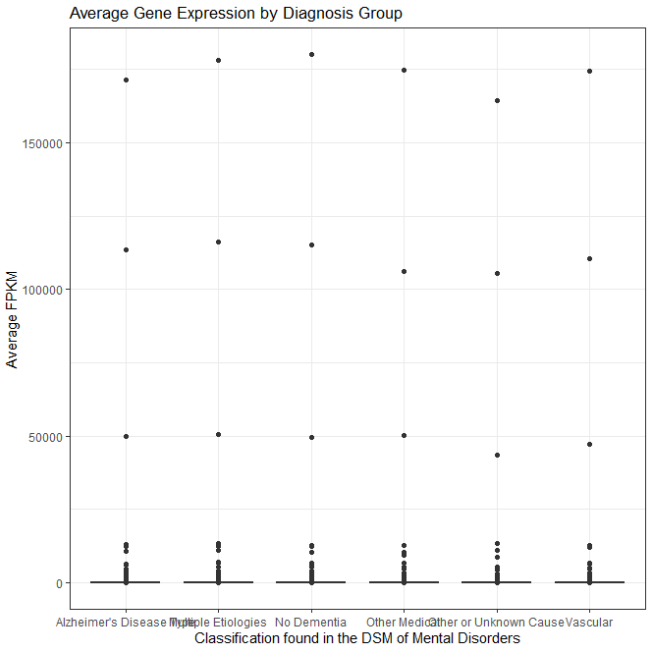
Aging Mean Melted:



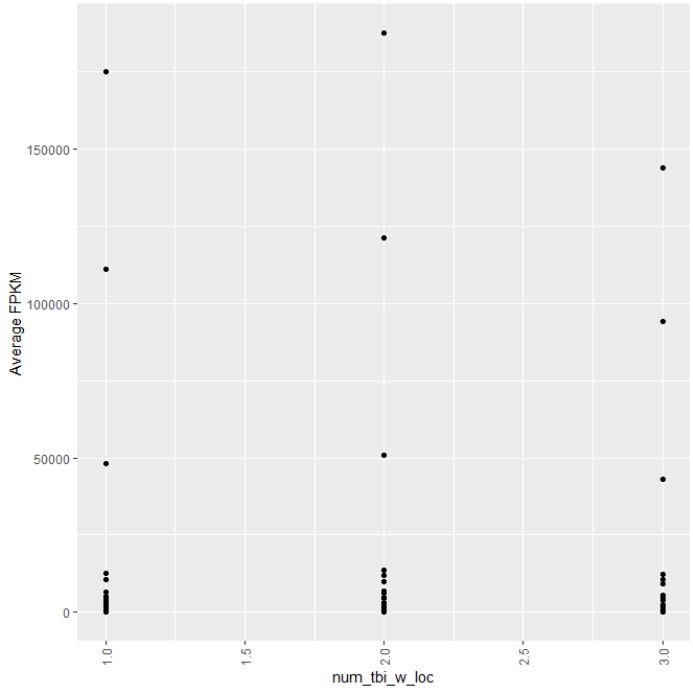
Cerad Mean Melted:



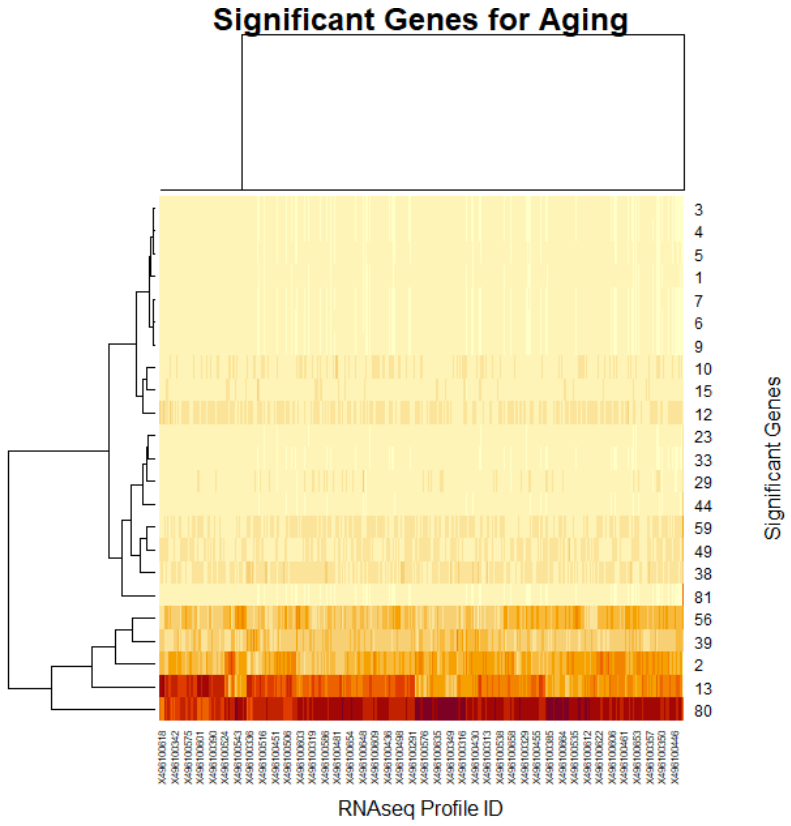
Average Gene Expression by Diagnosis:



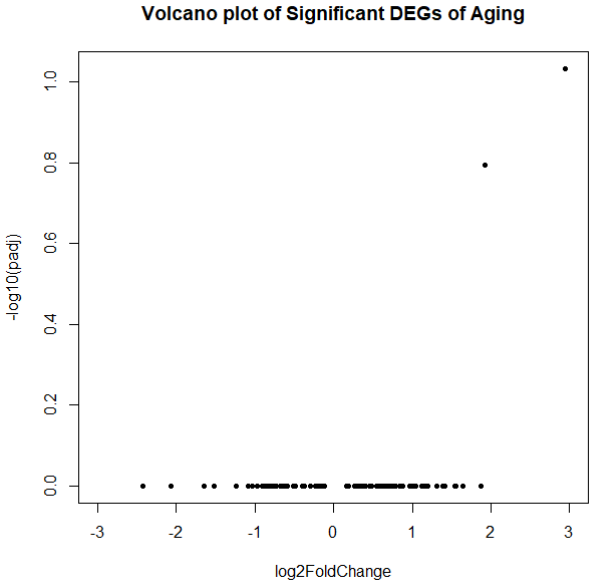
TBI Mean Melted:

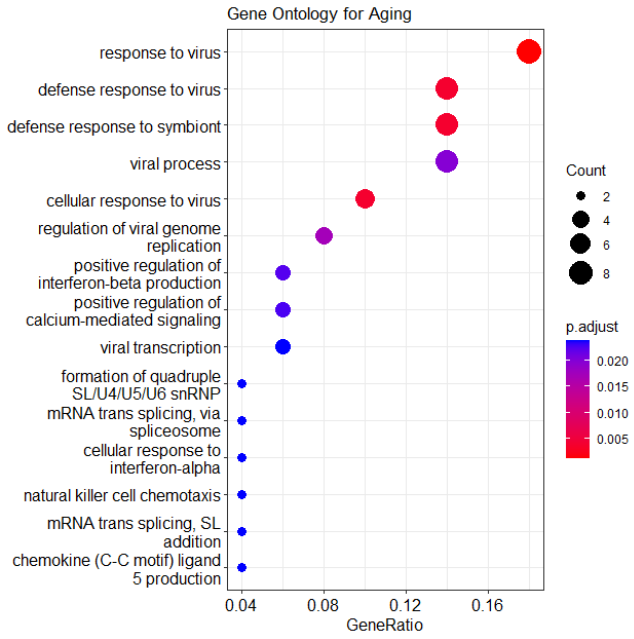


Significant Genes for Aging:

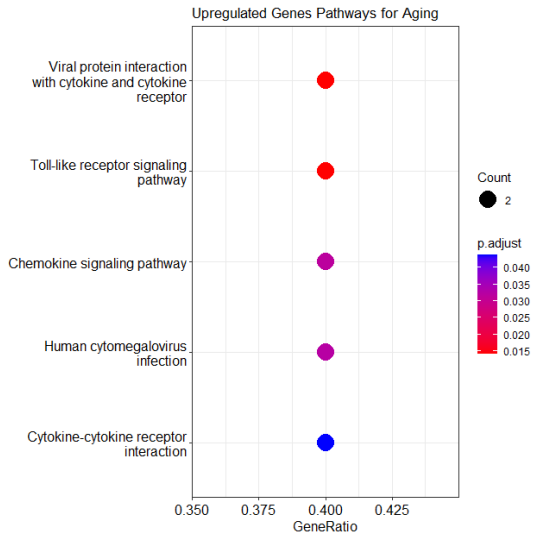


Volcano Plot for Significant DEGs of Aging:

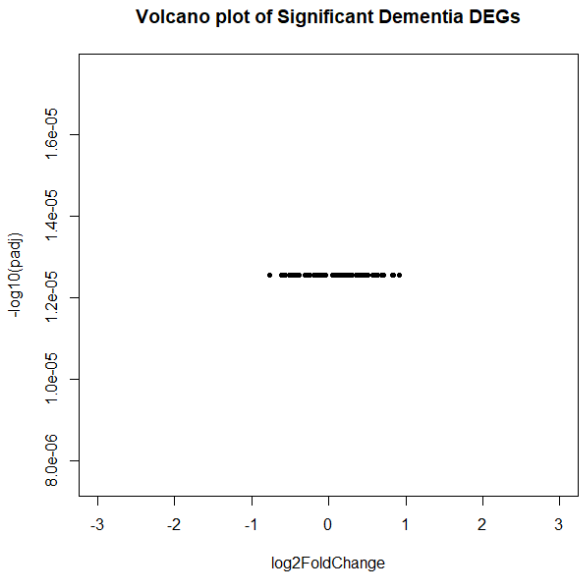




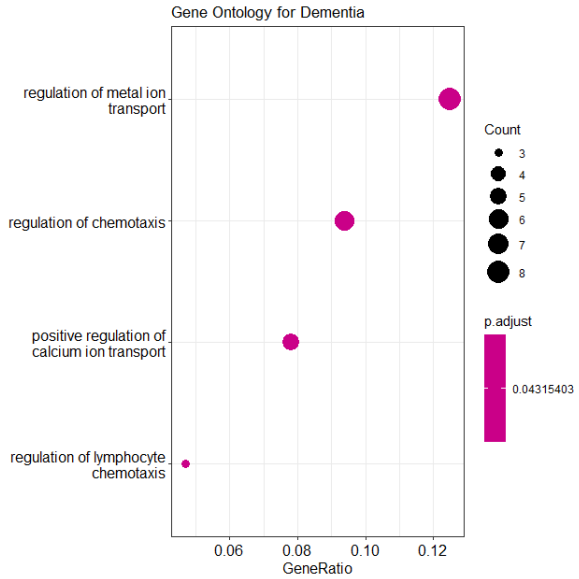
Upregulated Genes Pathways for Aging:



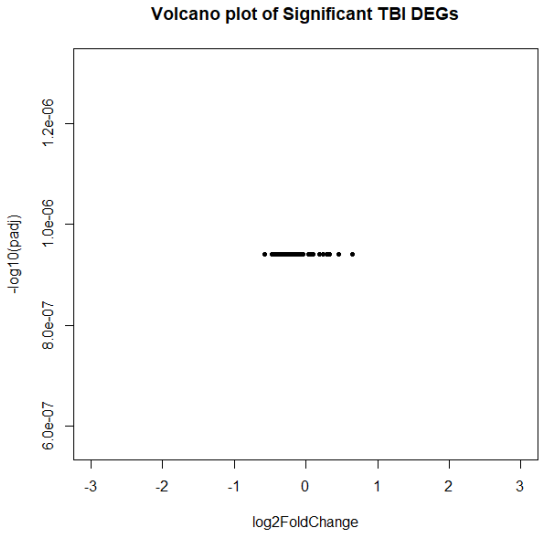
Volcano Plot of Significant Dementia DEGs:



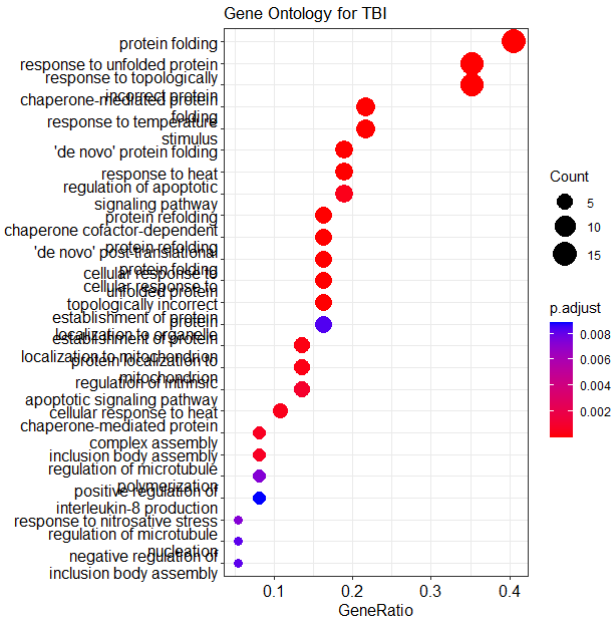
Gene Ontology for Dementia:



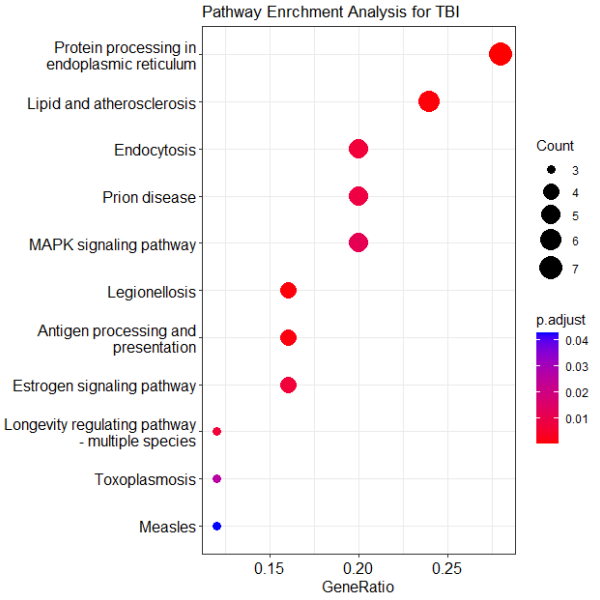
Functional Enrichment Analysis for TBI:



Gene Ontology for TBI:



Pathway Enrichment Analysis for TBI:



Results

For aging, a total of 88 DEGs were identified, with 15 upregulated and eight downregulated. Pathway enrichment analysis revealed that the upregulated genes were enriched in pathways related to viral infection and immune response. GO Analysis suggested that the DEGs were significantly enriched in biological processes related to response to virus and defense response to virus and symbiont, among other terms.

For dementia, a total of 104 DEGs were identified, with no upregulated and no downregulated. Pathway enrichment analysis suggests that none of the KEGG pathways are significantly overrepresented in the input gene list compared to the background set. GO Analysis suggests DEGs are enriched in 4 terms, that are related to positive regulation of calcium ion transport, regulation of metal ion transport, regulation of lymphocyte chemotaxis, and regulation of chemotaxis, which suggests that the input genes may be involved in cellular transport and chemotaxis processes.

For TBI, a total of 66 DEGs were identified, with no upregulated and no downregulated. Pathway enrichment analysis revealed that given DEGs are enriched in 11 KEGG pathways which are involved in various biological processes, such as protein processing in endoplasmic reticulum, antigen processing and presentation, and lipid metabolism. GO Analysis showed 111 enriched terms were found with respect to the 66 DEGs. The over representation of the genes suggests that these 66 DEGs for TBI are significantly enriched in functions related to protein folding, response to unfolded protein, response to topologically incorrect protein, chaperone-mediated protein folding and other biological processes.

Discussion

Based on the analysis performed in this project, it appears that there are no common pathways enriched across aging, dementia, and TBI. For aging, the upregulated genes are enriched in pathways related to viral infection and immune response, while for dementia, the input genes may be involved in cellular transport and chemotaxis processes. For TBI, the DEGs are enriched in KEGG pathways that are involved in various biological processes, such as protein processing in endoplasmic reticulum, antigen processing and presentation, and lipid metabolism, and the Gene Ontology Analysis suggests that the genes are significantly enriched in functions related to protein folding, response to unfolded protein, and chaperone-mediated protein folding, among others.

Therefore, it can be concluded that RNA sequencing data analysis provided insights into the DEGs and their associated biological pathways and functions in aging, dementia, and TBI. The results suggest that viral infection and immune response pathways are involved in aging, cellular transport and chemotaxis processes are associated with dementia, and protein processing and metabolism pathways are involved in TBI. But the pathways and biological processes enriched in each of these conditions are different, and there are no common pathways or processes that are enriched across all three conditions. The study provides a framework for further investigation of the underlying molecular mechanisms of these conditions which may reveal additional insights.

References

Akalin, A. (2020). RNA-Seq Analysis. In A. Akalin, *Compgenomr* *https://compgenomr.github.io/book/rnaseqanalysis.html* (p. Chapter 8). Online: Github.

Anders, S. H. (2010). Differential expression analysis for sequence count data. *Genome Biol 11*, R106. Retrieved from https://doi.org/10.1186/gb-2010-11-10-r106

Li, D. (2019). Statistical Methods for RNA Sequencing Data Analysis. In D. S. Editor Holger Husi, *Computational Biology [Internet] https://www.ncbi.nlm.nih.gov/books/NBK550334/* (pp. 85 - 100). Brisbane (AU): Codon Publications.

Maria Doyle, B. P. (2016, May 11). RNA-Seq Analysis in R. *Bioinformatics Core Shared Training https://bioinformatics-core-shared-training.github.io/RNAseq-R/*. Carlton, Australia: University of Cambridge.

RNAseq_DE_analysis_with_R. (n.d.). Retrieved from RNAseq-DE-analysis-with-R: https://monashbioinformaticsplatform.github.io/RNAseq-DE-analysis-with-R/RNAseq_DE_analysis_with_R.html

Rue-Albrecht, K. M. (2016). GOexpress. *BMC Bioinformatics 17*, 126.

Wu T, H. E. (2021 Jul). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb).*, 1;2(3):100141.

Yunshun Chen, A. T. (2014). Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR. In D. S. Somnath Datta, *Statistical Analysis of Next Generation Sequence Data https://www.researchgate.net/publication/263583950_Differential_Expression_Analysis_of_Complex_RNA-seq_Experiments_Using_edgeR.* Springer.

Zorzetto, A. (2022, March). *[RNAseq] Aging, Dementia and TBI*. Retrieved from Kaggle: https://www.kaggle.com/datasets/albertozorzetto/rnaseq-aging-dementia-and-tbi

Annexure

Code for this project comprises of 14 R files and a folder which are available in Github.

File named requirements.R comprises of the list of all the essential packages required.

Data_visualization.R performs the exploratory data analysis.

Diff_Exp_(name of the disease).R comprises of code for differential expression analysis and returns list of DEGs in csv file.

Func_Enrich_(name of the disease).R performs functional enrichment analysis and provides upregulated and downregulated genes.

Gene_Ontology_(name of the disease).R performs gene ontology analysis.

Pathway_Enrichment_(name of the disease).R performs the pathway enrichment analysis.

DEGs folder comprises of csv files obtained from Diff_Exp_(name of the disease).R