Tim Vaughn

CSC-303

Professor Kang

May 8, 2025

The Amygdala vs The Hippocampus: A Statistical Story

The story of the development of gene sequencing and the rise of computing are two narratives that are deeply interwoven with one another. Since antiquity, it has been the dream of scientists and physicians to understand the inner mechanisms of the body and the biological role that each organ plays. However, in recent times, there is one organ that has completely captivated the interest of science and medicine, and that is the brain. Advancements in new technologies such as MRI’s and EEG’s have lifted the curtain on the mysteries of the brain and given scientists the tools to peer directly into brain activity and conversely, the brain’s inner functions. And, at the same time, advancements in large scale data manipulation have given researchers the tools to sequence RNA and understand how gene expression is related to the tissue it is extracted from. This has allowed scientists to probe brain tissue with a biological function already in mind. This has also led scientists to ask very important questions, such as do expressed genes vary between tissue and if so, are there reliable patterns that can reveal which genes are essential to brain tissue function.

This research project will analyze how two different parts of the brain are similar and dissimilar on the level of gene expression: the amygdala and the hippocampus. Both tissues have been extensively studied and their structures mapped. The hippocampus is understood to be the key actor in memory formation while the amygdala is known to be the center of emotional responses. Both are located in the center of the brain and develop during the embryonic stage of fetal development. By focusing purely on gene reads of samples instead of structural differences, scientists can get a wholly quantitative understanding of the two tissues.

First, the gene reads of sample tissues from the hippocampus and amygdala were downloaded from GTEx and their sum totals across individuals tabulated. Genes that were listed but weren’t expressed above a threshold of 150,000 were dropped from the dataset. Many of these dropped genes were pseudogenes, or genes that had lost their biological function due to accumulated mutations and no longer served a discernable purpose. While these pseudogenes could contain biomarkers, none were expressed frequently enough to be reliable. The top twenty most expressed genes, or genes with the highest counts, were identified.

A high gene count means that the gene is encoded many times within a tissue, which is an indicator that it is functional and plays a mechanism. Looking at the gene reads of the hippocampus and the amygdala, the first thing that becomes apparent is the similarity of commonly expressed genes. They share the top five most expressed genes(in exact order) and the rest of the top twenty common genes are combination’s of each other. This must be because these genes are essential to the functioning of the respective tissues. While this may seem to be a good development at first, these commonly expressed genes are not unique to any particular tissue and are typically found to be most among the most commonly expressed throughout the brain. Profiling a tissue with only the knowledge of the genes most common would be like profiling the make of a car by asking whether or not it had doors and four sets of wheels. In order to make proper predictions as to what genes are essential to the two different tissues, the integer count of each must be carefully analyzed. This was accomplished by joining the two datasets together with an intercept, meaning that both genes are had to be present before being merged. This was necessary because while underrepresented expressed genes were removed from both sets, some genes that were just barely represented in one dataset were absent in others which caused several rows to be appear as having missing values, making an automated prediction model impossible. Once this was accomplished, a regression model using only the three most popular genes and their reads from two unlabeled samples could predict with 99.77% accuracy whether a sample was from the amygdala or the hippocampus. This works with nearly every any combination of gene reads fed into the model with two exceptions: genes with low reads(which were mostly removed) and genes that had extremely high variance in sample counts.

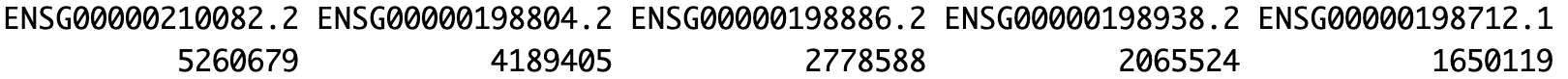
What does this development mean? How can the gene read count be so effective as a predictor of tissue, even when the ordering of commonly expressed genes is practically identical? The answer is in the consistency of values for individual samples which allows for gene reads to be differentially marked. The ranges of gene reads may differ by tens of thousands, but they will do so in a patterned manner across samples which speaks to the biological function of encoding a gene many times. Expressed gene reads rarely had such varying gene counts such that the range of reads fell into a different range of powers of ten. In fact, the count of a single gene (and possibly metadata such as age and sex) could probably be enough to give a researcher an idea as to which area of the brain the sample came from.

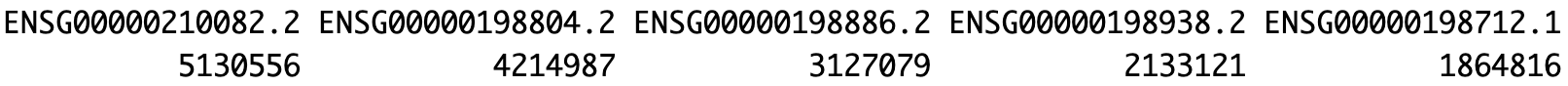
A graph with numbers and lines

AI-generated content may be incorrect.

A graph with numbers and a line

AI-generated content may be incorrect.





A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A black text on a white background

AI-generated content may be incorrect.