Project 2 Executive Report

Jiwoon Han, Hanbin Lee, Yue Yu, Carlyle Morgan

2025-10-10

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 14 genes from across two sites in the human brain. These genes have demonstrated statistically signficant expression between patients less than 70 and those 70 and older at the $\alpha = 0.05$ 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.

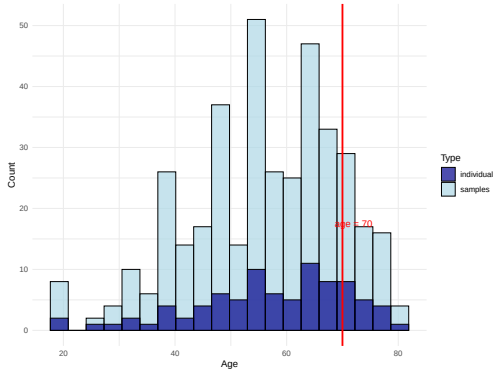

Figure 1: Breakdown of observations and samples by age, old/young cutoff (age 70) highlighted in red

Methodology

Simple Linear Regression

For each of the three brain regions (ANCG, CB, and DLPFC), we fit the following regression separtely:

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

The age indicator equals 1 if age ≥ 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^{β} on average (e.g., 2^{β_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 potential confounding and interaction effects with other covariates.

Our goal is to identify genes in which the expression differs by age within each region. Accordingly, we focus on β_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

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.

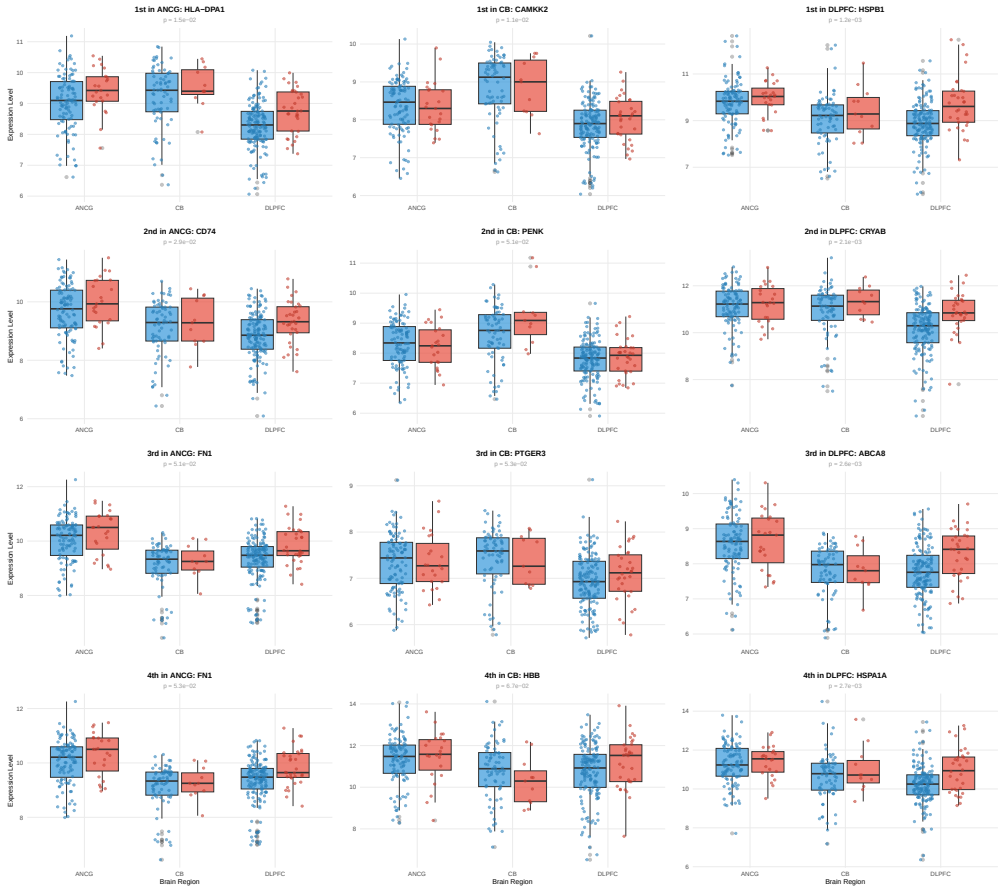

Figure 2: Visualization of the top significant genes from each region-specific model.

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.

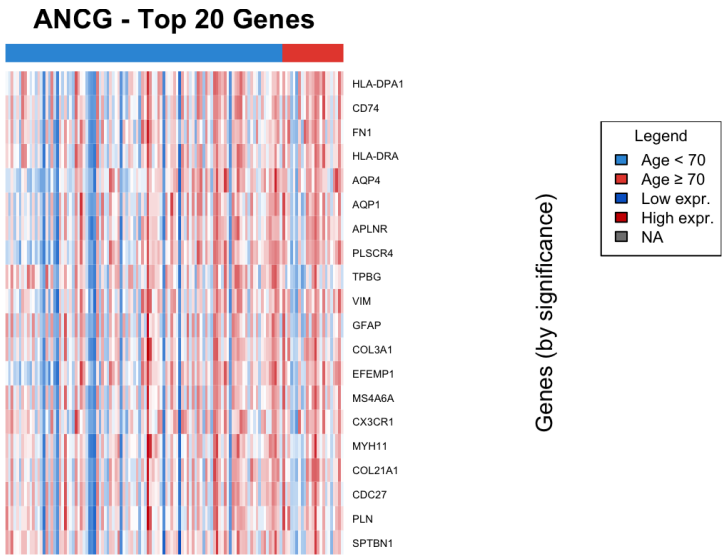

Figure 3: Heatmap of the top 20 age-associated genes in ANCG.

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.

Multiple testing adjustment

In summary, the empirical-Bayes multiple testing results reveal that age-related transcriptional effects vary substantially across brain regions.

The anterior cingulate cortex (ANCG) exhibits the strongest association between gene expression and age, with around 60 significant discoveries at a 5% FDR threshold under both the limma and NPMLE frameworks (Figure 6 and 7). This indicates robust and consistent age-dependent expression patterns in this region. In contrast, the cerebellum (CB) shows no significant findings, suggesting either biological stability or limited statistical power, while the dorsolateral prefrontal cortex (DLPFC) displays only a few significant genes, implying a weaker or more localized effect.

Comparing the two correction methods, the limma moderated t-test (which assumes a parametric inverse-Gamma prior on variances) and the NPMLE partially-Bayes approach (which nonparametrically estimates the prior via Kiefer–Wolfowitz EM) yield nearly identical results, highlighting that the simple parametric assumption is sufficient here. The slight increase in discoveries under NPMLE in ANCG suggests mild across-gene heteroscedasticity but not enough to change the overall conclusions. Together, these results indicate that aging-related transcriptional regulation is region-specific, statistically stable across EB methods, and most pronounced in the ANCG, where age variation and sample size likely confer higher detection power.

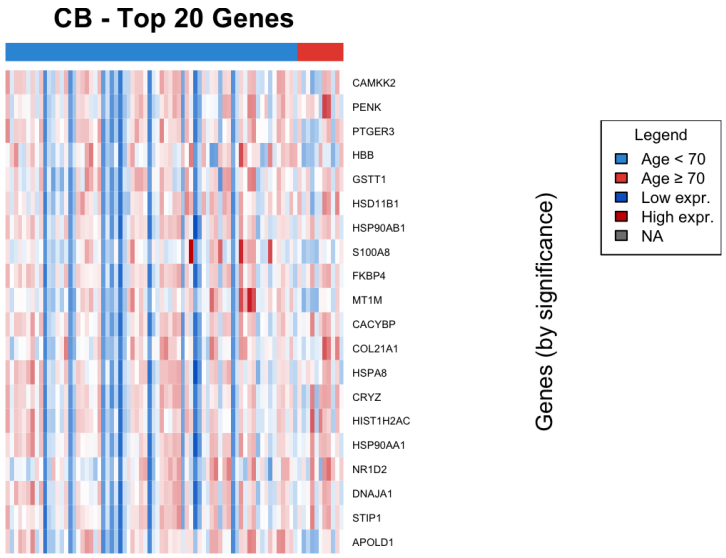

Figure 4: Heatmap of the top 20 age-associated genes in CB.

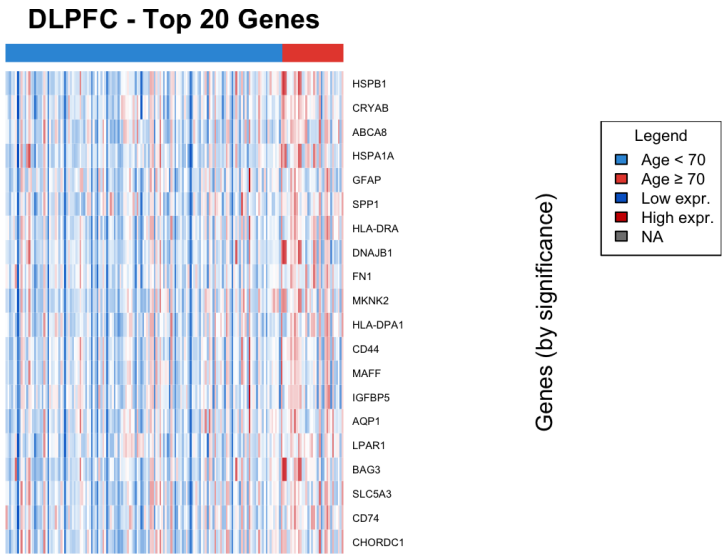

Figure 5: Heatmap of the top 20 age-associated genes in the DLPFC.

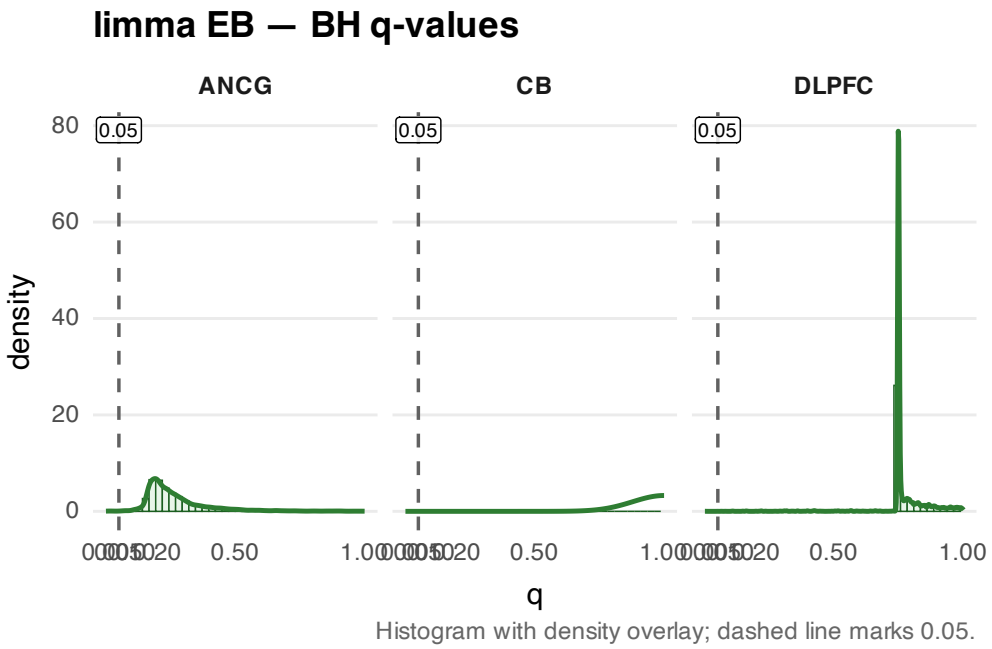

Figure 6: BH q-values from limma empirical Bayes.

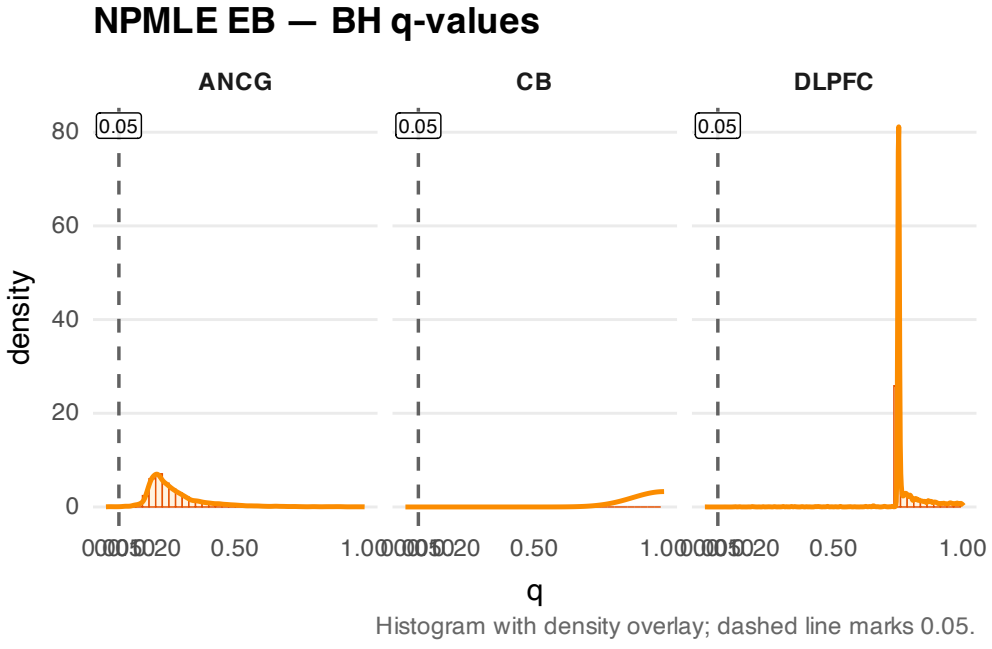

Figure 7: BH q-values from non-parametric MLE empirical Bayes.

Conclusion

We found no statistically significant genes in the CB region.

In the DLPFC region, we found GJA1 and GFAP to pass the threshold. According to the literature, GJA1 encodes a gap junction protein essential for direct cell-to-cell communication and rapid electrical signaling, particularly crucial for the synchronized contraction of the heart and GFAP encodes the main intermediate filament protein of central nervous system astrocytes, providing structural support to the brain and serving as a key marker for cellular response to injury (gliosis). Although GFAP has a known function in the brain, it is unclear why GJA1 appeared in our analysis. Nevertheless, the gene also appeared significant in the ANCG region.

In the ANCG region, we found GJA1, PLSCR4, AQP4, EFEMP1, ID4, NTRK2, AHCYL1, ERBB2IP, GFAP, AGXT2L1, RHOBTB3, AQP1, PPAP2B, and EDNRB to be significant. The two genes, GJA1 and GFAP that appeared in the DLPFC region were also found in this region. For other 12 genes, AQP4 and AQP1 are both genes that carry water molecules and are known to have roles in the astrocyte. NTRK2 encodes a receptor for brain-derived neurotrophic factor (BDNF) that, upon activation, plays a critical role in neuron survival, differentiation, and synaptic plasticity.

We did not find immediate relationship of other genes to the central nervous system.

Appendix

In the appendix, we elaborate the empirical bayes moderate techniques used. Here we use $\widehat{\beta}$ as a proxy of gaussian measurments with a bit loss of rigor. We adopt a unified *empirical partially Bayes* framework for large-scale testing problems where, for each unit $i = 1, \dots, n$, one observes a summary pair (Z_i, S_i^2) with

$$Z_i \mid (\mu_i, \sigma_i^2) \sim N(\mu_i, \sigma_i^2), \qquad S_i^2 \mid \sigma_i^2 \sim (\sigma_i^2/\nu)\chi_\nu^2.$$

The goal is to test $H_i : \mu_i = 0$ while accounting for uncertainty in the nuisance variances. The oracle partially Bayes analyst posits a prior G for σ_i^2 and defines the conditional two-sided p-value

$$P_G(Z_i, S_i^2) = \mathbb{E}_G\big[\, 2\{1 - \Phi(|Z_i|/\sigma)\} \mid S_i^2 \big],$$

which is uniform under H_i both conditionally and unconditionally on S_i^2. The `limma` method corresponds to the parametric choice

$$\frac{1}{\sigma_i^2} \sim \frac{1}{\nu_0 s_0^2}\chi_{\nu_0}^2,$$

leading to the *moderated variance*

$$\tilde{s}_i^2 = \frac{\nu_0 s_0^2 + \nu S_i^2}{\nu_0 + \nu}, \qquad \tilde{t}_i = \frac{Z_i}{\tilde{s}_i},$$

and the moderated-t p-value

$$P_{\text{limma},\,\nu_0,s_0^2}(Z_i, S_i^2) = 2\,\bar{F}_{t,\nu_0+\nu}(|\tilde{t}_i|),$$

where $\bar{F}_{t,\kappa}$ denotes the survival function of a t distribution with κ degrees of freedom. This is exactly equal to $P_G(Z_i, S_i^2)$ for the inverse–scaled–χ^2 prior, thus showing that `limma` is a parametric special case of conditional partially Bayes inference.

The paper generalizes this approach by replacing the parametric prior G with the nonparametric maximum-likelihood estimate (NPMLE)

$$\widehat{G} \in \arg\max_G \sum_{i=1}^n \log f_G(S_i^2), \qquad f_G(s^2) = \int_0^\infty p(s^2 \mid \sigma^2, \nu)\, dG(\sigma^2),$$

where $p(s^2 \mid \sigma^2, \nu)$ is the scaled-χ^2 density. The resulting plug-in p-values $P_{\widehat{G}}(Z_i, S_i^2)$ inherit the oracle's conditional uniformity asymptotically and can be computed efficiently since $\widehat{G}$ is discrete with at most n support points.

The complete multiple-testing pipeline (*Algorithm 1*) is as follows:

1. Estimate G from the sample variances $\{S_i^2\}$—parametrically (as in `limma`) or nonparametrically via NPMLE.
2. Compute the conditional p-values $P_i = P_{\widehat{G}}(Z_i, S_i^2)$.
3. Apply the Benjamini–Hochberg procedure at level α: reject all H_i such that $P_i \le P_{(k^\star)}$, where

$$k^\star = \max\Big\{\ell : P_{(\ell)} \le \alpha\ell/n\Big\}.$$

Under the hierarchical model where $\sigma_i^2 \overset{\text{iid}}{\sim} G$, these plug-in p-values are asymptotically uniform, so BH controls the false discovery rate at approximately $\alpha n_0/n$ with nearly parametric error rate $\mathcal{N}((\log n)^{5/2}/\sqrt{n})$. Even in the compound setting with fixed variances, the same algorithm achieves *average significance* and asymptotic FDR control. Thus, `limma` emerges as the parametric member of a broader empirical partially Bayes family, while the NPMLE variant provides model-robust and theoretically justified p-values for large-scale inference.