# Gene Expression Analysis in Brain Amygdala by Age Group

## Introduction

Aging is a complex biological process that involves changes at the genetic, cellular, and tissue levels. In the human brain, age-related shifts in gene expression can reveal important insights into development, neurodegeneration, and overall brain function. Among brain regions, the **amygdala** plays a key role in emotional regulation and memory, making it a particularly interesting area to explore.

In this project, we used publicly available data from the **Genotype-Tissue Expression (GTEx)** project to investigate **how gene expression in the brain amygdala varies across different age groups**. Our goal was to identify patterns in gene activity that are associated with age, using a combination of data wrangling, visualization, and summary analysis. The approach was exploratory, with a focus on building foundational bioinformatics skills and generating hypotheses for further research.

## Methods

### **Data Sources**

We worked with two key files from the GTEx v10 dataset:

* **Gene Expression File** (gene\_reads\_brain\_amygdala.gct): Contains normalized transcript counts (TPM) for each gene across individual amygdala samples.
* **Phenotype Metadata** (GTEx\_Analysis\_v10\_Annotations\_SubjectPhenotypesDS.txt): Includes important sample-level information such as **age group, sex, and cause of death**.

These files together allowed us to link gene expression profiles with biologically relevant sample characteristics. By analyzing gene expression within the context of this metadata, we were able to explore biologically meaningful subgroups, such as age-defined cohorts, and assess how transcriptional activity varies with age.

### **Data Wrangling and Preprocessing**

All preprocessing was completed in **R**, primarily using functions from the **tidyverse** and **data.table** packages. To streamline the process, we developed a custom R function to import and clean the data. Key steps included:

* **Removing header/comment lines** from both input files.
* **Transposing** the gene expression matrix so that each row represented a single sample.
* **Extracting GTEx sample IDs** and matching them to metadata using the SUBJID variable.
* Performing a **left join** to merge expression data with sample annotations.
* **Filtering out samples with missing age data**, which would have interfered with downstream group analysis.

After merging, we grouped the samples by the AGE variable and calculated the **mean gene expression** per age group for each gene. The data was then prepared for visualization by selecting a manageable subset of genes and applying a Z-score transformation.

### **Visualization**

To make the results more interpretable, we selected the **first 100 genes** in the dataset and performed a **Z-score transformation** to normalize expression levels across genes. This allowed us to highlight relative differences in expression regardless of absolute abundance. We used the pheatmap package to create a **heatmap** showing how these genes' expression varies across age groups. Clustering was disabled to maintain the natural order of the age groups, which helped in visually assessing trends in gene activity across the human lifespan.

## Results

The resulting heatmap displays a high-level view of gene expression across age groups in the amygdala. Some patterns we observed include:

* **Higher expression in younger age groups (20–29):** A subset of genes showed increased transcription in early adulthood, possibly reflecting ongoing brain development or neuroplasticity.
* **Increased expression in older groups (60–69):** Other genes were more active in older samples, which may be related to compensatory processes or cellular aging.
* **Middle age stability (40–59):** Some age groups displayed relatively stable expression across most genes, which could suggest a plateau phase in gene regulation during midlife.

These patterns reflect the dynamic nature of the brain’s transcriptome. While the heatmap was limited to the first 100 genes for simplicity, it served as a powerful tool to visualize meaningful biological trends. Expanding this approach to include all genes, or focusing on specific gene sets (e.g., mitochondrial genes or neural markers), could reveal even more specific age-related signatures.

## Discussion

This analysis demonstrates how publicly available transcriptomic data can be used to explore **biological questions related to aging**. While the scope of this project was exploratory, the process taught us a lot about **data cleaning, merging, and visualization techniques**—skills that are critical in bioinformatics research.

A few takeaways from our project:

* **Sample-metadata alignment is essential.** GTEx metadata is detailed but requires careful parsing to ensure proper linkage with expression data.
* **Z-score normalization** is helpful when comparing expression patterns across multiple genes with different expression ranges.
* The **heatmap provided an intuitive visualization**, but other techniques (e.g., PCA, clustering) could uncover additional structure in the data.

In addition to technical insights, the analysis raised several biological questions. For instance, are the genes upregulated in older age groups associated with inflammatory pathways or neurodegenerative processes? Do younger expression profiles reflect greater synaptic plasticity or developmental regulation? Further exploration with gene set enrichment analysis (GSEA) or pathway databases could help answer these questions.

This project could be extended in several ways:

* Apply the same workflow to **other brain regions**, such as the cortex or hippocampus, to see whether similar age-related expression trends appear.
* Investigate specific genes linked to aging, like **TP53**, **FOXO3**, or **APOE**, to determine their expression trajectories across age groups.
* Incorporate **sex-based comparisons**, or stratify by cause of death, to see if expression changes vary under different conditions.
* Use **unsupervised clustering** to identify co-expressed gene modules and relate them to biological functions.

## Conclusions

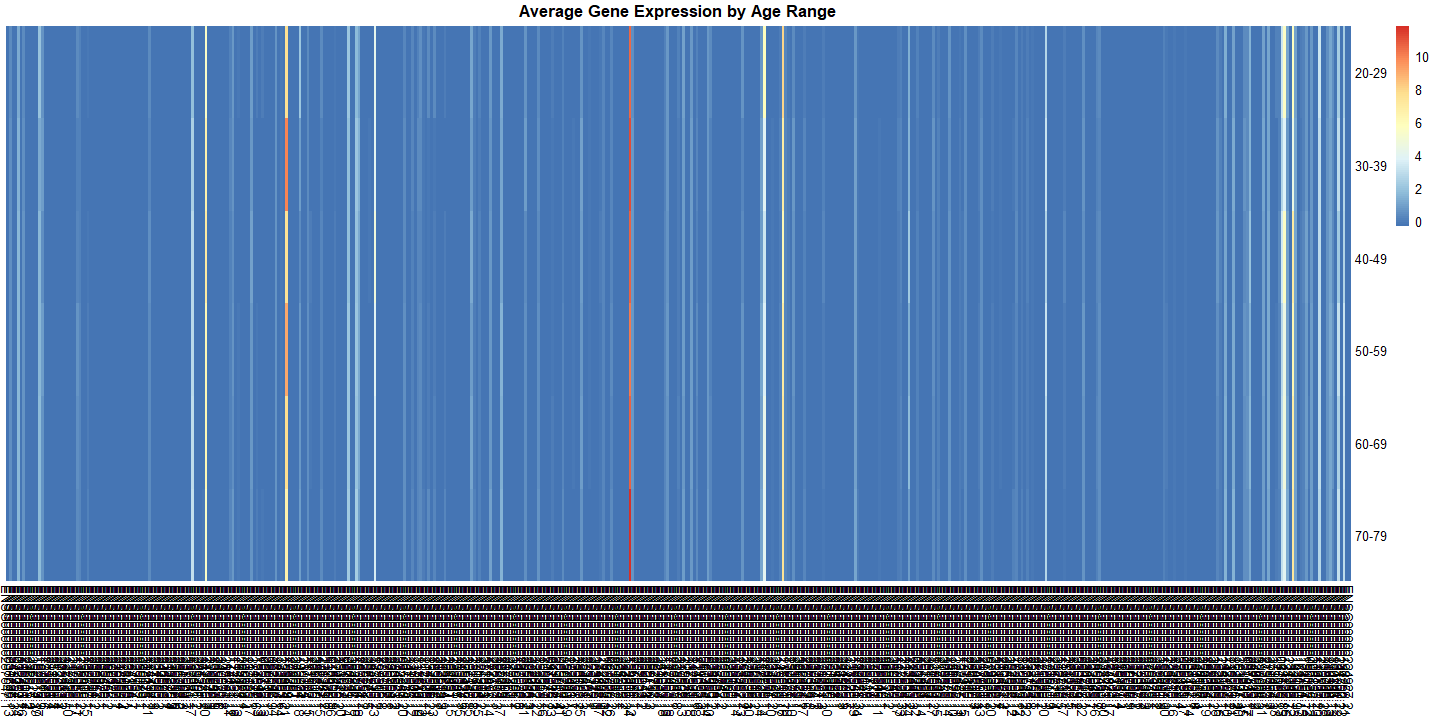
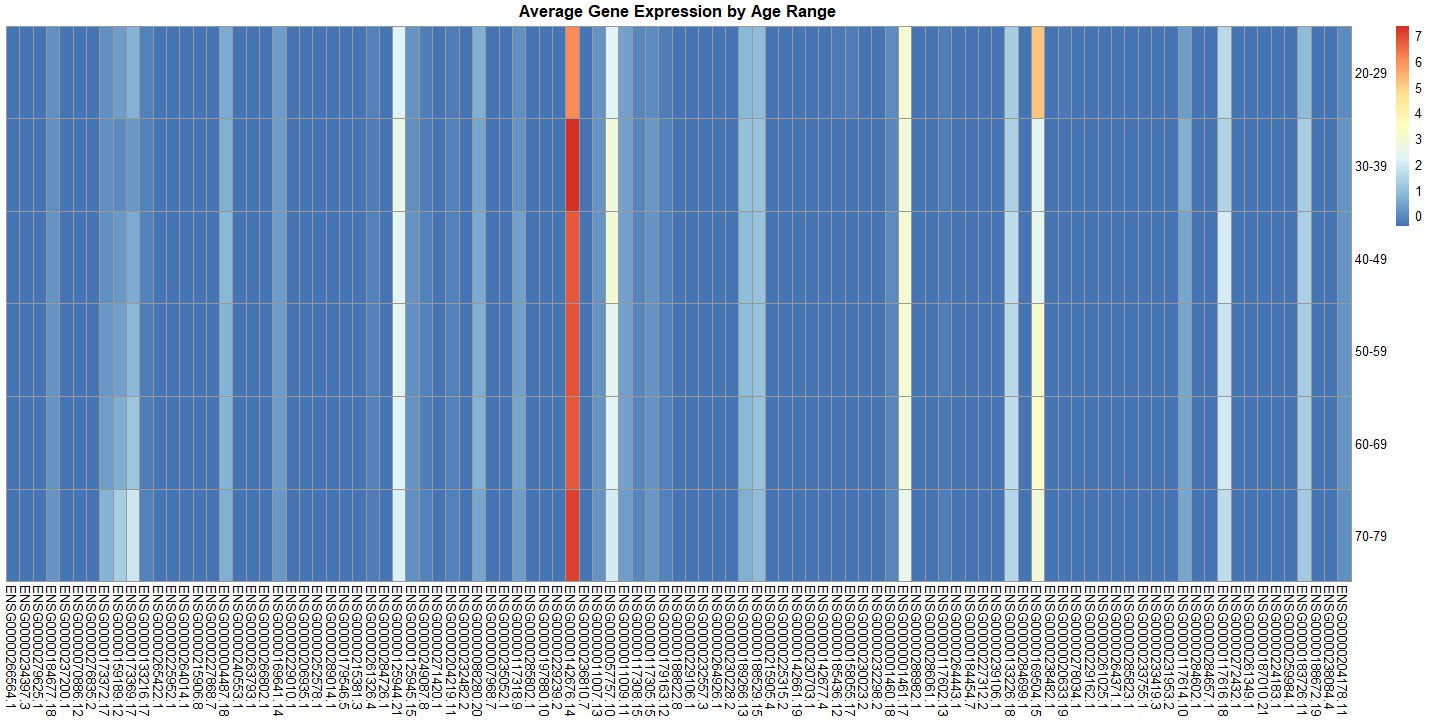
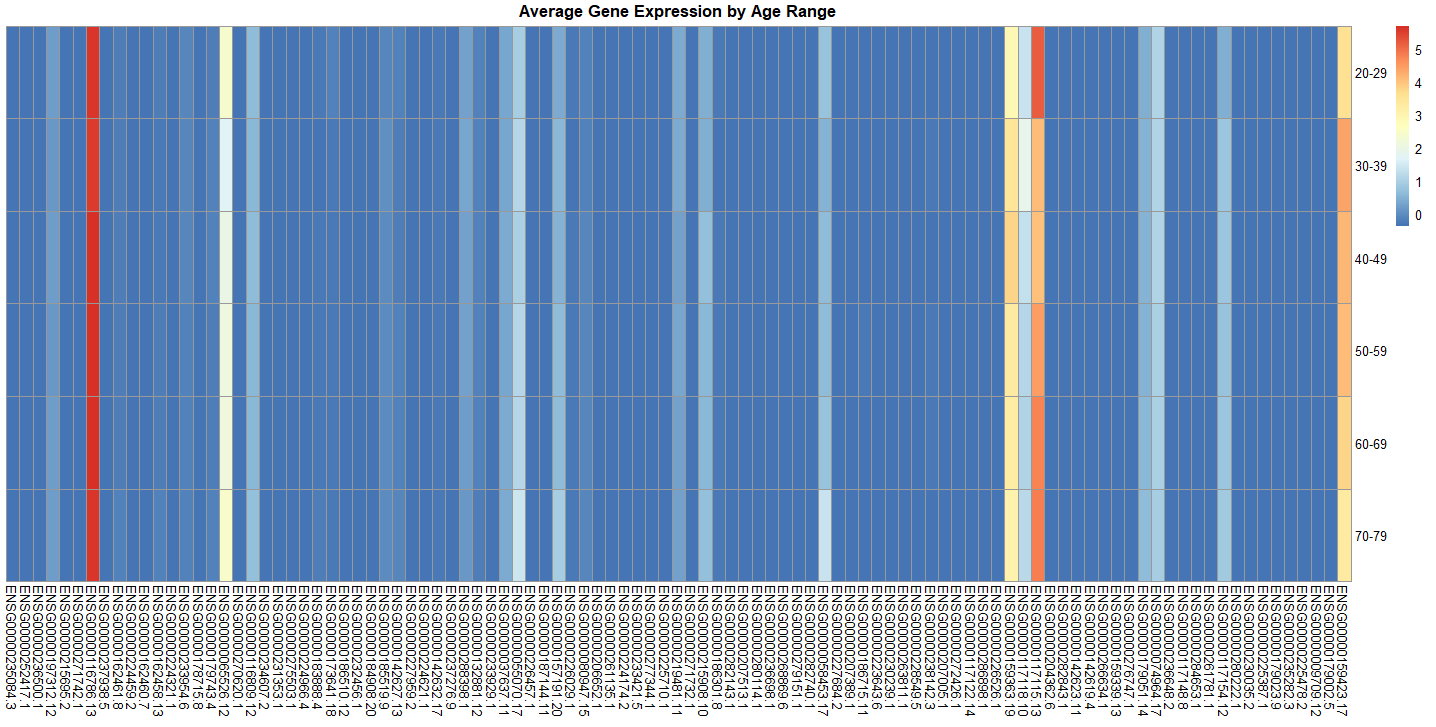
Through this project, we gained hands-on experience working with real-world RNA-seq data and learned how gene expression in the **amygdala** changes with age. Our results suggest that transcriptional activity shifts in distinct ways across the lifespan, and that age is an important factor to consider when analyzing brain gene expression.

The workflow we developed—cleaning raw GTEx files, merging metadata, and visualizing expression by age group—is adaptable and reproducible. We hope that this project serves as a stepping stone for more detailed analyses, such as identifying **tissue-specific biomarkers**, studying **age-related diseases**, or applying **unsupervised clustering** to uncover hidden structure in gene expression data.

These findings also underscore the value of large-scale datasets like GTEx, which provide rich resources for exploring gene regulation across human tissues. Our approach highlights how integrating biological context with statistical tools can lead to compelling and interpretable visualizations. While this study was limited to one brain region and used aggregated data, the same methodology could be scaled to individual-level data to support more personalized insights into brain aging.

## 

## Appendix: Figures

**Figure A1.** *Heatmap of Z-score scaled average gene expression (first 100 genes) across age groups in human brain amygdala tissue.* Rows correspond to genes; columns to age groups. Red indicates above-average expression, blue below-average expression.**

*Split 500-600, 700-800, and 2000-2500 respectively*