Midterm Project Report: Utilizing Gene Expression Data for Alzheimer’s Disease Analysis:

Introduction:

In this report, I outline the steps undertaken in our midterm project, focusing on the analysis of gene expression data to investigate potential molecular biomarkers and drug targets for Alzheimer's disease (AD). Our project aimed to contribute to the ongoing efforts in Alzheimer's disease research by examining gene expression patterns in human brain tissue samples from moderate and control subjects.

I did the codes in R:

Step 1 - Read in Data:

I initiated the analysis by importing two crucial datasets:

-Clinical data, obtained from "Blalock\_clin\_final.csv," which includes information on subjects' disease status.

-Gene expression data from "GSE62232\_Blalock\_geneexp\_final.tsv," which encompasses the gene expression profiles.

Step 2 - Clean/Filter Data:

With the clinical data in hand, I applied data filtering to select two primary groups:

- The "Moderate" group, representing AD subjects.

- The "Control" group, representing normal control subjects.

Step 3 - Identify Groups

I identified these groups by determining row labels corresponding to baseline and comparison group patients based on their disease status.

Step 4 - Sanity Check

To ensure data integrity, i conducted a sanity check to confirm that sample IDs in the clinical data matched those in the gene expression data. Additionally, i exported the features from the gene expression data for future reference.

Step 5 - Prep Data for T-Test

I checked the data's numeric nature and then proceeded to prepare it for statistical analysis.

\*\*Step 6 - Call Function for T-Test\*\*

To compare gene expression between the baseline and comparison groups, I implemented a t-test by calling a custom R function. We saved the results for further analysis.

Next Step - Sub-set Top Differentially Expressed Genes\*\*

In this step, i refined our results by shortlisting genes with a significant impact based on a p-value threshold (p-value <= 0.05). This allowed us to create a list of differentially expressed genes (DEGs) for further exploration.

- Imported the t-test results and further filtered them.

- Cleaned gene names by splitting them and removing duplicates and NA values.

- Loaded databases for the Enrichr R package, ensuring the connection was operational.

- Set the output file name and initiated the enrichment analysis using selected databases.

This report has outlined the essential steps of our midterm project, encompassing data import, cleaning, statistical analysis, and enrichment analysis. We hope that this analysis contributes to the ongoing research on Alzheimer's disease and the identification of potential biomarkers and drug targets. Further investigations into this complex disease may yield valuable insights for future studies and therapeutic interventions.

Then In this collaborative step, all three teams compared the DEG lists generated in Step 1, utilizing Venn diagrams and Excel to reveal shared and distinct genes among early, moderate, and severe stages of Alzheimer's disease. These visualizations highlighted the genes common to multiple stages, shedding light on crucial molecular players in AD progression. Further exploration of the shared genes using online resources deepened our understanding of their biological functions. This collective effort is a significant stride in unraveling the intricate mechanisms behind Alzheimer's disease and advancing our quest for potential biomarkers and drug targets.