**Analysis of Gene Expression Data Comparing Age, Sex, and Median Gene Expression Across Tissues**

Matthew Palmgren

It can be said the field of genetics first sprung into existence at the discovery of DNA in the 1860’s. As time passed, technology improved and a renaissance in genetics occurred in the mid-20th century as Watson, Crick, and Franklin helped to visualize the structure of DNA, while many others began making key discoveries in gene expression at the same time. These days it seems we are in the midst of a second genetic renaissance as computational biology and bioinformatic programs have turned years of analysis into seconds of waiting for a loading circle to finish spinning. These advancements have had massive implications in the medical field as symptoms can now be correlated to the expression of just a few genes. As the central dogma says, these genes produce proteins and with proteins, we can follow their metabolic path and make crucial lifesaving drugs. Because of the important implications associated with genetic data, there is ample funding to collect as much data as possible on genes, when they are expressed, where they are expressed and how much they are expressed. Websites like GTEx Portal are places where this data is compiled so that researchers from around the globe can analyze and contribute to our overall understanding of the genome and gene expression.

In this project, two datasets from GTEx portal and corresponding meta data were analyzed in R to try and gain some insights on gene expression across various tissues and what factors can influence gene expression. The first data set contained the median gene expression values of various tissues for a thousands of genes. The objective was to understand how much and how little genes are expressed across various tissues and to try and identify any correlation between gene expression profiles in different and similar tissue types. The second data set was a subset of subjects who donated hippocampus brain tissue and their levels of gene expression for various genes. From this the objective was to try and see if it was possible to make predictions on meta data variables such as Age and Sex based on corresponding gene expression data. The hippocampus is the region of the brain in charge of memory and learning and so identification of trends in gene expression across age and sex could not only have medical implications but even contribute to our understanding of how we live our day-to-day life.

Data was first read into R straight from URLs or from text files, and then cleaned and tidied to fit the needs of various statistical tests. This primarily consisted of transposing the hippocampus tissue data so that subjects became rows and genes became columns, as well as interjoining the meta data onto the transposed data table. It’s important to note that the meta data pertains to all tissue donors of the study and not just hippocampus tissue donors. Preliminary filtering of the meta data for just hippocampus tissue donors was necessary for interjoining the meta data with the hippocampus tissue data.

Next, exploratory data analysis of the median gene expression data began by looking at the structure of the data. Principle Component Analysis was run and a scree plot (Figure 1) was used to visualize and interpret the results. Roughly 40% of the variability in the data can be explained by PC1 which suggests a slight trend in gene expression for certain genes among tissues. PCs 2,3, and 4 account for roughly another 40% of the data combined. Although on the lower end of what could be considered sufficient explained variability that is worth analyzing, further investigation was conducted for as to what this trend could be. A biplot (Figure 2) was created to further visualize variability explained by PCs 1 and 2. The biplot showed different tissues in black and different genes in red. Both groups were clustered close together toward the 0.0 intersection of the biplot indicating very few genes influenced by PC1 and PC2. Despite the overwhelming majority being clustered into this region of the plot, there are a few longer arrows pointing to the upper right-hand corner of the graph demonstrating that they may be key drivers in the explained variance seen in PC1. The top 10 loading values were then taken and the driving genes (PRSS1, HBB, INS, PRSS2, REG1A, CPA1, CLPS, CELA3A, HBA2, PNLIP) were further researched for their function. The 10 genes identified play important roles in blood cell formation and regulation as well as metabolic pathways associated with digestion (mainly serine proteases) ([GeneCards.org](https://www.genecards.org/)). This would make sense as many enzymes created for digestion are highly specific and only function for certain regions in the body. Gene expression for these enzymes would have to trend higher in certain tissues associated with digestion compared to other tissues such as the brain amygdala or ovaries. Interestingly apart from HBB and HBA2 which are genes that code for Beta globin and Alpha globin subunits of Hemoglobin, the rest of 10 identified genes are all genes that code for enzymes and hormones that are released from the pancreas.

Further evidence for this line of reasoning was obtained from basic summary statistical tests. Tests to determine which tissues expressed the most and least number of genes were run and found that Testis sample tissue had the most genes expressed while Whole Blood samples had the least genes expressed. Thinking back to the top 10 genes driving PC1, both Alpha and Beta globin which are blood cell subunits appear. This could potentially mean that Whole Blood cell gene expression is highly specialized and the genes that are expressed are mostly unique to this tissue type. Further testing analyzed tissues with the greatest ranges in expression values and identified Pancreas Islets to have the greatest range in expression values and Artery Aorta to have the smallest range in expression values. Pancreas Islets being the tissue with the greatest range in expression values is worth noting especially since the majority of the top 10 PC1 driving genes are expressed in the pancreas. This could possibly signify that in addition to expressing highly specific genes the pancreas also has control over a wide majority of other metabolic functions as well. Further literature review of the function of the pancreas and blood cells is recommended to validate these ideas and queries.

Additional testing also revealed the most commonly expressed gene across all tissues which was identified to be the WASH7P gene. The gene is a pseudogene and therefore is not a protein coding gene however it plays a role in the regulation of the WASP gene family which is active in a multitude of cellular processes ([NCBI](NCBI%20WASH7P)).

The data set containing the median gene expression values of various tissues was also used to look at the similarities and differences in the gene expression profiles of sex organs. Both play a role in human reproduction yet serve very different purposes and so examining their gene expression profiles could potentially reveal an interesting correlation. A data frame of the two tissue’s gene expression values was created and a scatter plot was then used to visualize the similarity in gene expression between the two (Figure 3). An identity line was plotted in the center of the graph to see which genes have similar expression in both tissues. Analyzing the graph, it appears that the gene expression profiles of the two tissues are more different than similar. The majority of the points cluster together where the two axis’s meet indicating low expression for a majority of the genes. There are quite a few points that are plotted nearly along the y-axis meaning that there are many genes that are moderately expressed in the testis that are hardly expressed in the ovaries. On the contrary, there are very few genes plotted closely along the length of the x-axis yet they show high expression levels of these genes. These few genes are therefore highly expressed in the ovaries yet not expressed at all in the testis. A few other genes are scattered throughout the plot at varying levels of expression for both ovaries and testis but none fall too close to the identity line. From this it appears that gene expression between the two tissues is very dissimilar, which indicates that the two serve very different purposes despite both being sex organs. Despite the initial observation from the scatter plot, a calculated correlation coefficient (ρ = 0.775753) seems to indicate the opposite. However, based on the visual results of the graph, the majority of the genes appear to have low expression values for both the ovaries and testis so it is possible that these values are driving the correlation coefficient to be higher than reality.

Finally, a correlation heatmap (Figure 4) for the median gene expression value of each tissue was created to further visualize the relationships between tissues and their gene expression profiles. Since the dataset contains tens of thousands of genes, in order to comply with R memory restrictions, only the top 100 most variable genes in the data set were used for the heatmap. Once again it appears that the vast majority of the tissues share similar expression profiles (red) with the exception of a few select tissues which appear to be different. Once again, the culprit seems to be pancreas tissues which are more dissimilar in their expression profiles when compared to the rest. Further analysis revealed that the two most correlated expression profiles both come from Brain Basal Ganglia tissue and the two most dissimilar expression profiles are of the Whole Blood Cell and Liver Pancreas tissue.

The second data set used was an interjoined data set of hippocampus tissue gene expression profiles for several tissue donors and corresponding metadata for the tissue donors. The hippocampus is the region of the brain responsible for memory and learning and so interesting patterns might arise in Sex and Age meta data. First, data was cleaned and normalized to ensure that the ensuing models would run smoothly. The overall structure of the data was first visualized using box plots of the most widely expressed gene in the data set across all tissues (WASH7P). Since the gene is commonly expressed in all tissue types the expectation is that there is little variation within its expression in hippocampus tissue across age and sex. The first boxplot comparing gene expression of WASH7P by sex in the hippocampus (Figure 5) showed extremely similar plots for both male and female with near identical medians. A t-test comparing the two groups confirmed no significant difference in expression (p=0.7948). The box plot figure comparing gene expression of WASH7P by age (Figure 6) showed a decreasing trend in the median gene expression as age increased. It should be noted for this plot that fewer tissue samples of brain hippocampus tissue exist for younger age groups (20-29, 30-39, 40-49) and this lack of data could be represented in the plots. A linear regression model was then created to further analyze trends in WASH7P gene expression across age groups. The model identified moderate change in expression for the 60-69 and 70-79 age groups (p=0.0425, p=0.0291), however small multiple and adjusted R squared values (Multiple R^2=0.03076, Adjusted R^2= 0.008218), as well as a large p-value (p=0.2388) indicate that the model is a poor fit, likely due to a lack in data for the younger age groups.

Heat Shock Proteins (HSPs) are proteins found in the brain that respond to stressful stimuli and some studies have found that sex hormone levels can affect HSP expression ([Romani, W., Russ, D., 2013](https://pubmed.ncbi.nlm.nih.gov/23821238/) ). A total of 15 HSP genes were identified and the difference in expression between male and female was visualized using a boxplot. Out of the 15 genes identified only one (HSPE1P8) was seen to show any real difference in box plots of male and female expression (Figure 7). A linear regression model was run for HSPE1P8; however, no conclusive information could be gathered as once again the model proved not to be a great fit for the data.

Next, PCA was run for both sex and age using the top 1000 most variable genes and a scatter plot was made using R graphing features. The plot shows a dense cluster of similar expression profiles but as variability increases so does the randomness of the points, and neither sex (Figure 8) nor age (Figure 9), showed any clustering by their respective meta data. A similar story was told with K-means clustering of 3 centers using a PCA plot (Figure 10). The plot is shown to be fanning outward as variability increases indicating a gradient or spectrum-like structure in expression across samples but the clusters seem to be organized as just as variability of the samples increases. Hierarchical clustering was run to confirm these results and demonstrated the nested nature of the data but again the variability increases dramatically as the second and third clusters are plotted (Figure 11). Tables created tested clustering against meta data did not appear to show any bias in sex or age across clusters.

Lastly, classification modeling using Random Forest classification was used to try and predict age based off of gene expression data. Due to the large nature of the dataset, a random sample of 100 subjects and 50 genes were used to train the data. The model also used 500 trees in the data training. The results indicated that the model could predict age with a degree of accuracy hovering just over 50 percent. A lack of samples in the age classes 20-29, 30-39, and 40-49 was a real limitation to the model and further tests should be conducted with a larger sample size. Interestingly, a Variable Importance Plot (Figure 12) highlighted two genes, CACNG7 and WWC2 to have high mean decrease accuracy and mean decrease Gini. values indicating that these two genes were most important in the model’s predictive capabilities. The gene CACNG7 corresponds to the calcium voltage ion channel and has been linked to certain psychological disorders ([GeneCards.org](https://www.genecards.org/cgi-bin/carddisp.pl?gene=CACNG7)). The exact function of the WWC2 gene is currently unknown but some studies show that it may be correlated with cancer ([NIH.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC7829927/)). It’s interesting that the model picked out these two gene when determining age, especially as cognitive function decreases and cancer probability increases as age increases.

The data analyzed in this report helped look at how gene expression can vary from tissue to tissue and even looked to see if it could identify and biological trends in gene expression for age and sex. Time constraints and computational strength constraints restricted what could have been accomplished for this project and so it is recommended that future studies build off of these results and analyze more data and research more into potential correlations between tissues and gene expression profiles, especially for pancreas and whole blood tissue samples. Data sets such as the ones provided by GTEx portal give us invaluable information that can help us make discoveries that not only impact the medical community but also better our understanding of the human body.

**Appendix**

A graph with green dots

AI-generated content may be incorrect.

Figure 1

A graph with black and red lines

AI-generated content may be incorrect.

Figure 2

A graph of a graph showing the difference between male and female

AI-generated content may be incorrect.

Figure 3

A red and black text

AI-generated content may be incorrect.

Figure 4

A graph of a diagram

AI-generated content may be incorrect.

Figure 5

A graph with numbers and a line

AI-generated content may be incorrect.

Figure 6

A graph with a number of dots and lines

AI-generated content may be incorrect.

Figure 7

A diagram of a number of dots

AI-generated content may be incorrect.

Figure 8

A diagram of a number of dots

AI-generated content may be incorrect.

Figure 9

A diagram of a graph

AI-generated content may be incorrect.

Figure 10

A diagram of a graph

AI-generated content may be incorrect.

Figure 11

A graph of numbers and letters

AI-generated content may be incorrect.

Figure 12