


The background of the slide is a blurred image of a financial market data screen. It features various stock indices and their values, such as 'OMX COPENHAGEN 25 INDEX' with a value of 1172.94, 'OMXRG1' with 10916.69, and 'OMX18' with 6230.9. There are also line graphs showing price fluctuations and 'Buy' or 'Sell' indicators. The overall color scheme is dominated by blue and red, typical of financial data visualizations.

MATH7343 Applied Statistics
Spring 2023

RNA Sequencing Statistical Data Analysis

Abhilasha Jain, Ailin Dolson-Fazio, Mukund Kudluga, Nikhil Anand, Prathamesh Thite

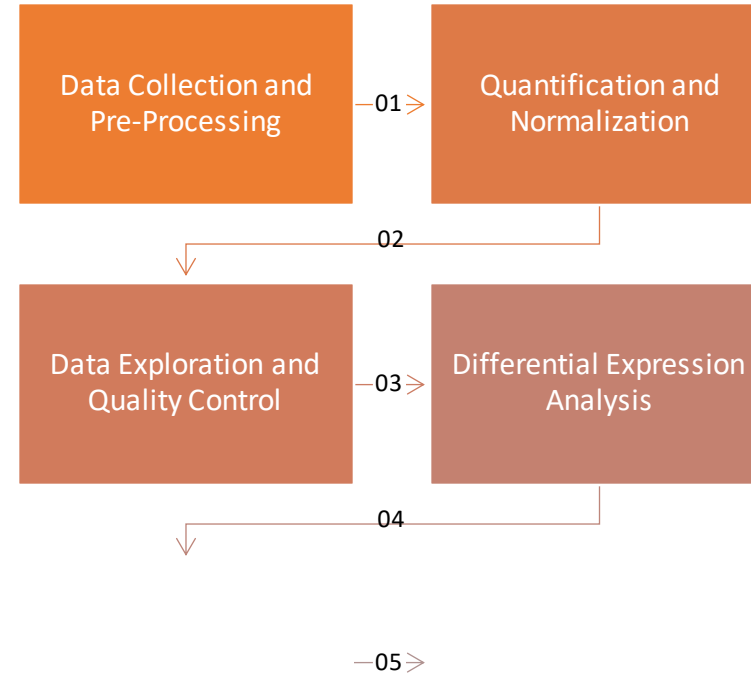
- 
- Topic in Functional Genomics
 - Uses next-generation sequencing to analyze the transcriptome of a cell
 - Used for analyzing gene expression and separation of classification of samples

Background

Objective

- To perform a statistical analysis of RNA sequencing data
 - To identify potential biomarkers, pathways, and gene signatures
- Data on Aging and Dementia and Traumatic Brain Injuries (TBI).

Methods and Methodology

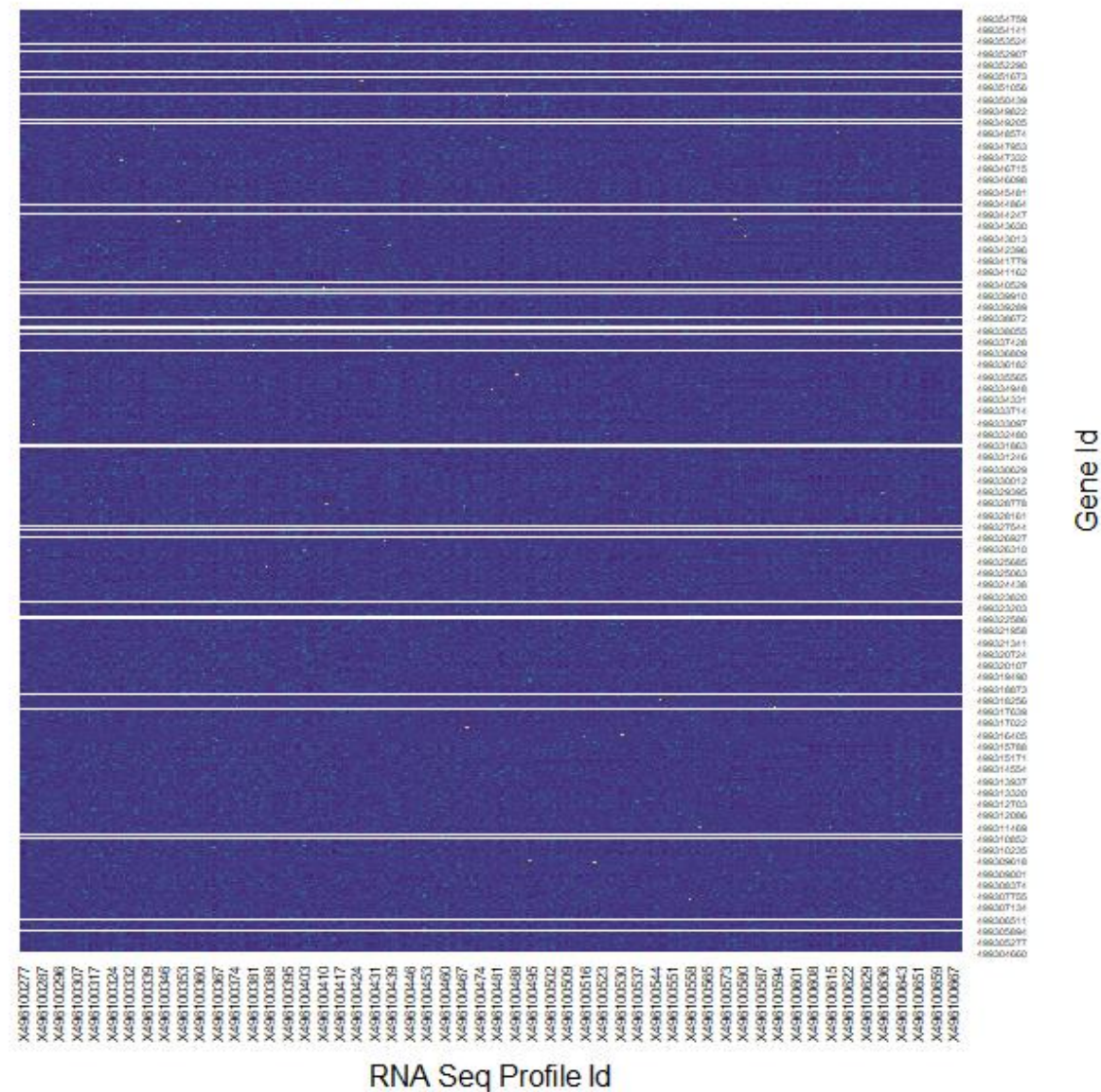


Dataset

- The Aging, Dementia and Traumatic Brain Injury Study "Adult Changes in Thought" (ACT)
- From the University of Washington, Kaiser Permanente Health Research Institute, and the Allen Institute for Brain Science
- A longitudinal study of aging, dementia, and Traumatic Brain Injuries (TBI) within the Seattle region
- Detailed neuropathologic, molecular and transcriptomic characterization of brains plus control group

Raw Gene Expression

Gene Expression Unnormalized



Data Collection and Pre-Processing



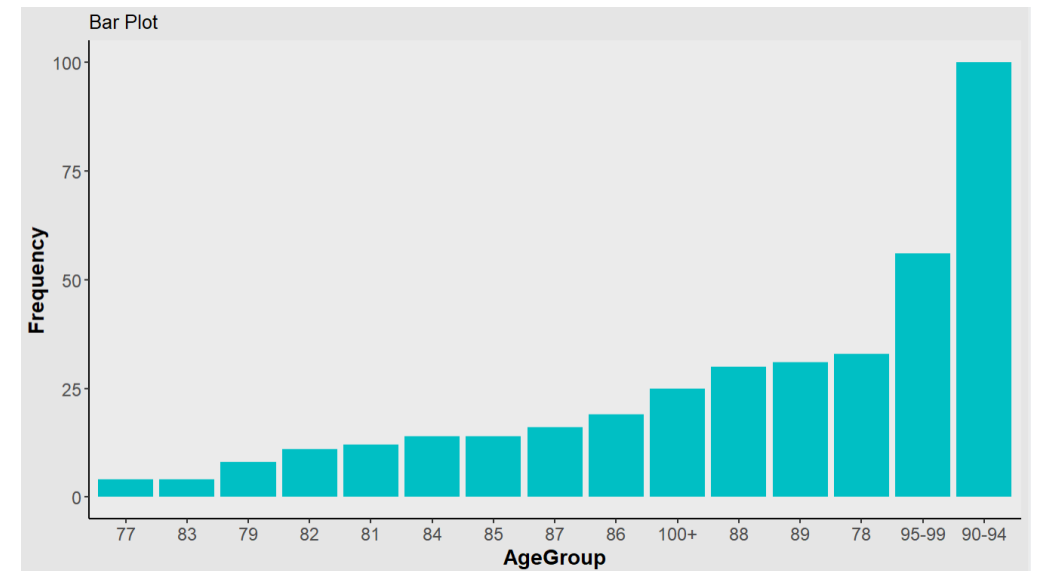
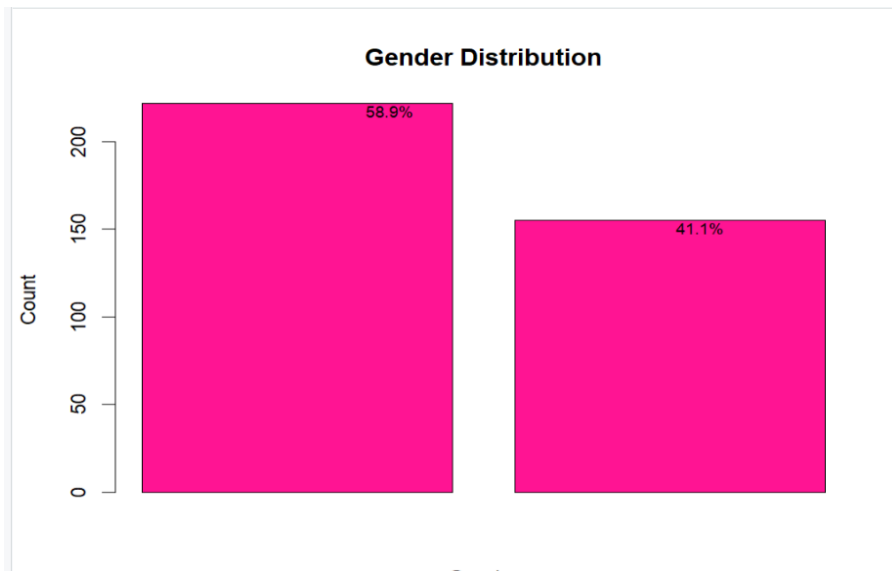
**Datasets can be sourced publicly or
through private research**



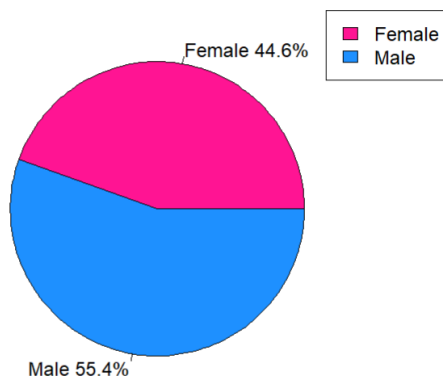
Pre-Processing

Gender Distribution

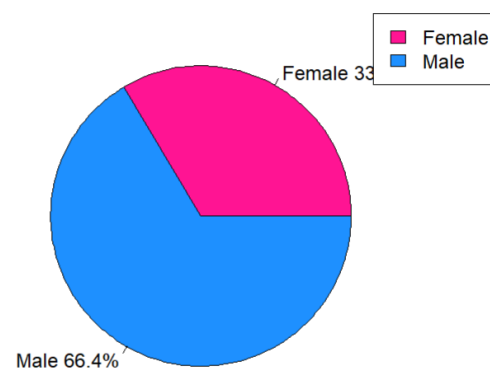
Age Distribution



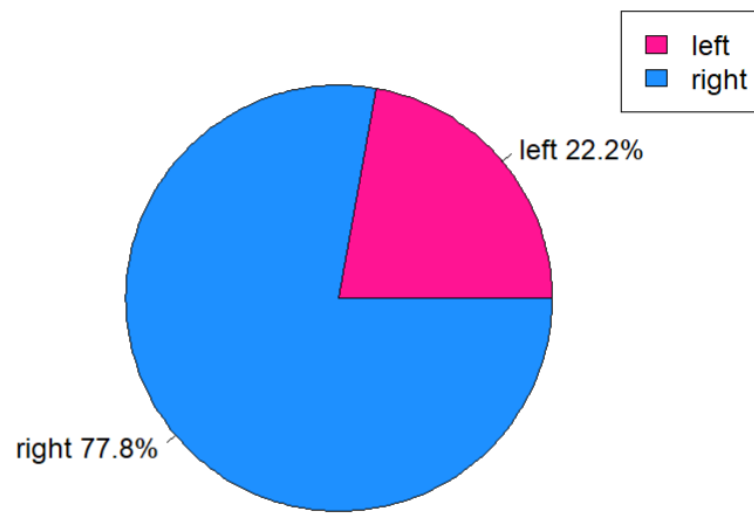
Right Hemisphere Gender Distribution



Left Hemisphere Gender Distribution

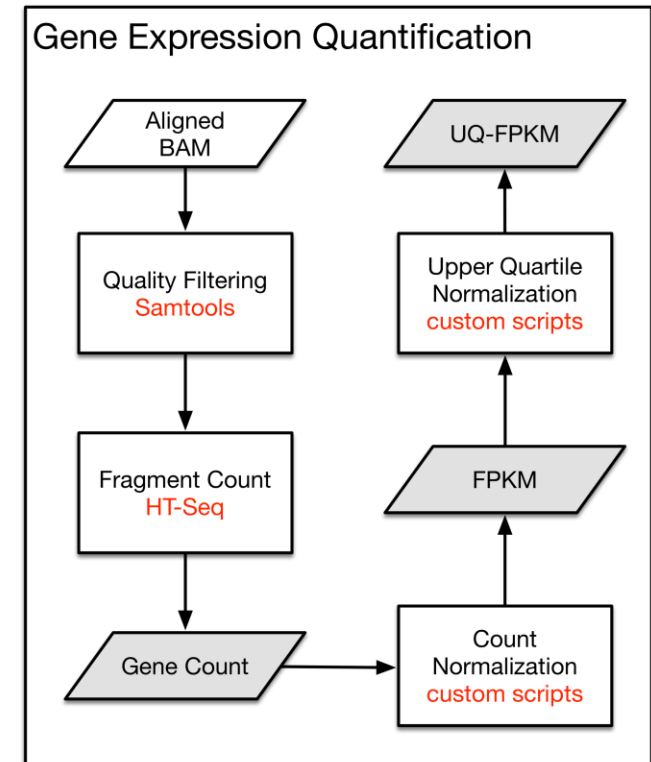
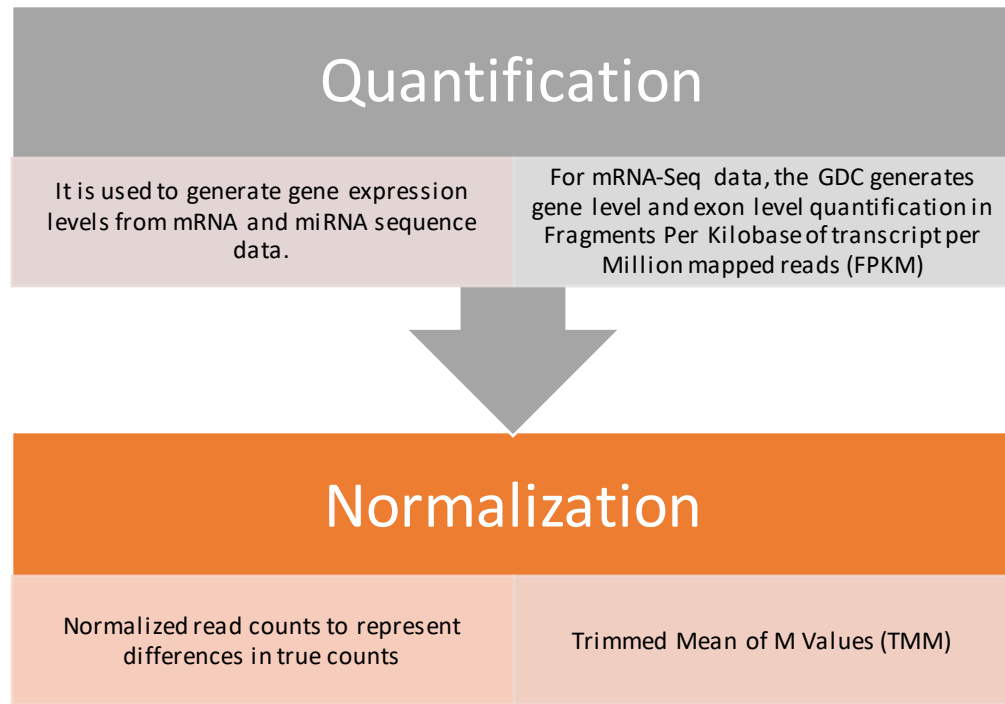


Hemisphere Distribution

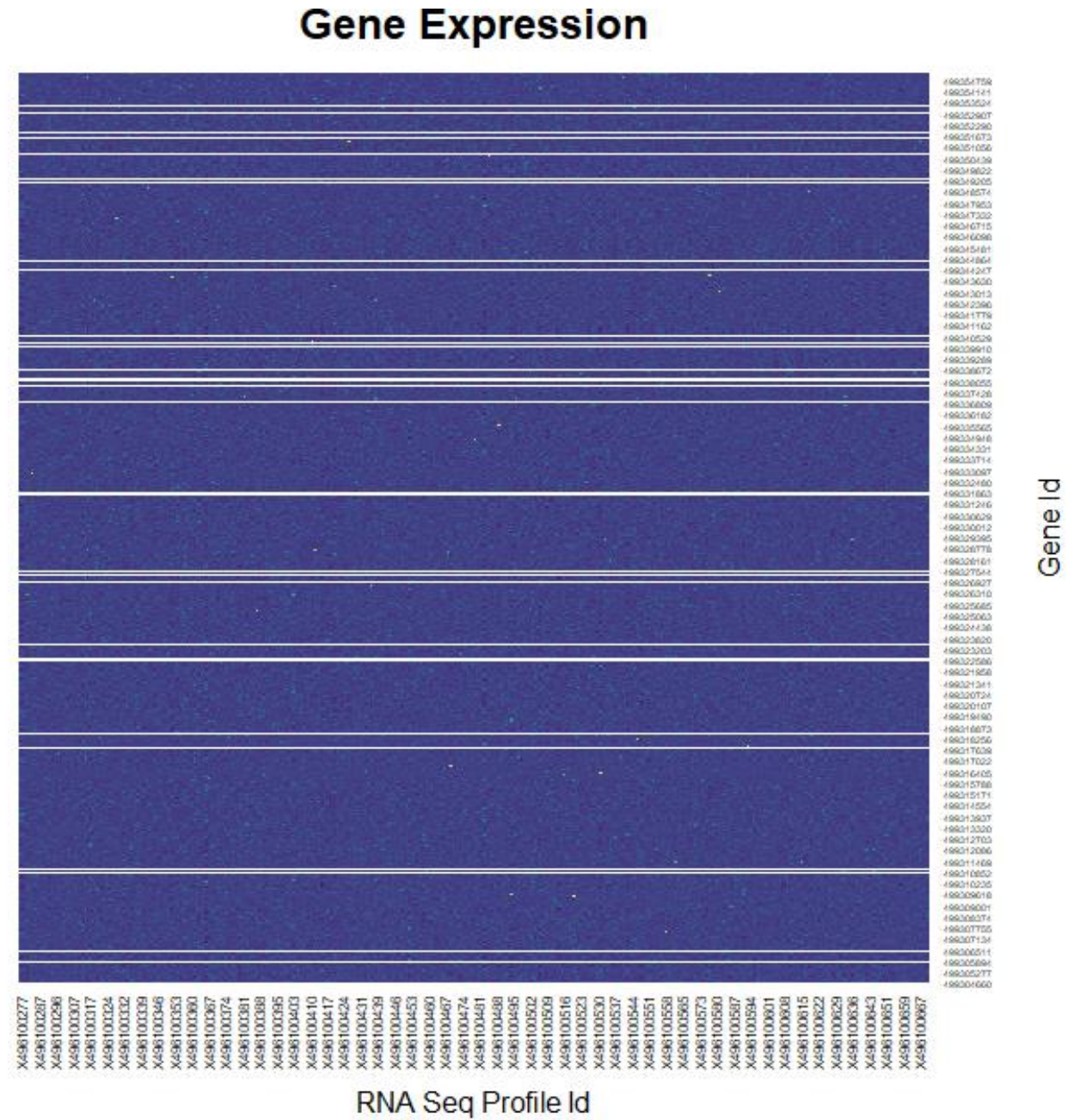


Brain Hemisphere
Distribution for Dataset

Quantification and Normalization



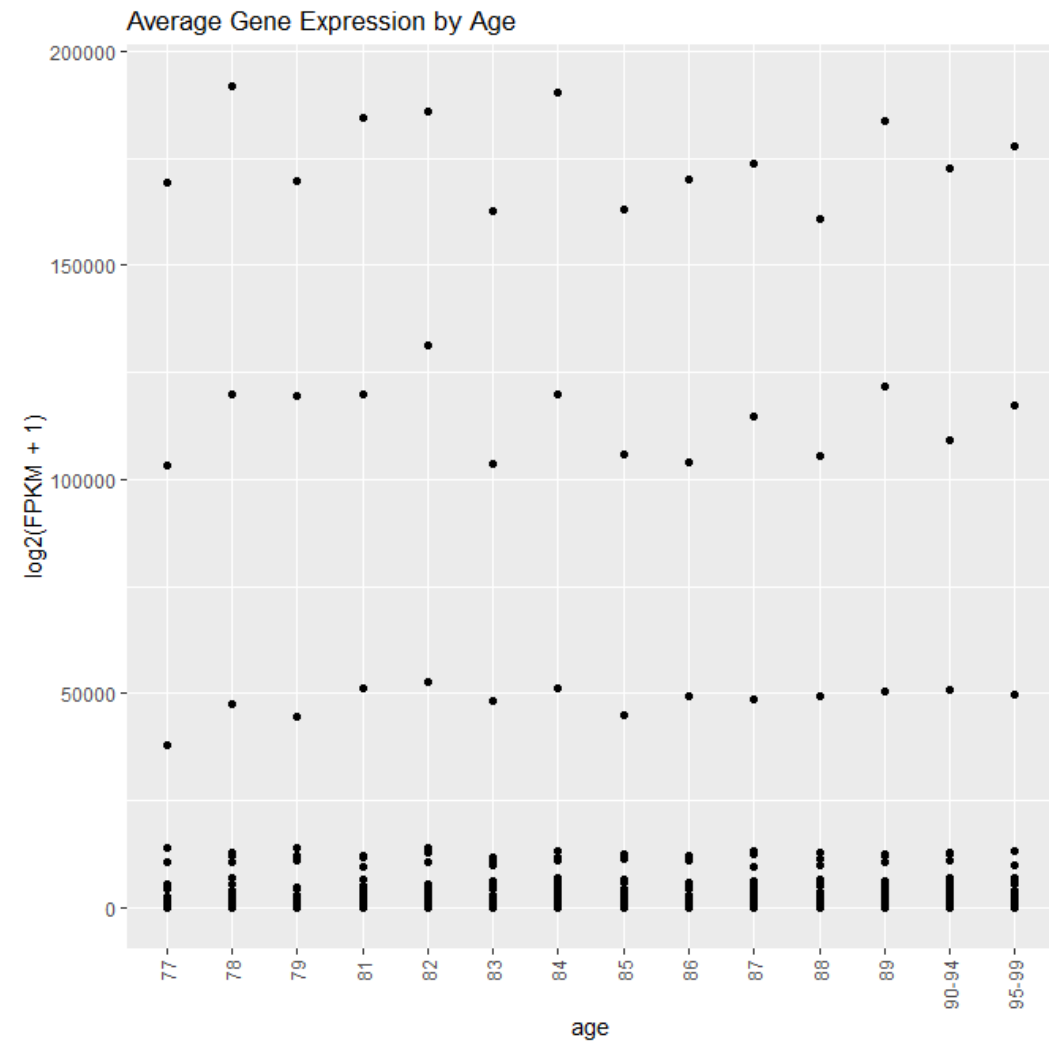
Gene Expression After Normalization



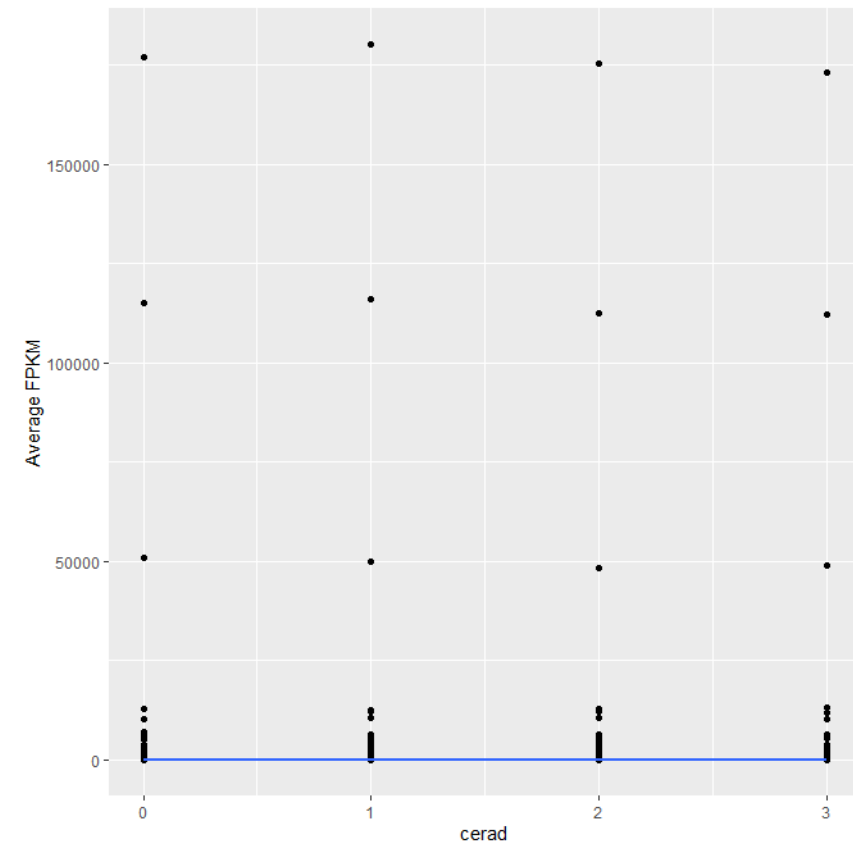
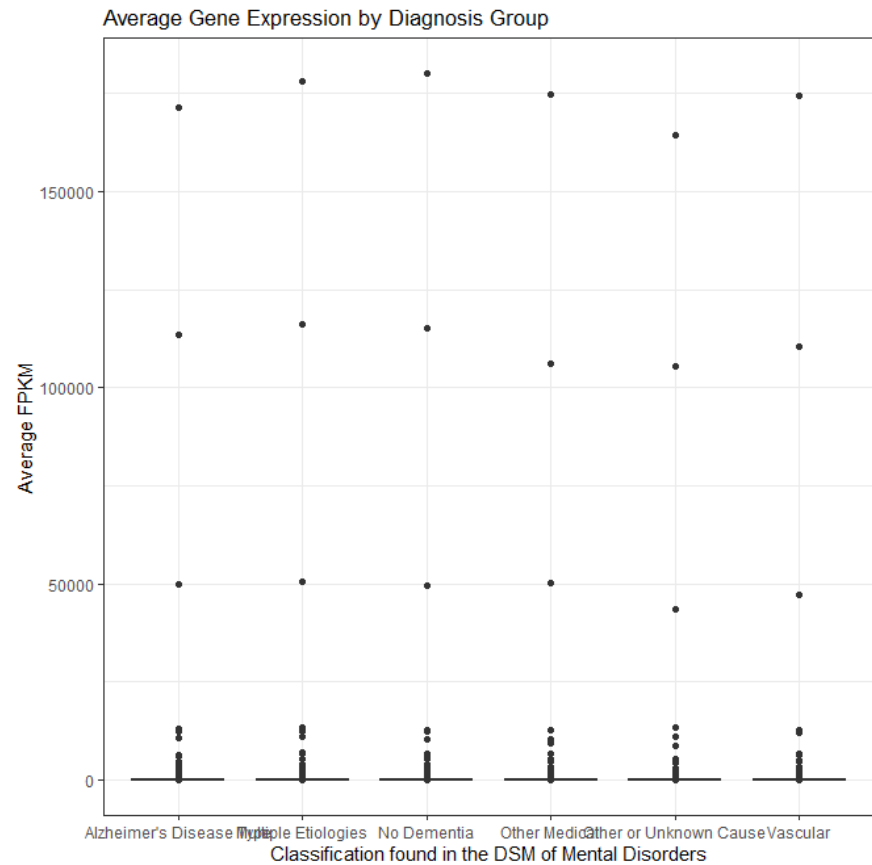
Data Exploration and Quality Control

- Ensures accuracy and reliability of pre-processed and normalized data
- Explore relationships between samples
- Identifies outliers
- Addresses overall quality of the data
- Available Tools
 - FastQC
 - Principal Component Analysis (PCA)
 - Multi—Dimensional Scaling (MDS)
 - Heatmaps

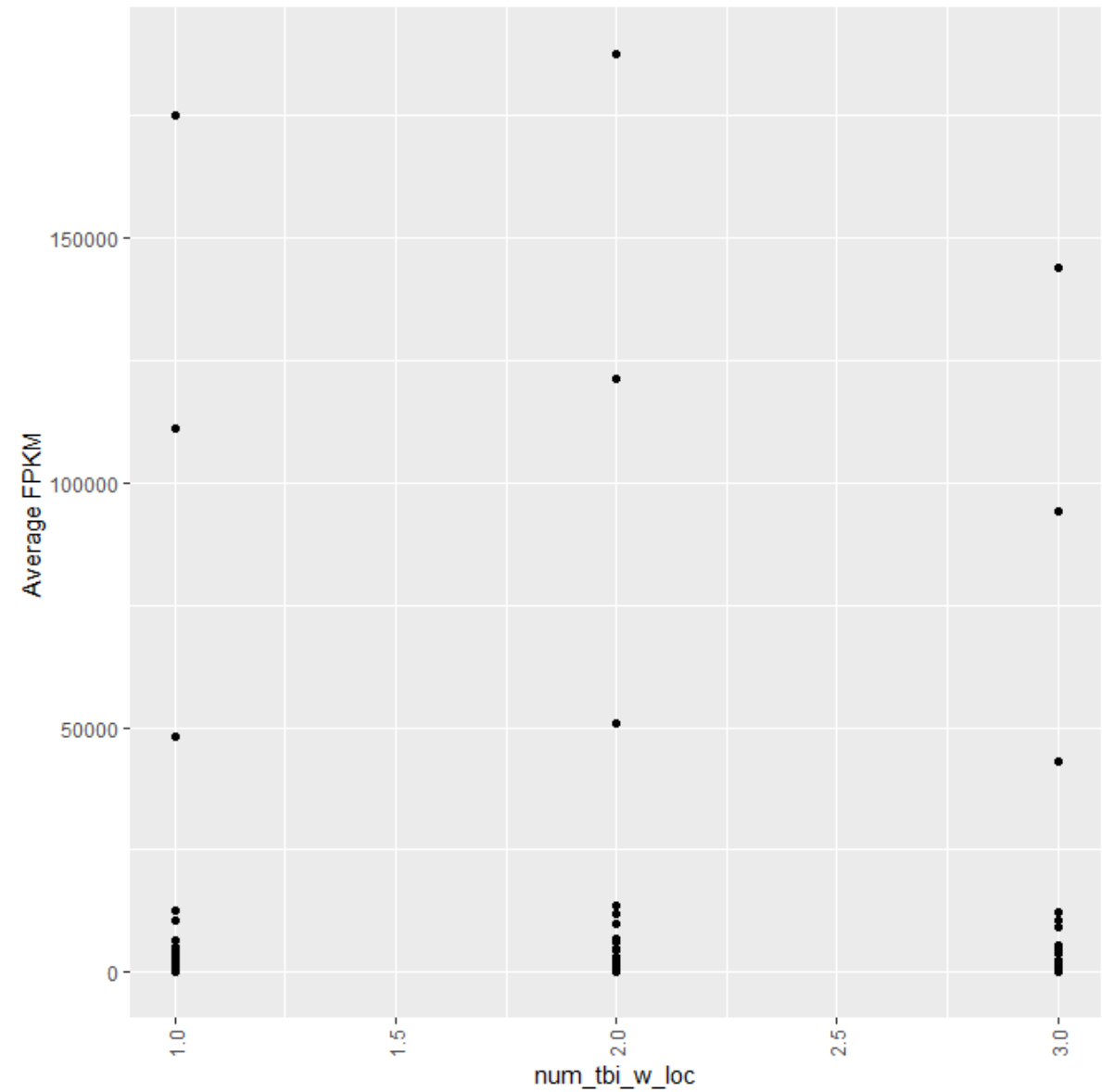
Gene Expression w.r.t. Age



Gene Expression by Dementia Group and CERAD Score



Gene Expression w.r.t. Number of TBIs



Differential Expression Analysis



Identifies differently expressed genes between samples of interest



Specifically looks for genes with significant changes in expression levels between different conditions or groups



Typically involves estimating fold changes

Represents magnitude of expression changes, p-values, and statistical significance

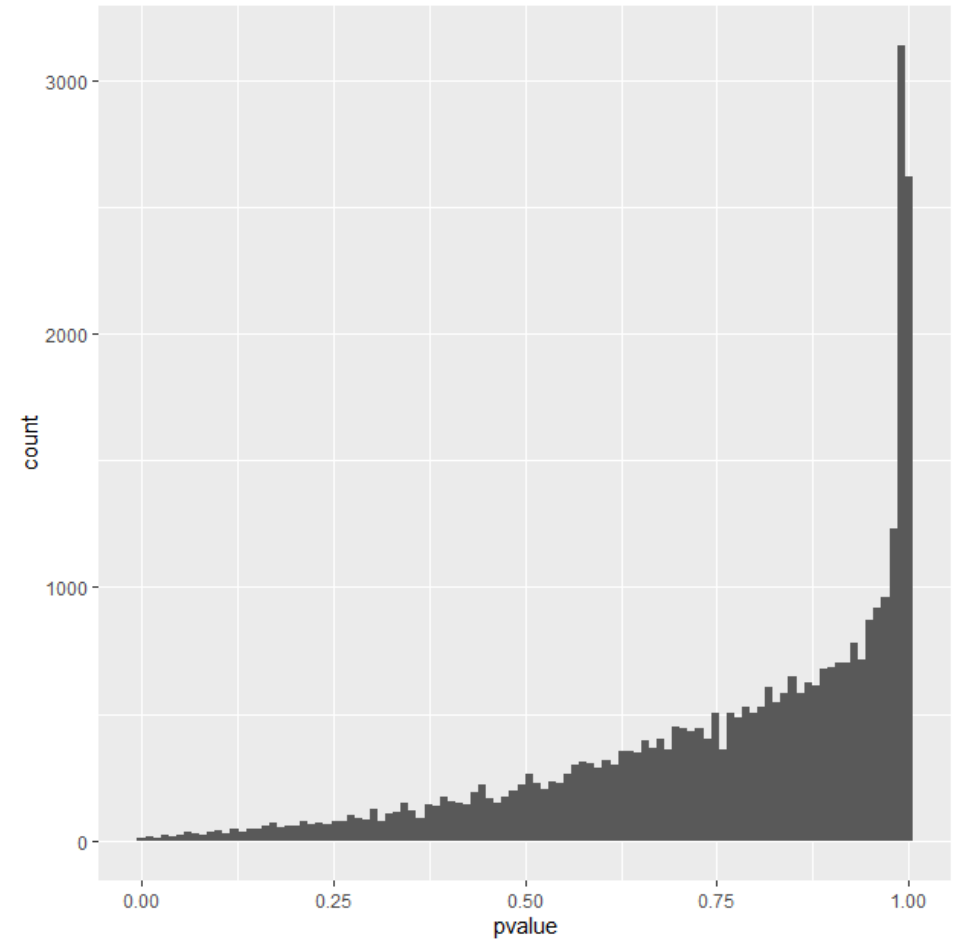
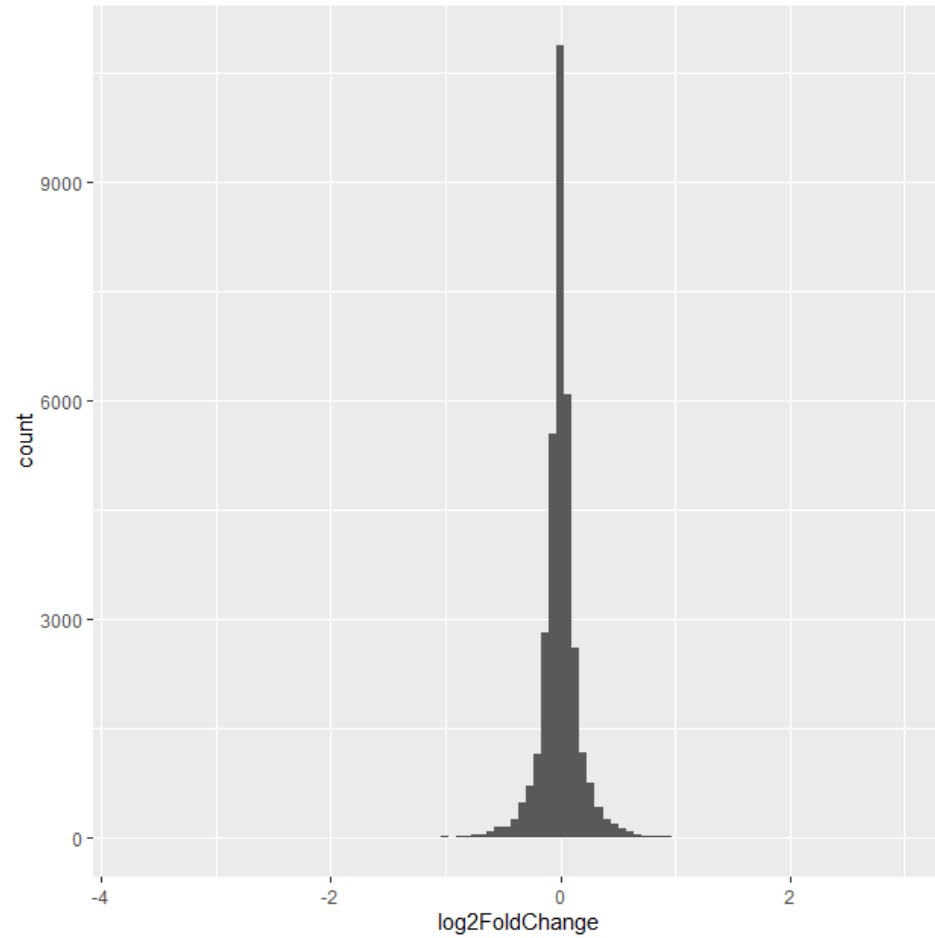


Available Tools

Methodology

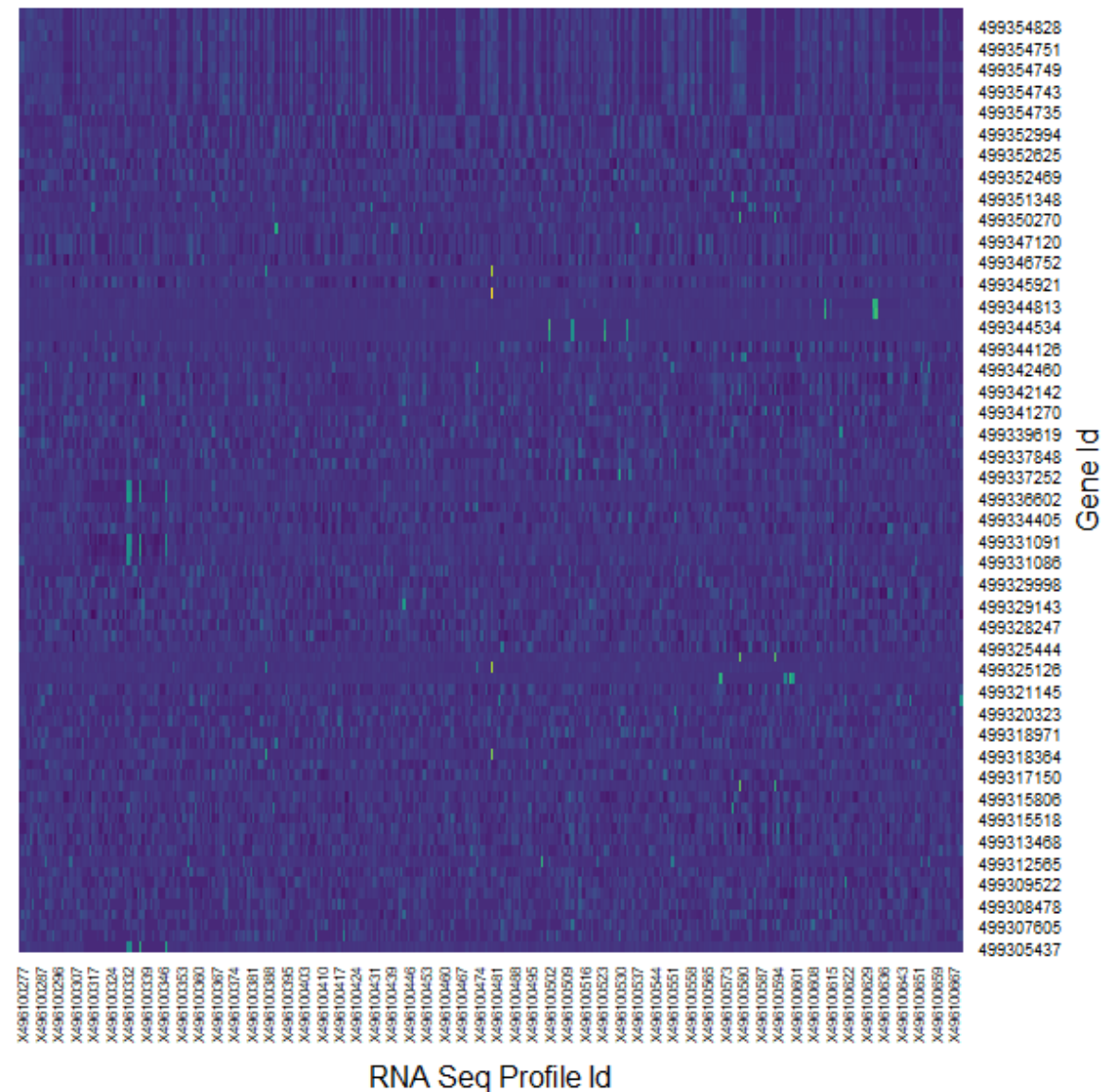
- Read counts for each gene follows a Negative Binomial Distribution
- Evaluate μ and σ for each gene for different conditions.
- Perform Log-likelihood ratio test (LRT) to determine test statistic
- Chi-squared distribution with 1 DF

P-value and LFC for Aging

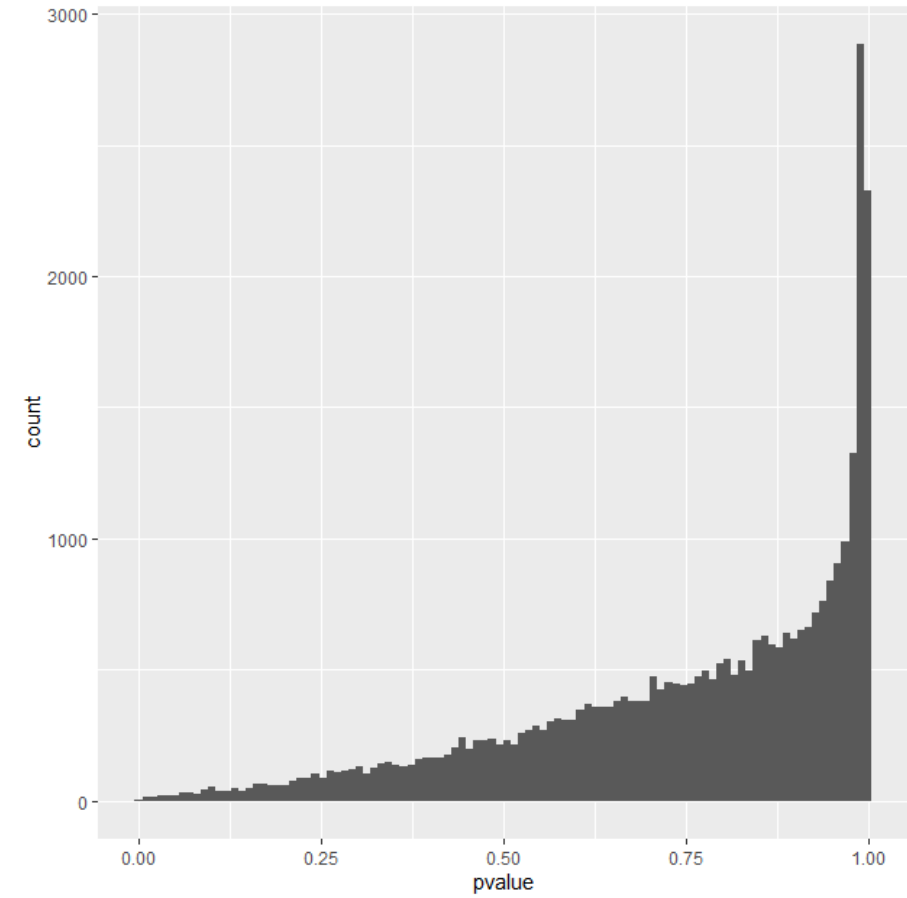
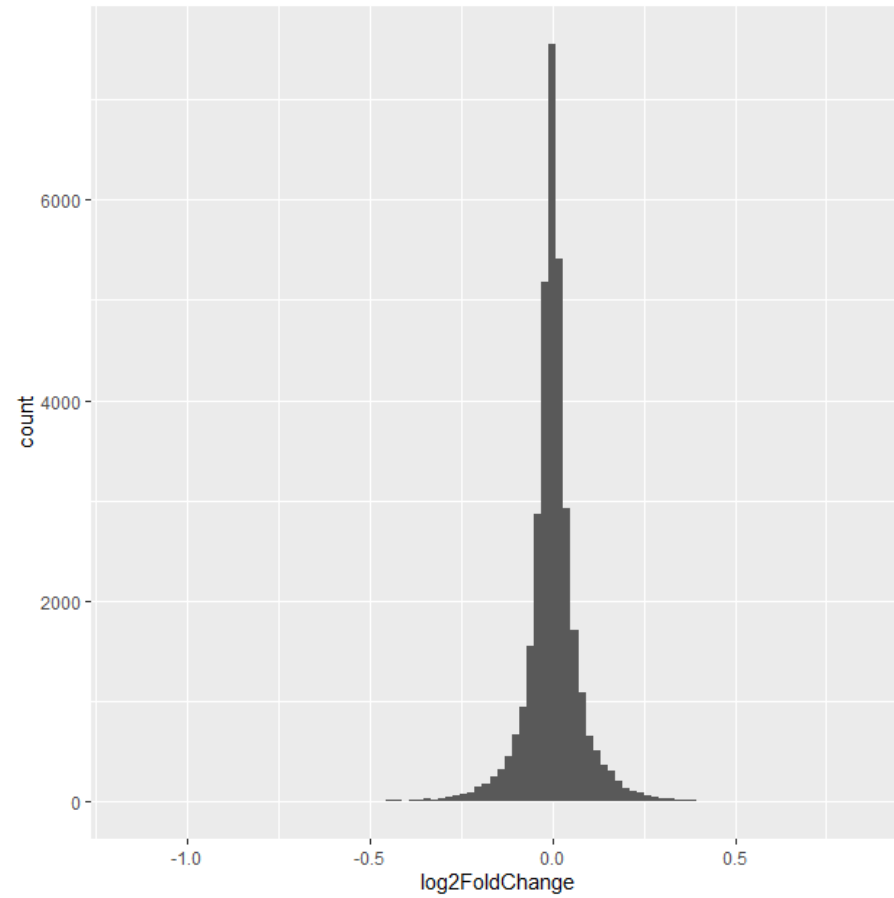


Gene Expression of DEGs for Aging

Gene Expression of Significant Genes

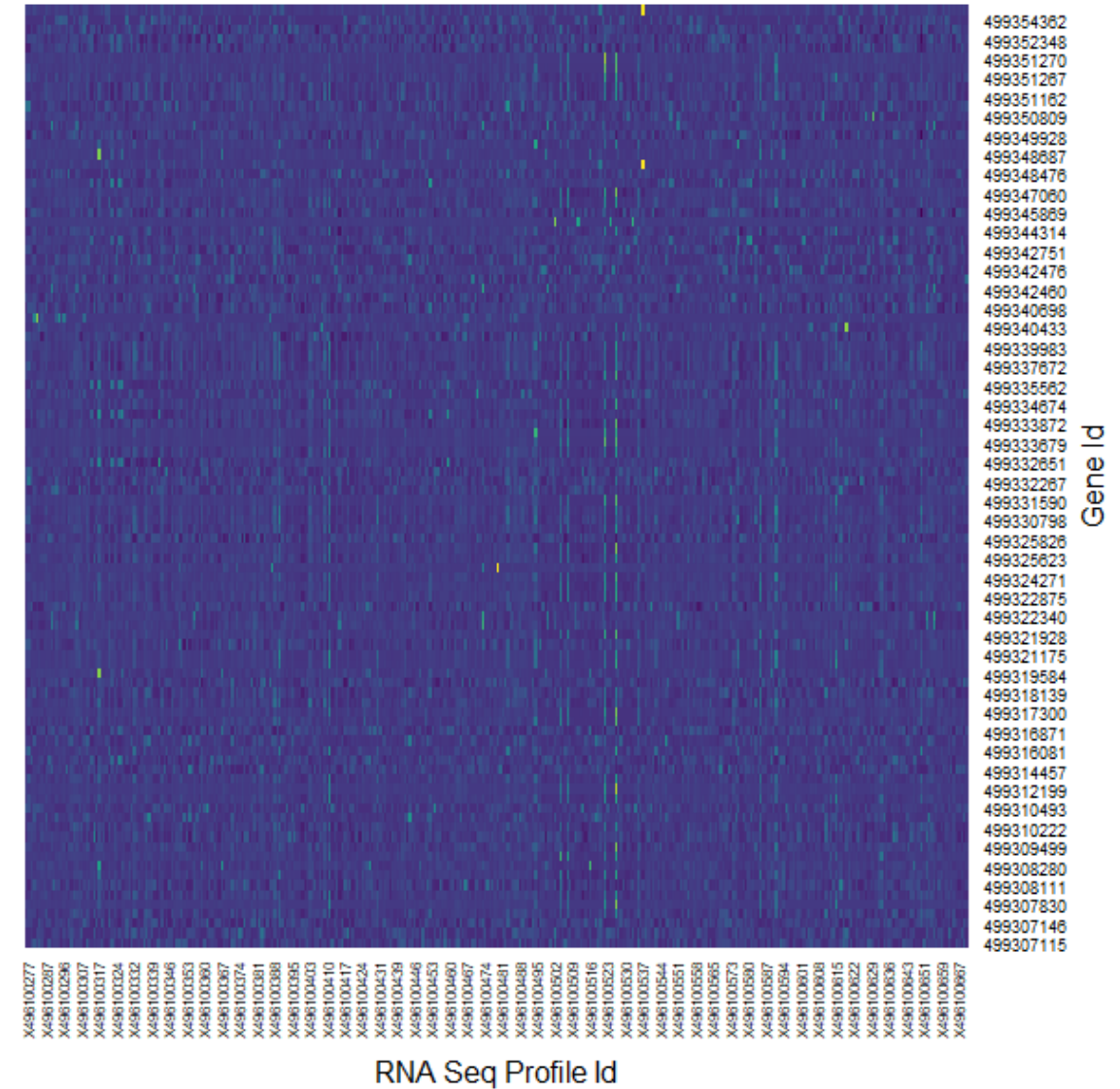


P-value and LFC for TBI

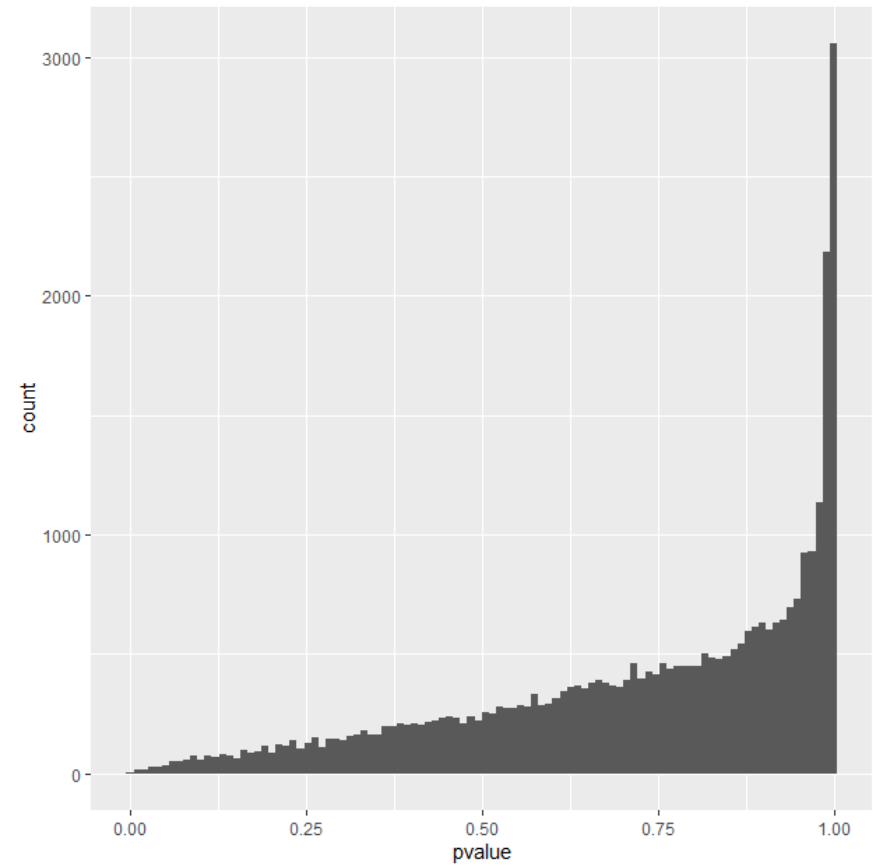
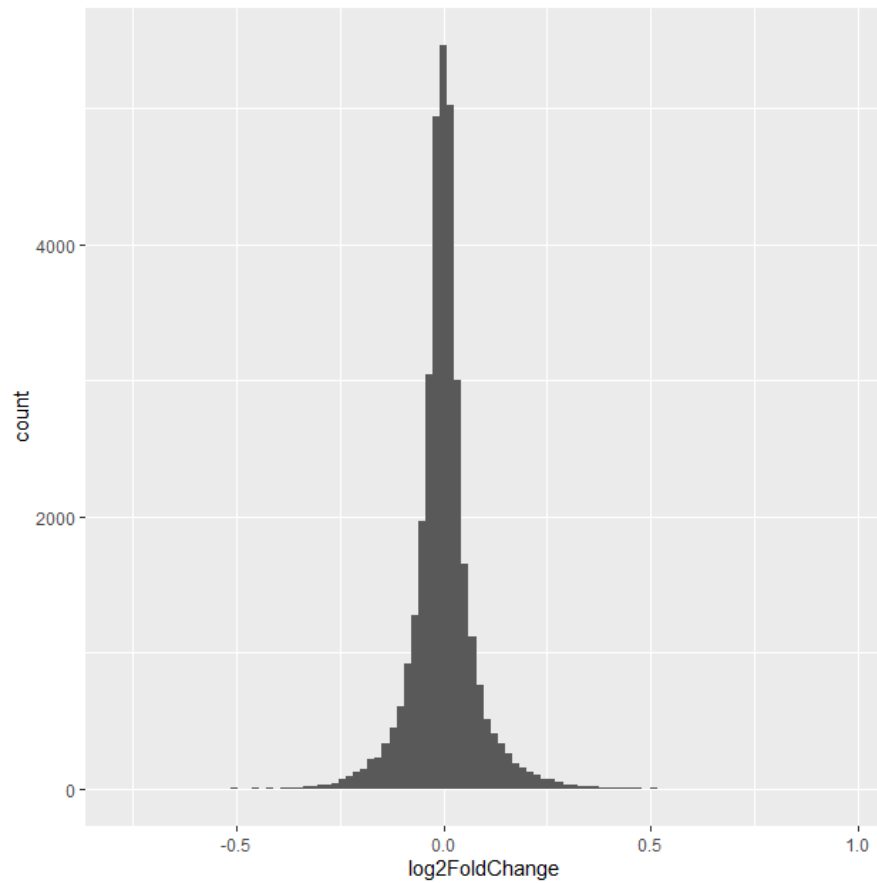


Gene Expression of DEGs for TBI

Gene Expression of Significant Genes

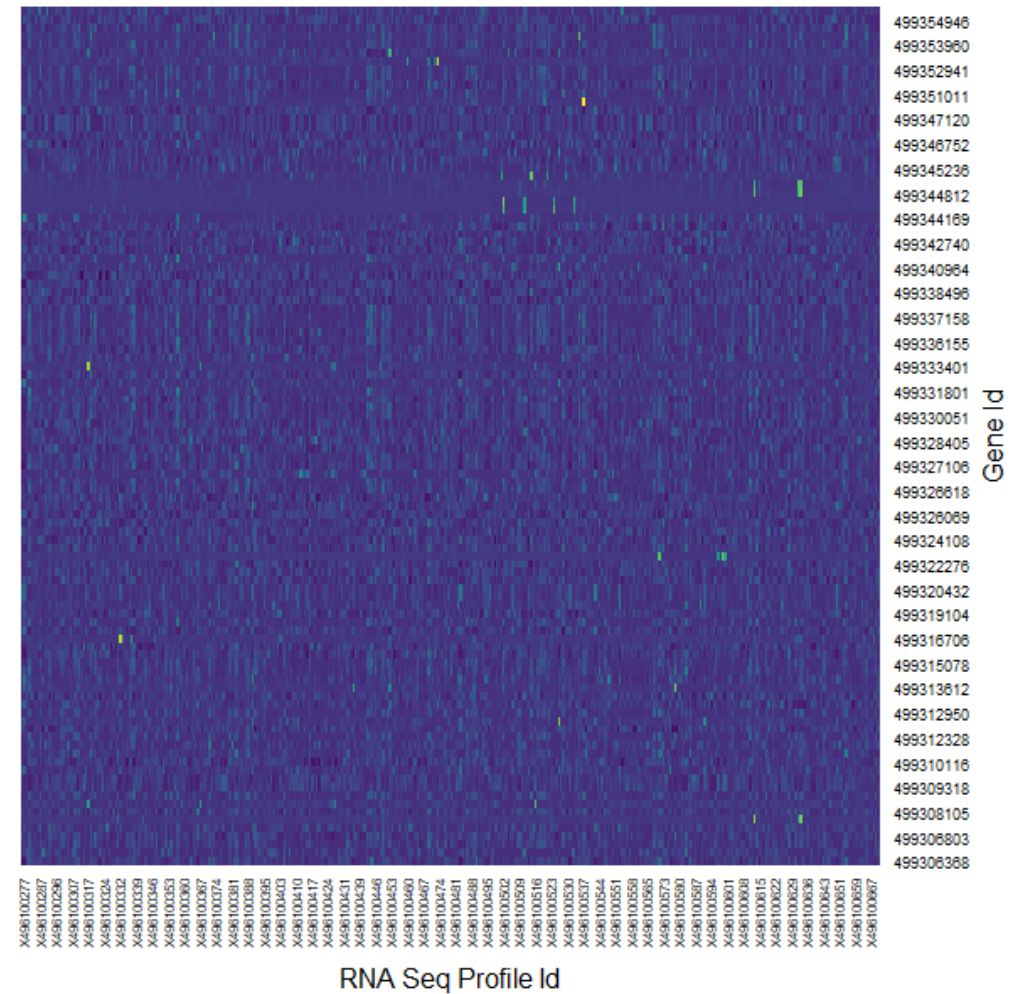


P-value and LFC for Dementia



Gene Expression of DEGs for Dementia

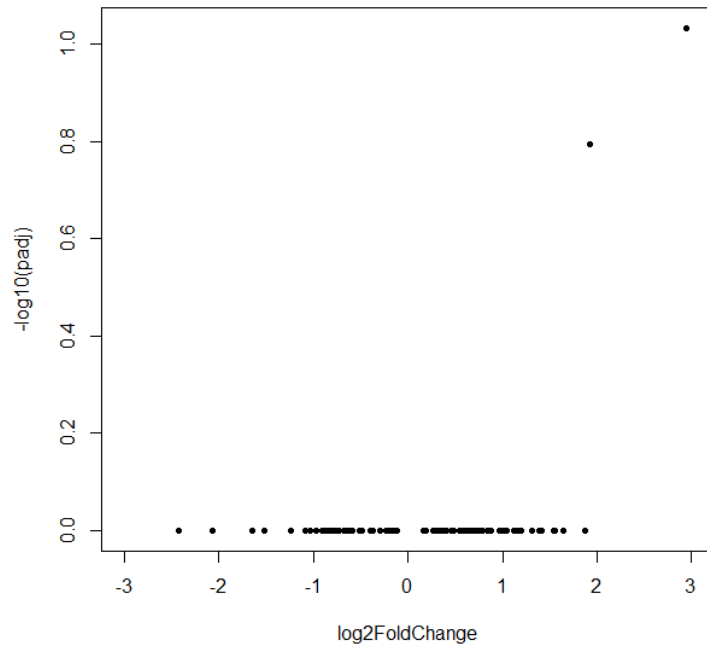
Gene Expression of Significant Genes



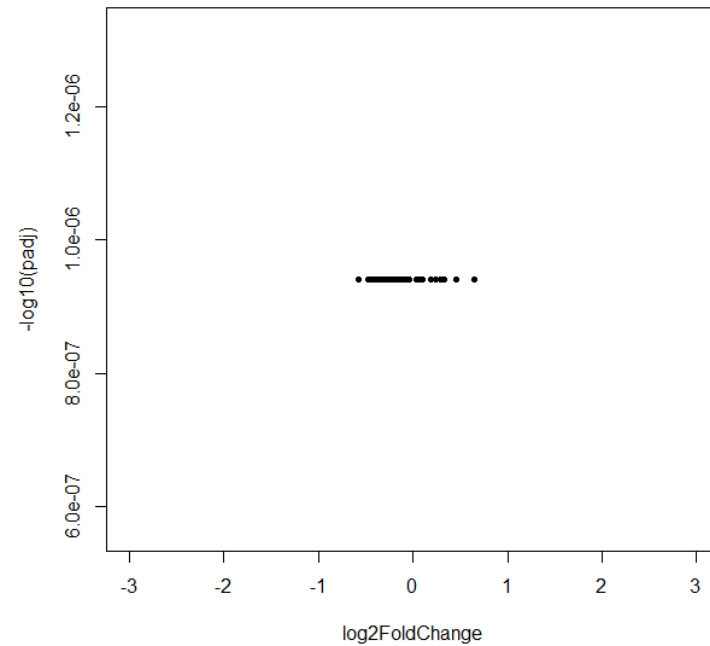
LFC of Differentially Expressed Genes



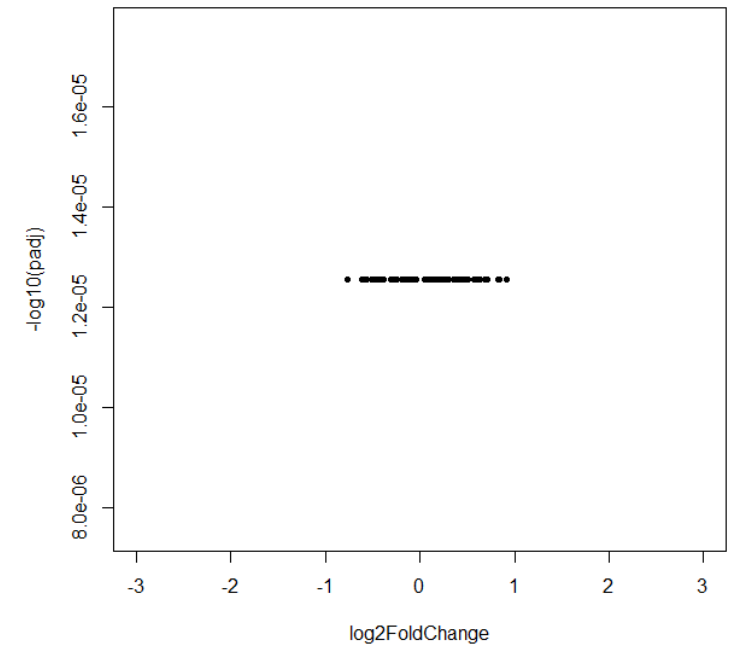
Volcano plot of Significant DEGs of Aging



Volcano plot of Significant TBI DEGs



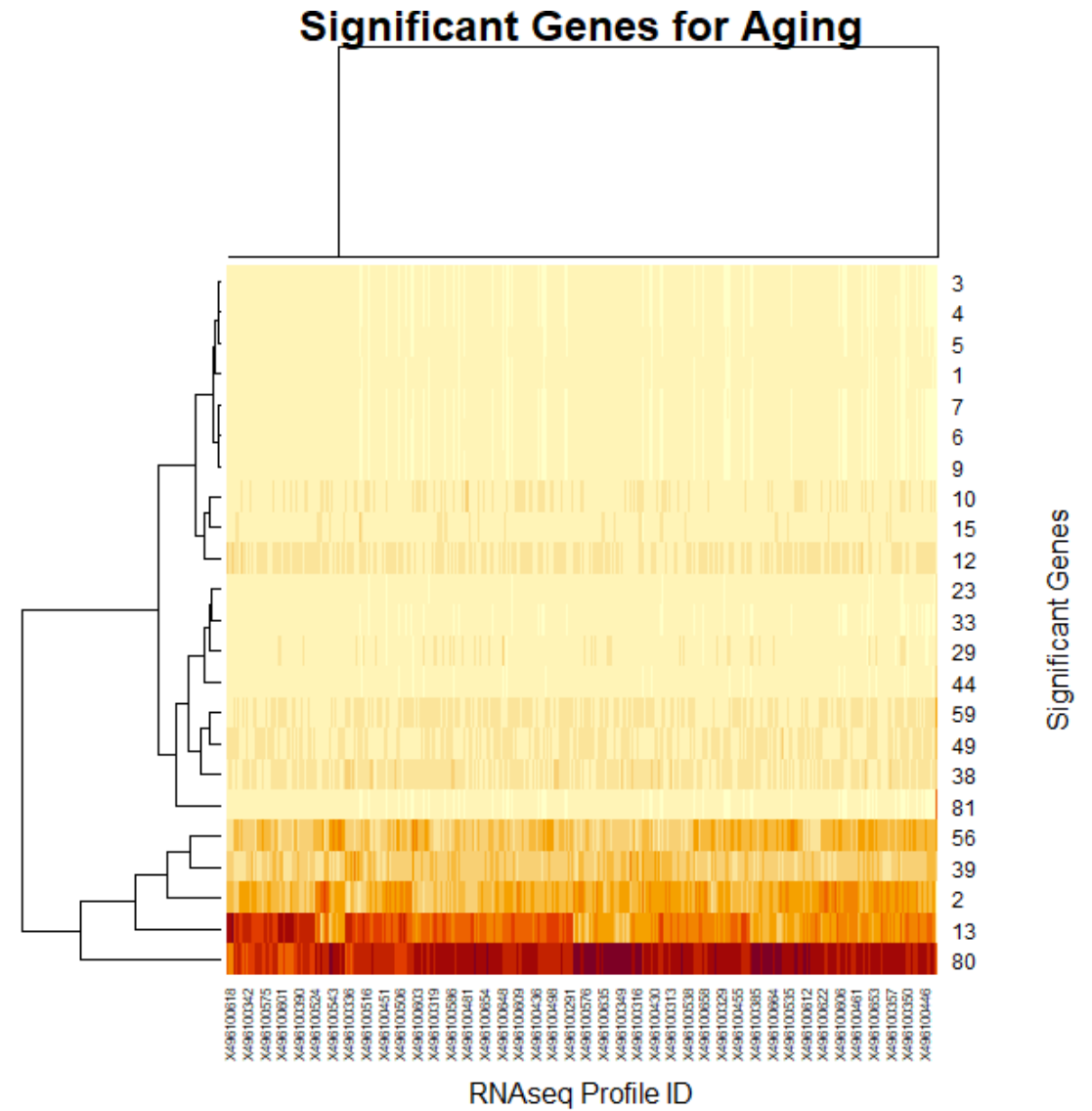
Volcano plot of Significant Dementia DEGs



Functional Enrichment Analysis

- To study bunch of genes together and figure out what they might be doing in the body
- Used to gain insights into the affected biological functions and pathways
- Mainly sets up for next steps
- Looks for over-represented terms
- Available Tools
 - clusterProfiler
 - Gene Set Enrichment Analysis (GSEA)

Expression of Upregulated Gene & Downregulated Gene



Gene Ontology and Pathway Analysis

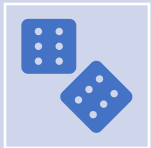
- Annotates differently expressed genes with functional information
- Used to gain a deeper understanding of the biological roles
- Helpful for interpreting correlation between results and pre-existing biological knowledge
 - Gives direction for potential future research
- Available Tools
 - GO Database
 - KEGG

Visualization and Interpretation



Helpful for better understanding and communication of results

Can be represented as volcano plots, heatmaps, and gene expression plots



Summation of research

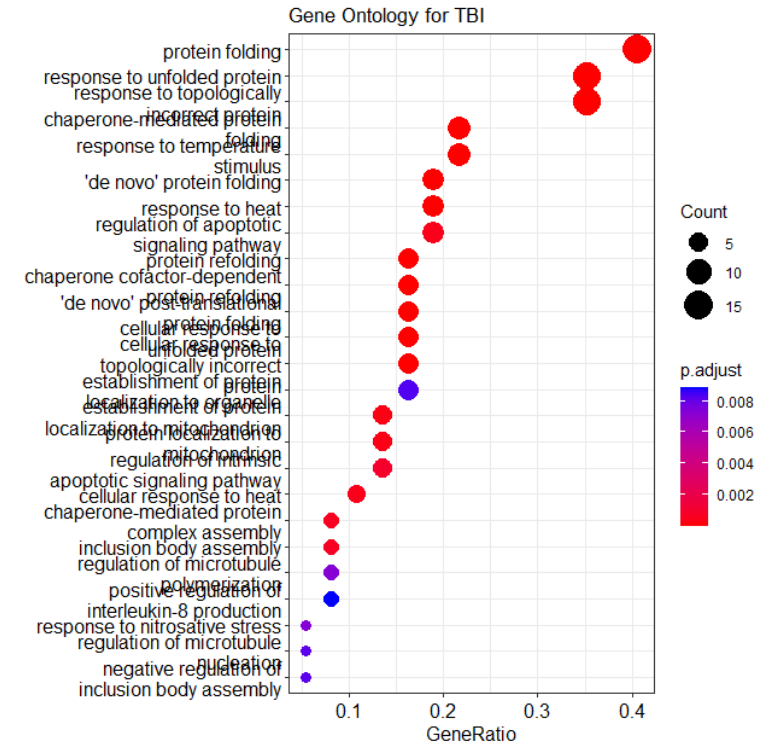
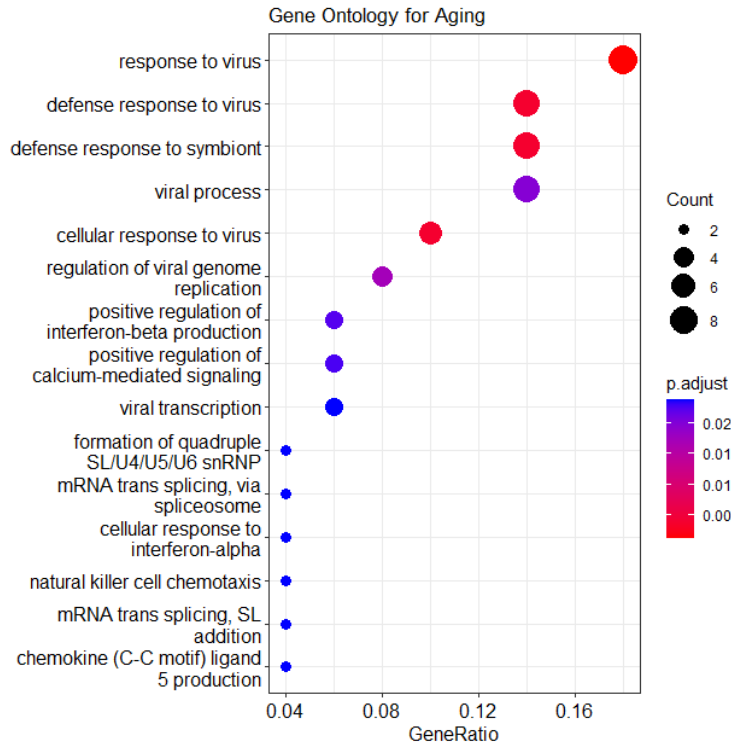
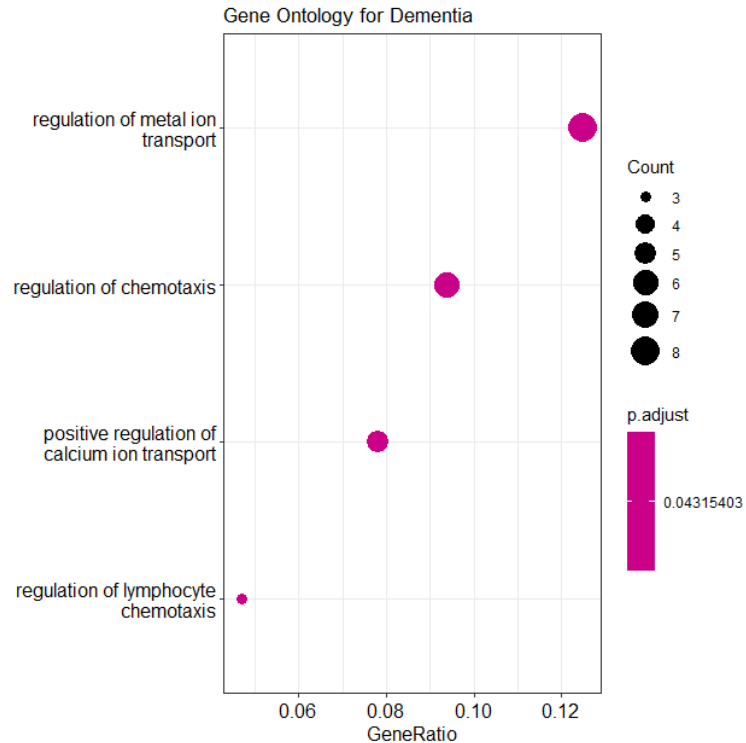
With the help of figures, tables and text



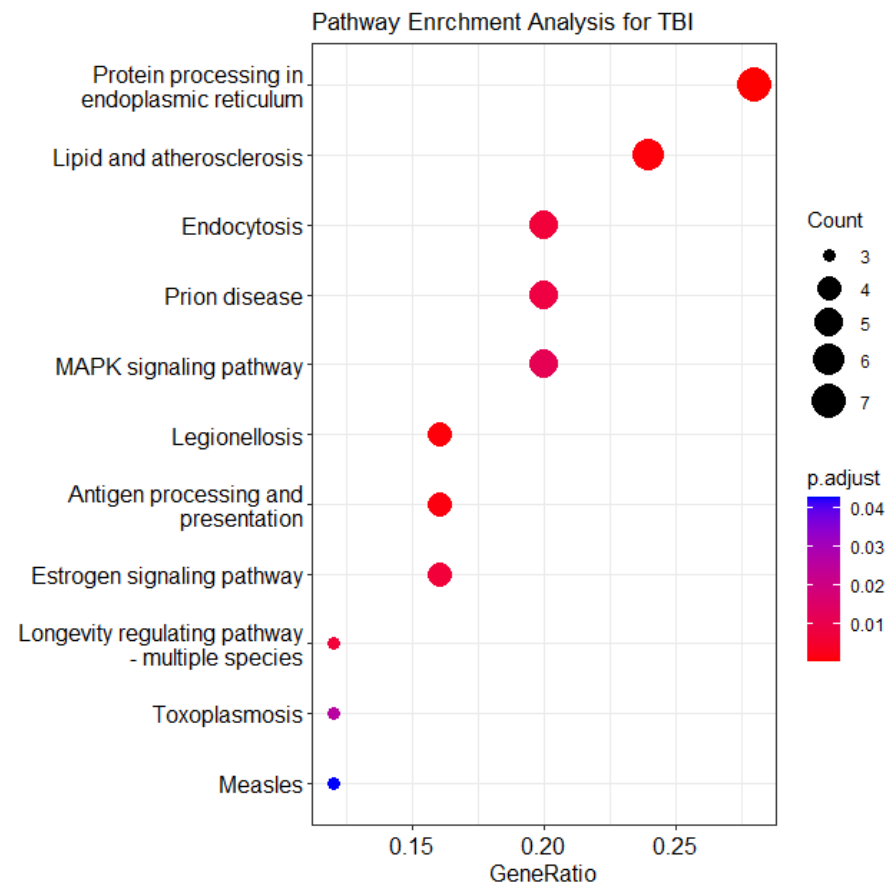
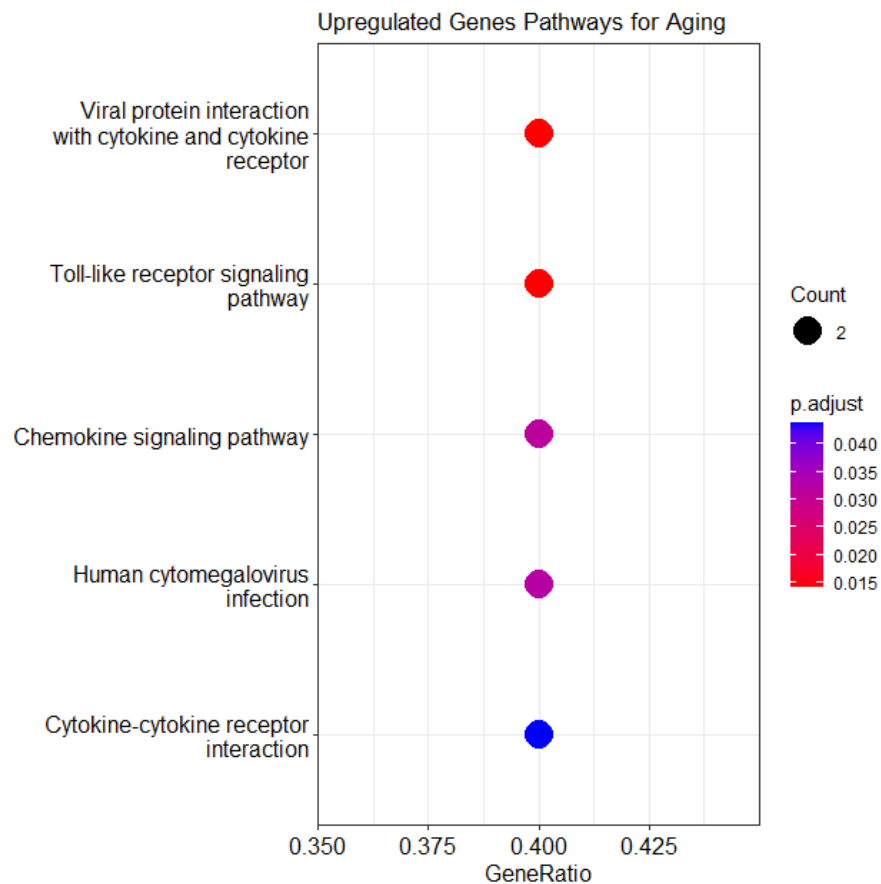
Available Tools

Ggplot, qcplot

Gene Ontology Results



Pathway Enrichment Analysis Results





Aging

- Total DEGs – 88
- Upregulated genes – 15
- Downregulated genes – 8
- Pathway Enrichment – viral infection and immune response
- Gene Ontology - Enrichment related to Biological processes – Response to virus and defense response to virus and symbiont

Dementia

- Total DEGs – 104
- No Upregulated and downregulated genes were found
- Pathway Enrichment – no significant overrepresentation
- Gene Ontology - Enrichment related to Biological processes – positive regulation of calcium ion transport, regulation of metal ion transport, regulation of lymphocyte chemotaxis, and regulation of chemotaxis
- Gene responsible for cellular transport and chemotaxis processes

Traumatic Brain Injury



- Total DEGs – 66
- No Upregulated and downregulated genes were found
- Pathways Enrichment – 11 KEGG pathways related to protein processing in endoplasmic reticulum, antigen processing and presentation, and lipid metabolism.
- Gene Ontology - Enrichment related to Biological processes – protein folding, response to unfolded protein, response to topologically incorrect protein, chaperone-mediated protein folding

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 Huse, *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>



THANK YOU



Any Questions??

