# Project 2 Executive Report

2025-10-09

## Introduction

We identified several promising genes for future study into aging in the human brain. According to our analysis, we would suggest further examiniation of 20 genes from across three sites in the human brain. From the ACG region... These genes have demonstrated statistically signficant expression between patients less than 65 and those 65 and older at the $alpha = ...$ level.

These results were obtained through versions of a linear model adapted to each brain region studied in our sample, as we observed stark differences between the expression of genes based on their location in the brain. The justification for and specification and interpretation of these models will follow in the following sections.

### Exploratory Data Analysis

The distribution of the available data makes pairwise analysis difficult. If we wished to compare gene expressions for older and younger people for the same brain region and lab site using this method, we would run into complications quickly. Referring to the table below, one can see glaring sparsity at the region-lab level. For example, genes from the ANCG region are completely absent from analysis conducted at Davis lab.

| Region | Age | Davis | Irvine | Michigan | Total |
|---|---|---|---|---|---|
| CB | $< 70$ | 15 | 38 | 17 | 70 |
| CB | 70+ | 6 | 4 | 1 | 11 |
| ANCG | $< 70$ | 0 | 56 | 48 | 104 |
| ANCG | 70+ | 0 | 10 | 13 | 23 |
| DLPFC | $< 70$ | 69 | 9 | 68 | 146 |
| DLPFC | 70+ | 16 | 0 | 16 | 32 |

Table 1: Meta array observations by region, age group, and location

A possible solution would be to ignore possible confounding coming from the lab or brain region variables. On further analysis, this would be a large mistake. In this dataset, the same gene is observed to have wildly different expression levels depending on the region of the brain the observation comes from. This corroborates the idea that different parts of the brain have different functions, as these disparate functions would likely draw on differently expressed combinations of genes.

A related problem is the correlation structure between observations coming from the same individual. Since multiple observations from the same individual are analyzed in either separate arrays or different labs, it is important to understand the balance between the number of observations and the number of individuals these observations come from. By the below figure, we can see that there is a tangible imbalance in both the number of observations and individuals among the two groups. Missing and corrupted data analysis was also conducted by examining the control genes, but no systematic issues were found.

Our proposed solution for adjusting for these data distribution imbalances is the use of separate linear regression models for each brain region. These will lack some of the inferential power that could be acheived by ignoring brain region as a significant confounder, but it is more justifiable from a modelling perspective and causal inference perspective.

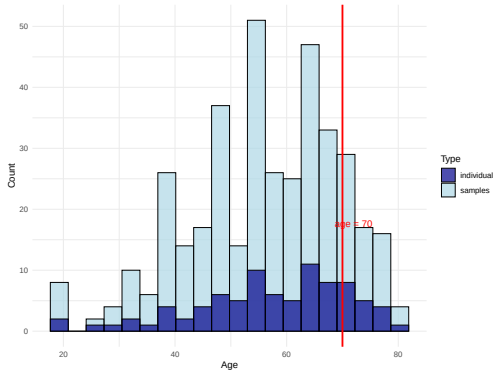


Figure 1: Breakdown of observations and samples by age, old/young cutoff (age 70) highlighted in red

# Methodology

## Simple Linear Regression

Based on the three brain regions (ANCG, CB, and DLPFC), we fit separate models for each region dataset:

$$\text{gene expression} = \beta_0 + \beta_1 \text{age indicator} + \beta_2 \text{lab} + \beta_3 \text{sex} + \beta_4 \text{array version} + \varepsilon.$$

The age indicator equals 1 if age $\geq 70$ and 0 otherwise. Gene expression is measured as $\log_2$ fold change, so a one-unit increase in a predictor multiplies expression by $2^{\beta}$ on average (e.g., $2^{\beta_1}$ for the age indicator).

We analyze regions separately because differential expression depends on tissue: ANCG lies deep in the front–midline of the brain, CB sits at the back/bottom of the head, and DLPFC is on the outer upper frontal lobe. These regions differ in cell-type composition and function, so stratifying by region avoids confounding and implicit interaction effects with other covariates.

Our goal is to identify genes whose expression differs by age within each region. Accordingly, we focus on $\beta_1$ and its associated p-value to assess whether the age indicator is a significant predictor of expression.

# Results

## Simple Linear Regression

By checking the coefficient of age, the numbers of gene expressions that show significant difference with respect to the age were shown as:

Table 2: Number of genes with significant age-associated differential expression by region.

| ANCG | CB | DLPFC |
|---|---|---|
| 2 | 1 | 98 |

**Volcano Plots.** We visualized this result with volcano plots of region-wise effect sizes, which plots the (coefficient, -log(p-value)).

**Boxplots.** Figure 2 shows boxplots of the top significant genes identified from each region-specific model. Each column corresponds to one brain region (ANCG, CB, and DLPFC), and within each panel the expression levels are compared between young (blue) and old (red) groups across all three regions.

For ANCG, genes such as HLA-DPA1 and CD74 show clear difference in the older group. In CB, gene CAMKK2 exhibit small age-related effects. In DLPFC, a larger number of genes display noticeable age-associated differences.

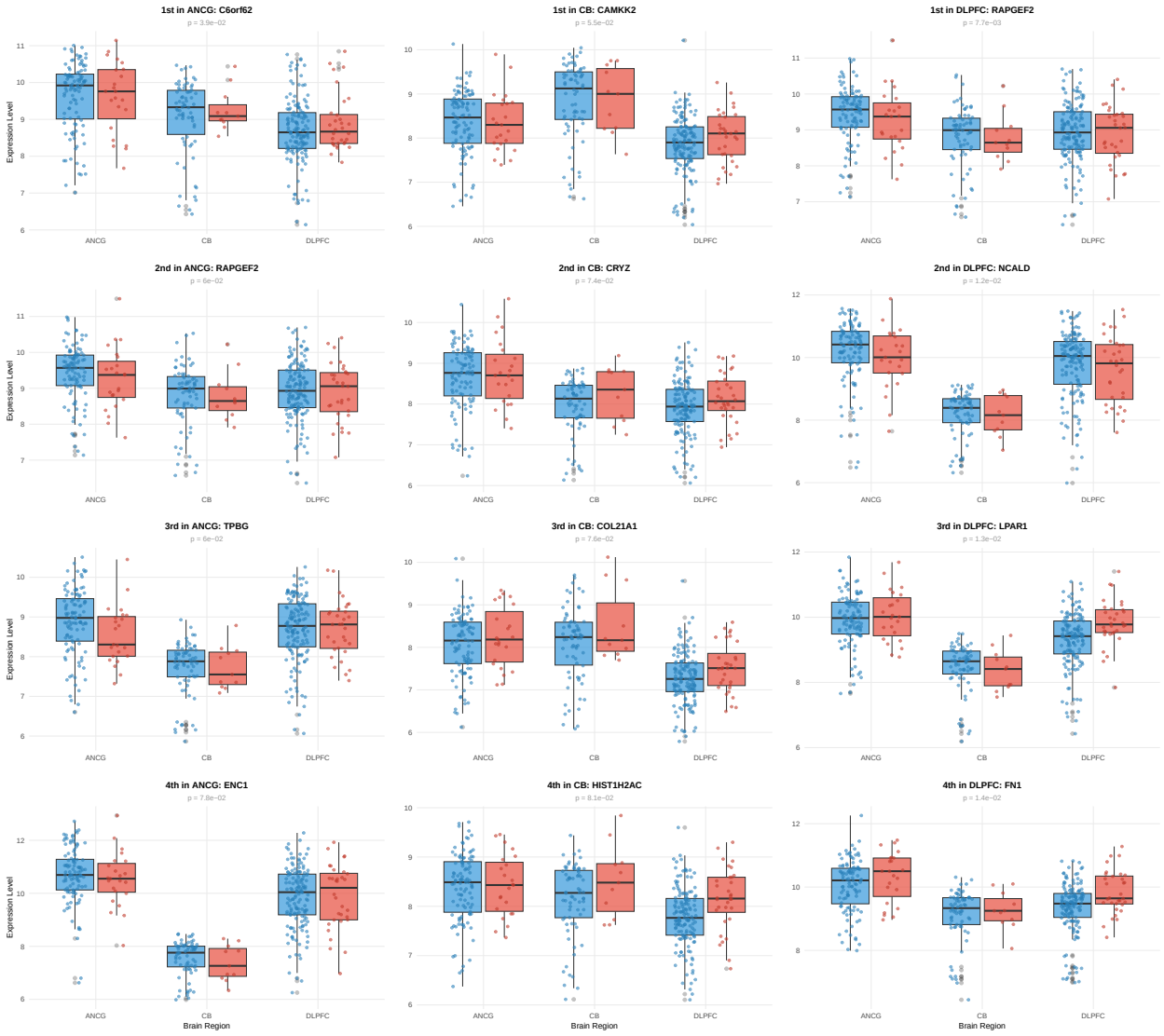


Figure 2: Visualization of the top significant genes from each region-specific model.

Overall, these region-specific patterns suggest that the DLPFC is most transcriptionally sensitive to aging, whereas ANCG and CB show more limited sets of age-related genes. Also, the genes identified as significant within each region were generally not significant in other regions, suggesting that age-related expression changes are largely region-specific.

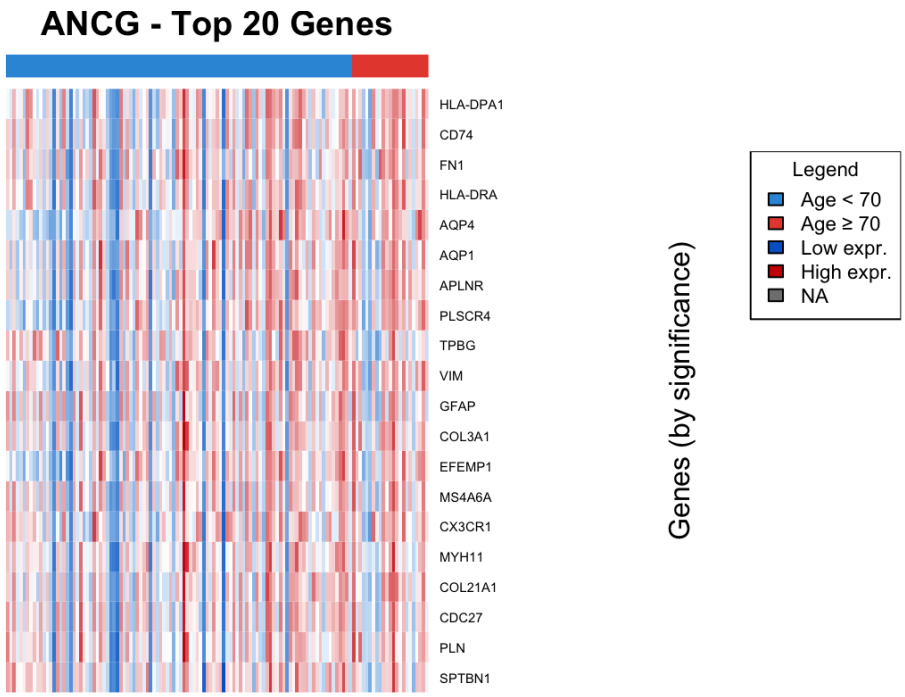


Figure 3: Heatmap of the top 20 age-associated genes in ANCG.

**Heatmaps.** We draw three heatmaps for each model (Figures 3–5), illustrating region-specific expression changes associated with aging. From each model, the top 20 significant genes were selected. Each column in the heatmap represents an individual sample ordered by age, and each row corresponds to one of the selected genes, ordered by statistical significance (most significant at the top).

The DLPFC exhibits the strongest differential expression, ANCG shows moderate immune-related changes, and CB remains relatively stable.

## Conclusion

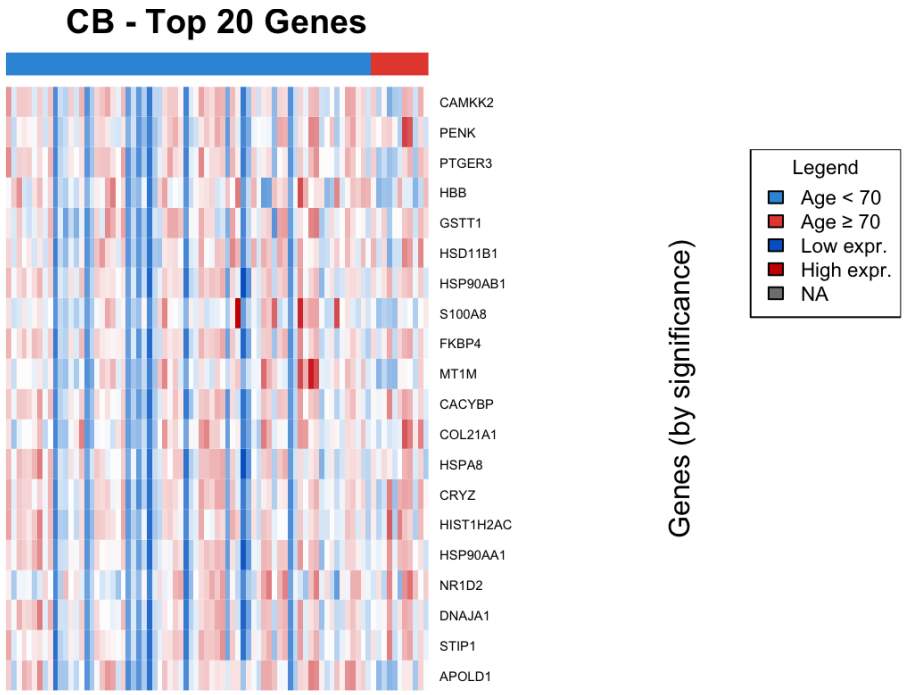


Figure 4: Heatmap of the top 20 age-associated genes in CB.

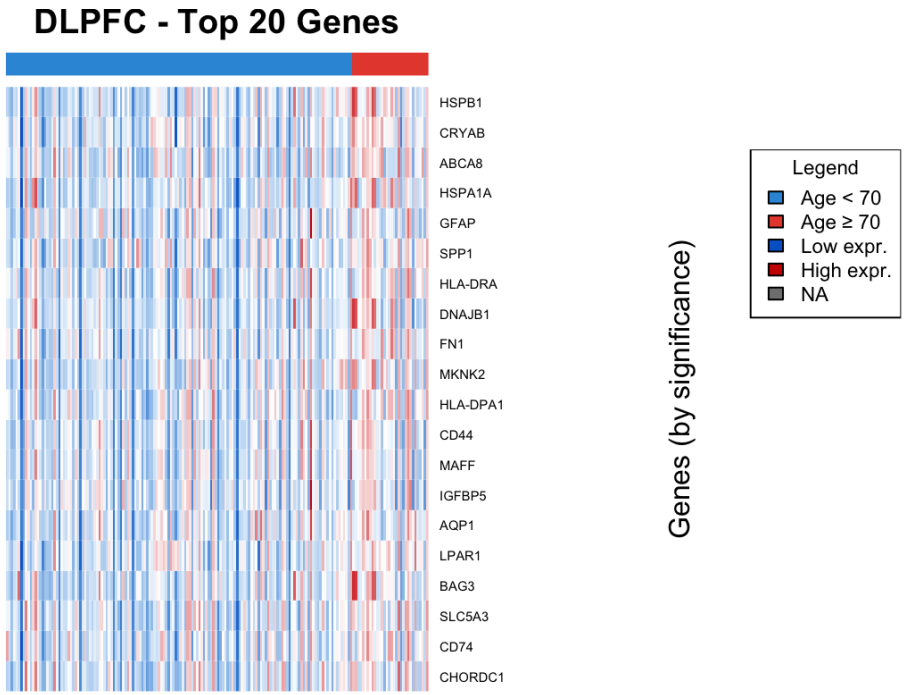


Figure 5: Heatmap of the top 20 age-associated genes in the DLPFC.