Henry Lock

Final Project

STT 303

Analyzing tissue-specific gene expression from GTEx through linear regression, logistic regression, and PCA k-means clustering

Introduction

The ability to organize and analyze large amounts of genetic information is essential in growing our understanding of the human genome, characterizing significant genes, and identifying patterns related to genetic diseases. With hundreds of thousands of genes in certain datasets, such as those found in the GTEx Portal, choosing a select, significant set to dive deeper into and research is one of the first challenges. One way to do this is by identifying the most variable genes in a set. This means selecting the genes that differ the most between individuals when it comes to how often they are expressed and transcribed. Too much or too little expression from certain genes can be directly linked to diseases, disorders, and conditions. Therefore, looking at genes with high expression variability could help identify underlying causes, or increase our ability to identify diseases before the symptoms manifest. Even with this initial gene sorting, there could still be too much information to process and further categorization could be needed.

Sorting genes into what tissues they originated from is also an effective and meaningful way to sort gene data. Knowing what tissue in the human body the gene is linked to can help researchers investigate specific disorders such as myopathies in skeletal tissues and hormonal

imbalances in pituitary tissues. The genes with the highest variability within each tissue type can be the cause of these disorders. After these genes are selected, identified, and characterized, the goal is often to find trends across the gene dataset sample and be able to reliably apply those trends to the greater human population. Linear regression, classification through logistic regression, and clustering through PCA and k-means models are all statistical techniques that attempt to find these key trends.

In this study, we aim to do exactly that. The top ten most variable genes were identified in five different tissue groups. Linear regression was used to try and find a pattern between variability, age, and sex; logistic regression attempted to classify genes into sexes based on expression; and PCA/k-means explored if expression could be used to cluster genes into age, sex, and/or tissue type. These three techniques and the results attempt to discover those relevant patterns to disease, and do it with a sufficient dataset.

Methods

The main resource used in this project were the gene TPMs(Transcripts per million) by tissue files in the GTEx Analysis V10 release dataset. To start, five gene TPMs files for five different tissue types were randomly selected. This resulted in the selection of gene TPMs for: the brain - frontal cortex (BA9) (gene_tpm_v10_brain_frontal_cortex_ba9.gct.gz) , the heart - left ventricle (gene_tpm_v10_heart_left_ventricle.gct.gz), the lung (gene_tpm_v10_lung.gct.gz), the muscle - skeletal (gene_tpm_v10_muscle_skeletal.gct.gz), and the pituitary (gene_tpm_v10_pituitary.gct.gz). As stated, these tissue groups were chosen randomly. After the selection, the files were uploaded to RStudio Version 2024.12.1+563 where the data was organized and analyzed.

First, the gene TPM files, along with the metadata files that describe the genes, were read and stored. Next, the top ten most variable genes with sufficient samples (20) were determined and displayed on a bar graph for each tissue type. For each of the top ten genes, a linear regression was then run comparing age to gene expression, separated by sex. The level of significance between the slopes of regression lines for male and female of each gene was then determined using a t-test to compare the interaction coefficients of each linear regression slope. These results were displayed on a volcano plot. Next, a logistic model was run for each tissue type, attempting to classify each gene sample as male or female. The results of this classification were compared to the actual data labels of each sample using a confusion matrix. This same classification model was then run for all of tissue samples, across all tissue types. Following the classification model, three PCA and k-means clustering method analyses were administered on the top 1000 most variable genes from all five tissue sample files. These genes were filtered for variability and usable data(NA and 0 variance data removed), transposed to view each gene as a variable, and z-score standardization was used to scale the data. This was run for two, three, and five clusters, and the results of each were compared with the actual sex, age, and tissue labels to see if the clustering was effective and grouped the genes into pre-established groups.

Results

The results for the top ten variable genes for each tissue type can be seen in figures 1-5. The linear regressions run for each of these genes can be seen in figures 6-10 and the volcano plots showing their significance can be seen in figures 11-15. These plots show no significant difference between male and female regression lines except for four skeletal muscle tissue genes. The results of the logistic regression can be seen in the confusion matrices figures 16-21 and are summarized as follows: the brain - frontal cortex (BA9) (196/259 male predictions 75.68%, 5/10

female predictions 50%) , the heart - left ventricle (295/420 male predictions 70.24%, 19/32 female predictions 59.38%), the lung (408/594 male predictions 68.69%, 4/10 female predictions 40%), the muscle - skeletal (523/765 male predictions 68.37%, 24/53 female predictions 45.28%), the pituitary (227/307 male predictions 73.94%, 4/6 female predictions 66.67%), and all tissues combined (1676/2387 male predictions 70.21%, 41/69 female predictions 59.42%). Following that are the results of the k-means clustering on expression PCA graphs for two, three, and five clusters. These are seen in figures 22-24. The confusion matrices plotted for these clustering graphs are seen in figures 25-33 and show how the model's clustering compares to the pre-established age, sex, and tissue type labels.

Conclusions

Overall, this project showed varied success in finding significant patterns across the top variable genes of the five selected tissue types. The linear regression model showed almost no significant difference in how gene variability changes over time between males and females. The only exception was the skeletal muscle tissue samples. The patterns between sexes have the potential to reveal sex-specific aging effects which could help to establish different treatment plans and testing for males and females. This would be beneficial for genes that cause diseases which show their effects later in life. Further research should be done into the significantly different skeletal muscle genes found to determine if this is the case.

For the classification model, the male data was much more revealing than the female data. Having only 69 female samples across all the tissue types for the top variable genes is simply not enough. A more robust dataset containing a larger female representation is needed to see if the overall 59.42% prediction rate would change for a larger population. For the male samples, a 70.21% correct prediction rate of sex for 2387 samples shows much more promise.

This shows there is potential for individual genes to be strong predictors of sex. If those same genes are linked to a disease, having this strong prediction metric could mean the mechanism of the disease is sex-related. This could lead to a better understanding of the disease and a potential way to help affected individuals.

All three clustering method trials compared to the pre-existing labels show a similar trend. Tissue type strongly drives clustering while age and sex are much looser. The confusion matrices show tissues were almost always put into the same cluster, and as clusters increased, the tissue types were also isolated into a cluster. This shows how gene expression varies more for these genes across the tissue types than it does across age or sex. One potential take away from this clustering ability is understanding organ-specific diseases.

Within each of the tissue types, one of the most variable genes was selected and studied in how it relates to a condition in that organ. The brain frontal cortex BA9 gene ENSG00000198712.1 aka MT-CO2 encodes a component of the cytochrome c oxidase protein which plays a role in cellular respiration, key in ATP production. A mutation in this gene can result in MELAS (Mitochondrial Encephalomyopathy, Lactic Acidosis, and Stroke-like episodes). One of the most variable genes across multiple tissue types (lung, skeletal muscle, and brain) was ENSG00000198804.2 aka MT-ND1. This and the gene ENSG00000198886.2 aka MT-ND2 are both mitochondrial DNA that encode a subunit of the NADH dehydrogenase protein which is essential in ATP production. Mutations in these genes can cause a variety of issues including MELAS and Leber Hereditary Optic Neuropathy (LHON). For the skeletal muscle tissues, gene ENSG00000143632.15, aka ACTA1, encodes alpha-skeletal actin which helps skeletal muscle fibers contract. Mutations in this gene can lead to various myopathies. Finally, the pituitary gene ENSG00000172179.13, aka PRL, encodes the prolactin hormone

which plays key roles in lactation signalling and reproduction functions. Unbalanced levels of PRL expression can lead to infertility, menstrual disturbances, and other reproduction-related consequences.

This disease information for high-variable genes links to this project because the proven potential of a clustering model for tissue type could help lead to sub-clusters within these tissue types that predict organ-specific diseases. If a similar model was run for the genes linked to diseases such as PRL, could clusters predict which genes are more at risk, and which genes respond better to different treatments? This project suggests that there is a potential for clustering tissue types, and further investigation in future studies should be conducted to test this. In conclusion, the clustering method seems to have the most potential for real-world application modeling techniques. While logistic regression showed some potential in skeletal muscle tissue, more research should be done to see if these seemingly sex-linked expression changes over time could help with disease detection and regulation. For classification, more female samples need to be tested to prove the efficacy of female classification, but male prediction could help study sex-linked diseases. This project may only show an introductory approach to modeling gene expression data, but it highlights the potential for future studies in early genetic disease detection, underlying sex-linked disease mechanisms, and cluster-specific treatment plans.

Appendix

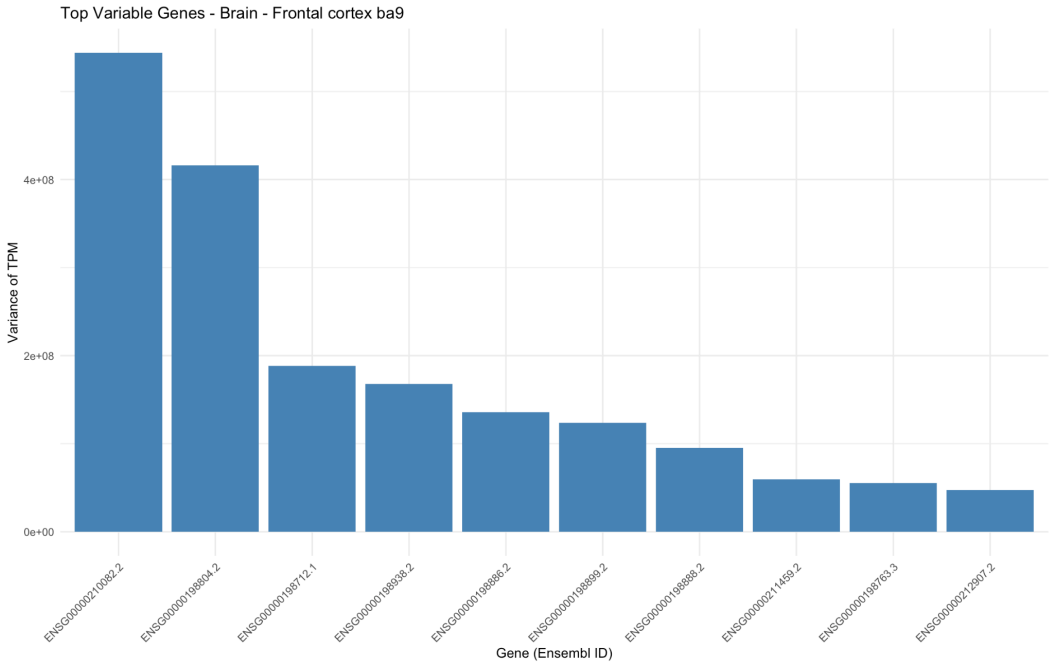

Figure 1. Top ten variable genes by variance of TPM in the GTEx gene TPMs for Brain - Frontal Cortex BA9

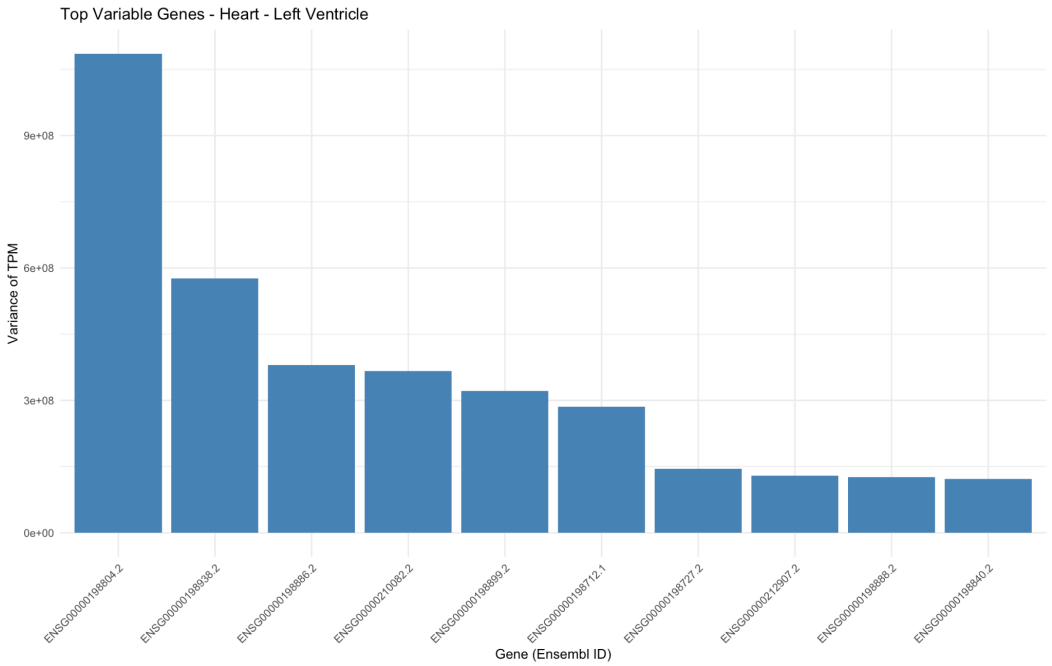

Figure 2. Top ten variable genes by variance of TPM in the GTEx gene TPMs for Heart - Left Ventricle

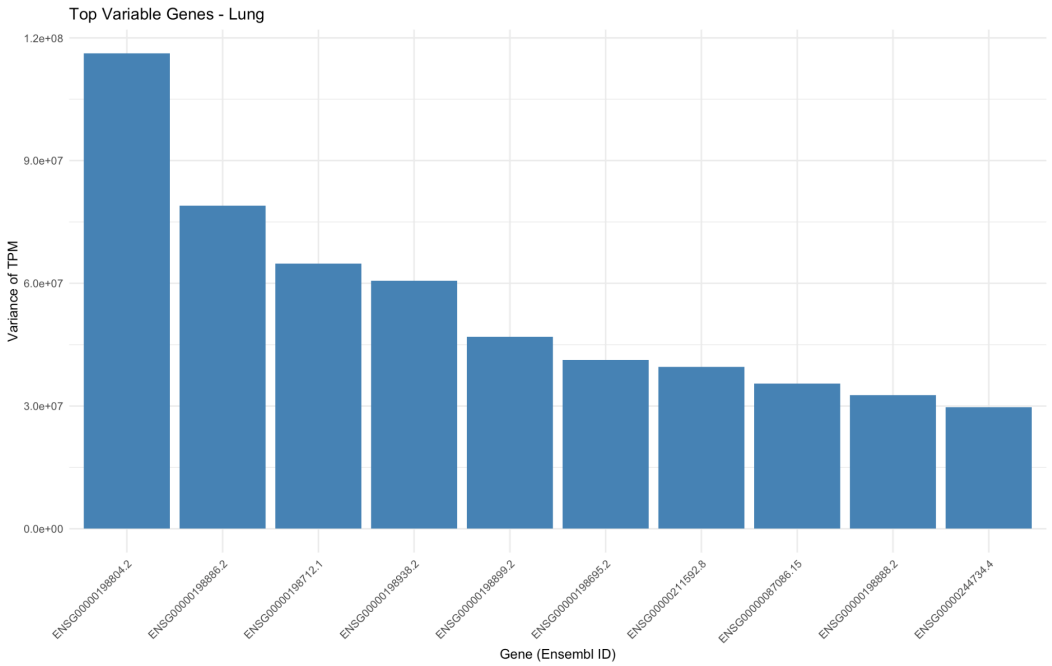

Figure 3. Top ten variable genes by variance of TPM in the GTEx gene TPMs for Lung

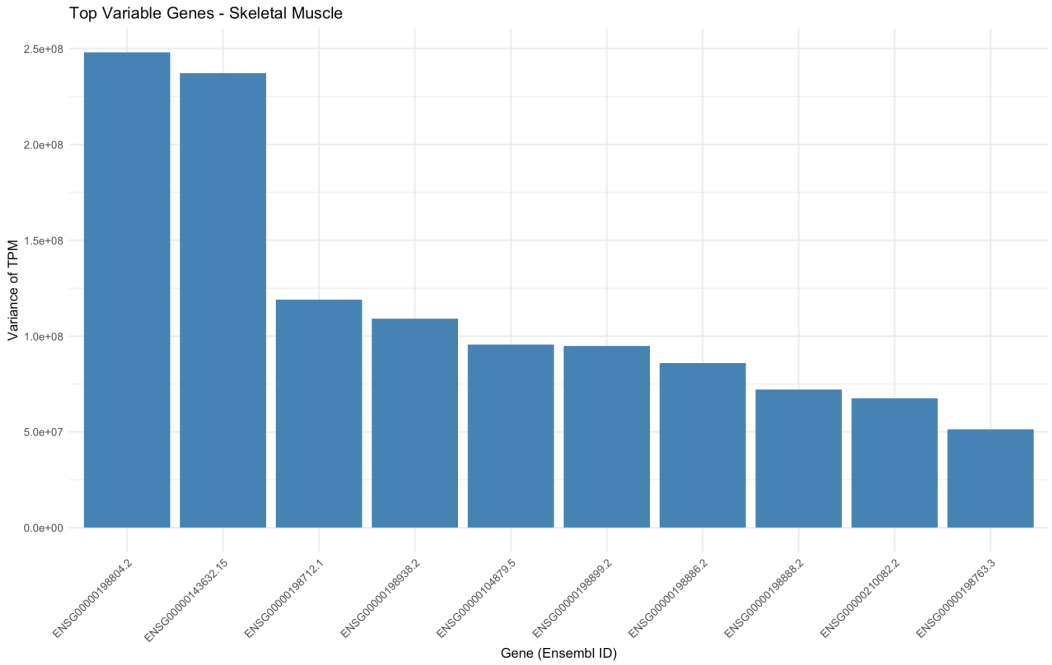

Figure 4. Top ten variable genes by variance of TPM in the GTEx gene TPMs for Skeletal Muscle

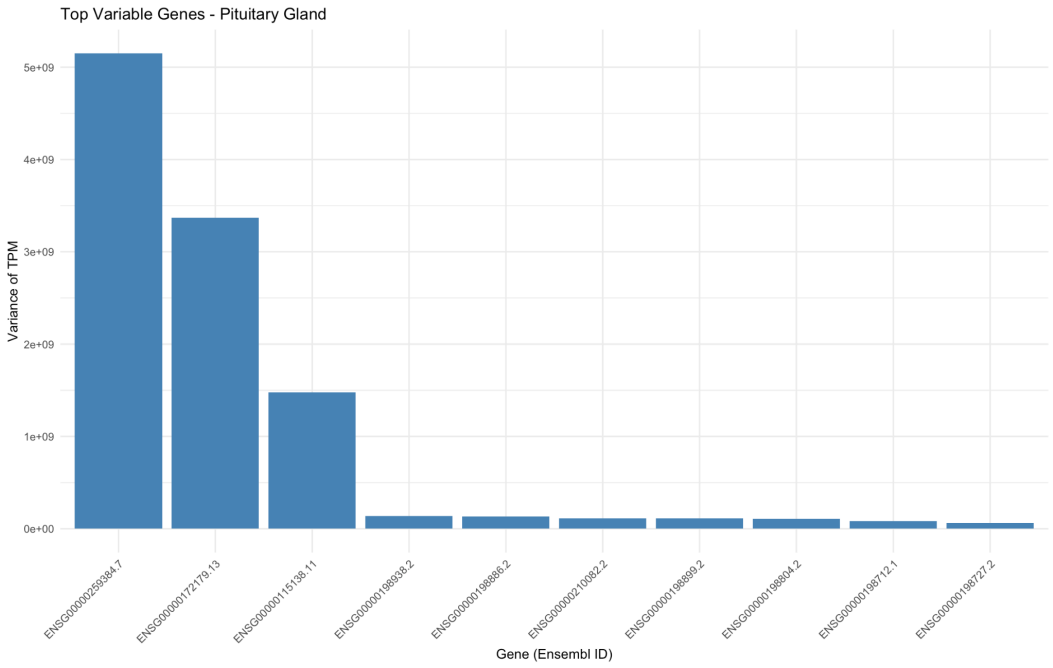

Figure 5. Top ten variable genes by variance of TPM in the GTEx gene TPMs for Pituitary

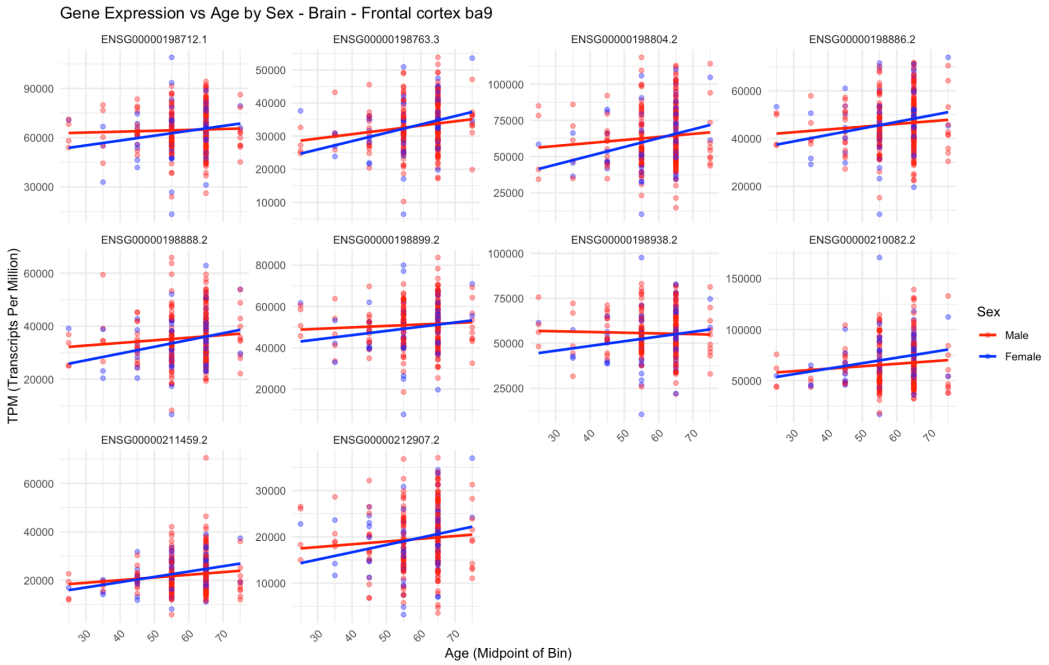

Figure 6. Level of gene expression in TPM based on age from the midpoint of age bin for brain frontal cortex BA9 tissue samples. Includes linear regression line for male and female sexes in each gene.

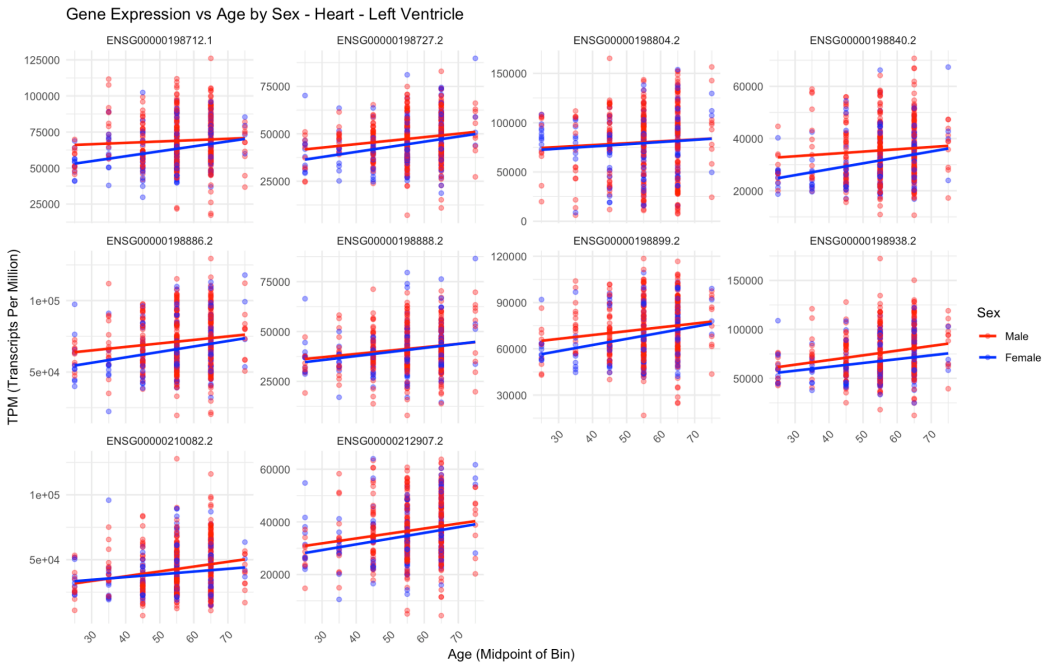

Figure 7. Level of gene expression in TPM based on age from the midpoint of age bin for heart - left ventricle tissue samples. Includes linear regression line for male and female sexes in each gene.

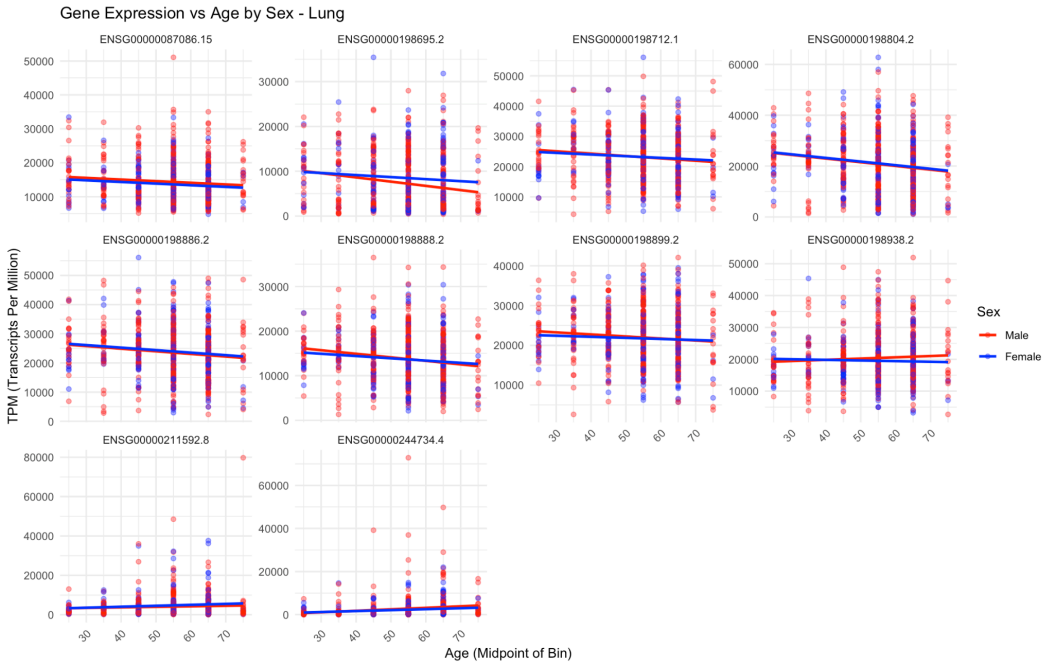

Figure 8. Level of gene expression in TPM based on age from the midpoint of age bin for lung tissue samples. Includes linear regression line for male and female sexes in each gene.

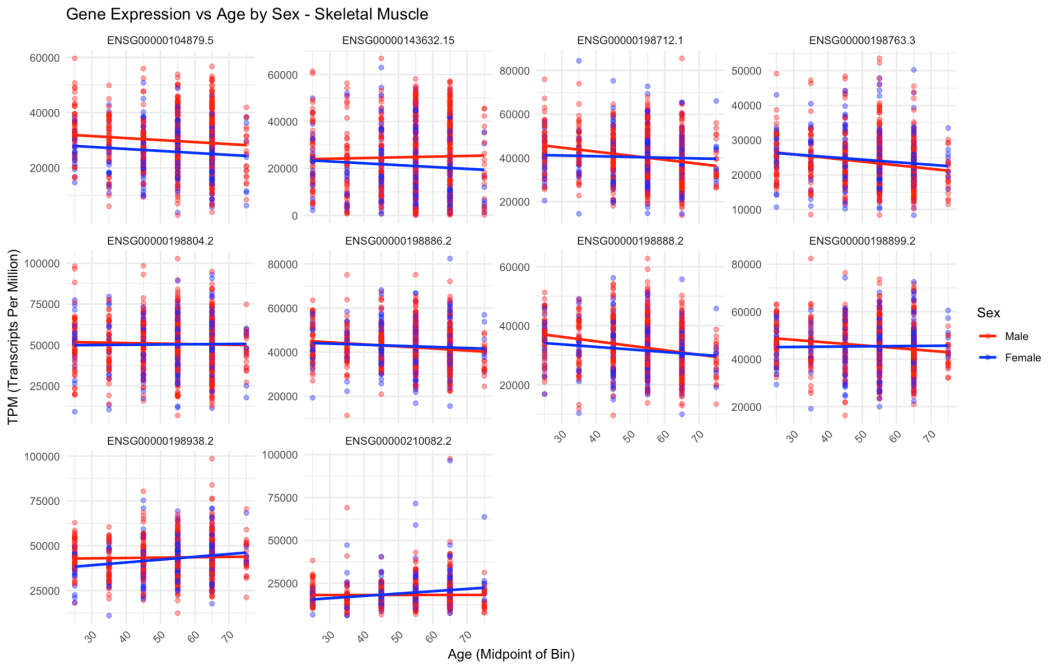

Figure 9. Level of gene expression in TPM based on age from the midpoint of age bin for skeletal muscle tissue samples. Includes linear regression line for male and female sexes in each gene.

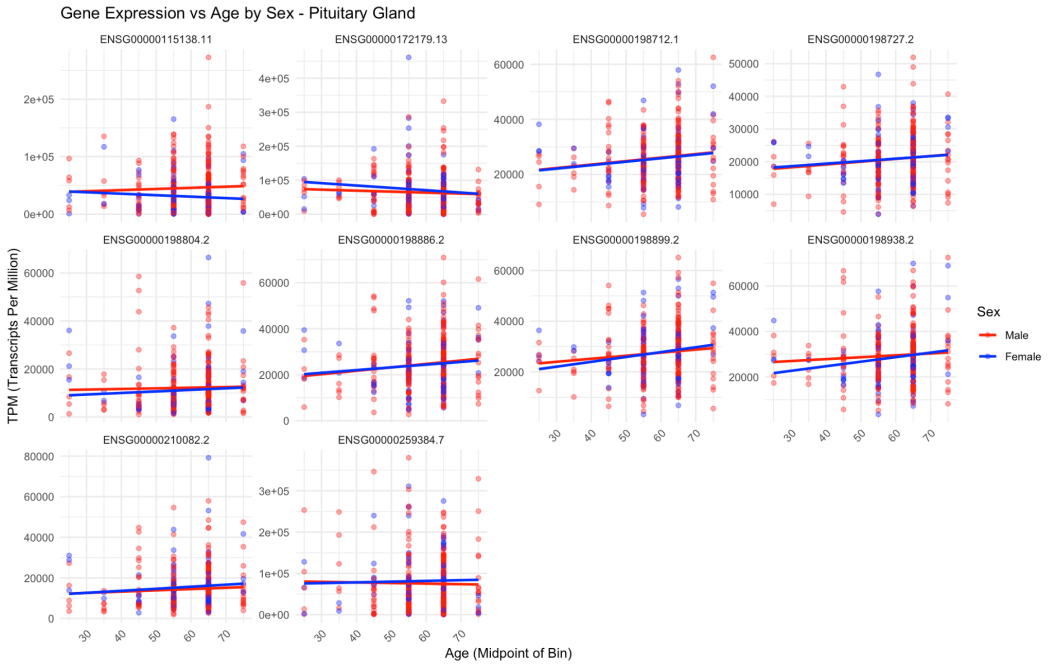

Figure 10. Level of gene expression in TPM based on age from the midpoint of age bin for pituitary gland tissue samples. Includes linear regression line for male and female sexes in each gene.

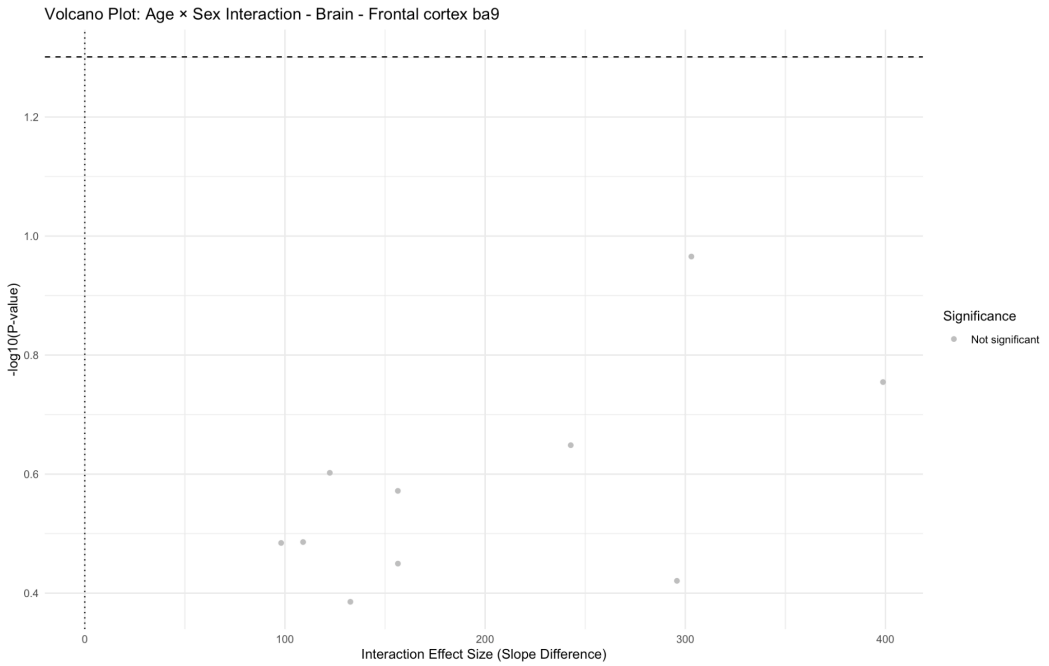

Figure 11. Volcano plot showing the significance of the linear regression slope differences by sex for each of the top ten brain - frontal cortex BA9 genes.

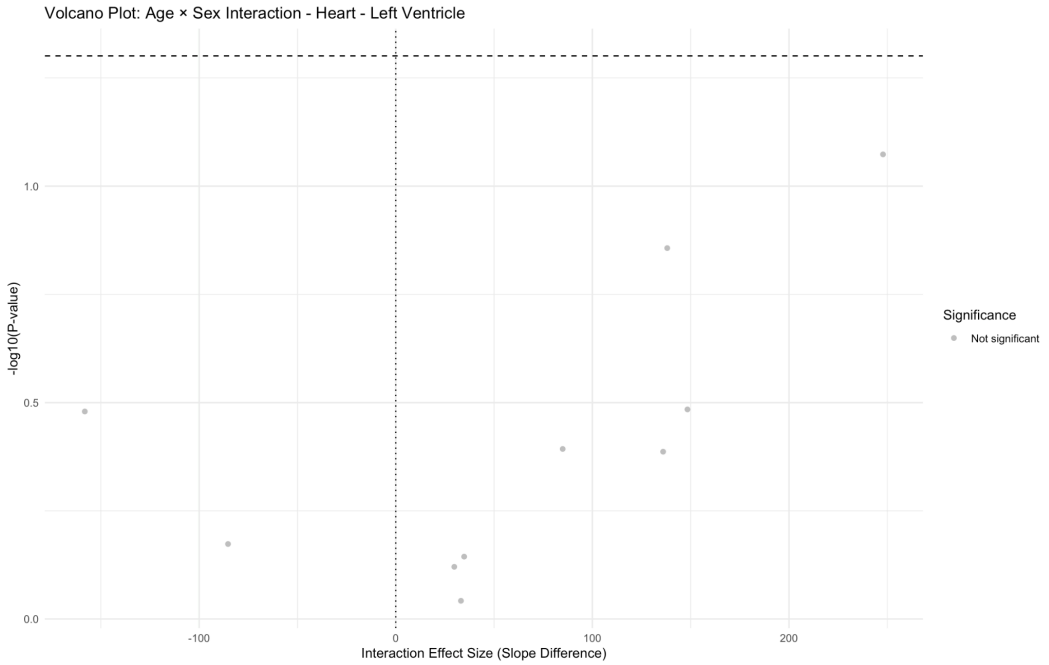

Figure 12. Volcano plot showing the significance of the linear regression slope differences by sex for each of the top ten heart - left ventricle genes.

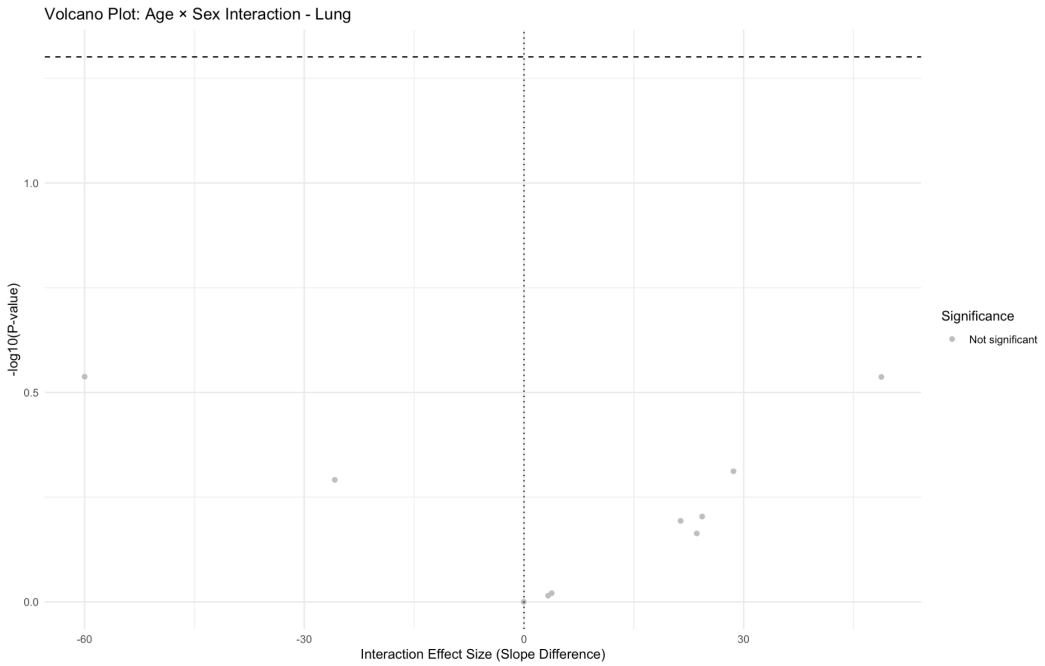

Figure 13. Volcano plot showing the significance of the linear regression slope differences by sex for each of the top ten lung genes.

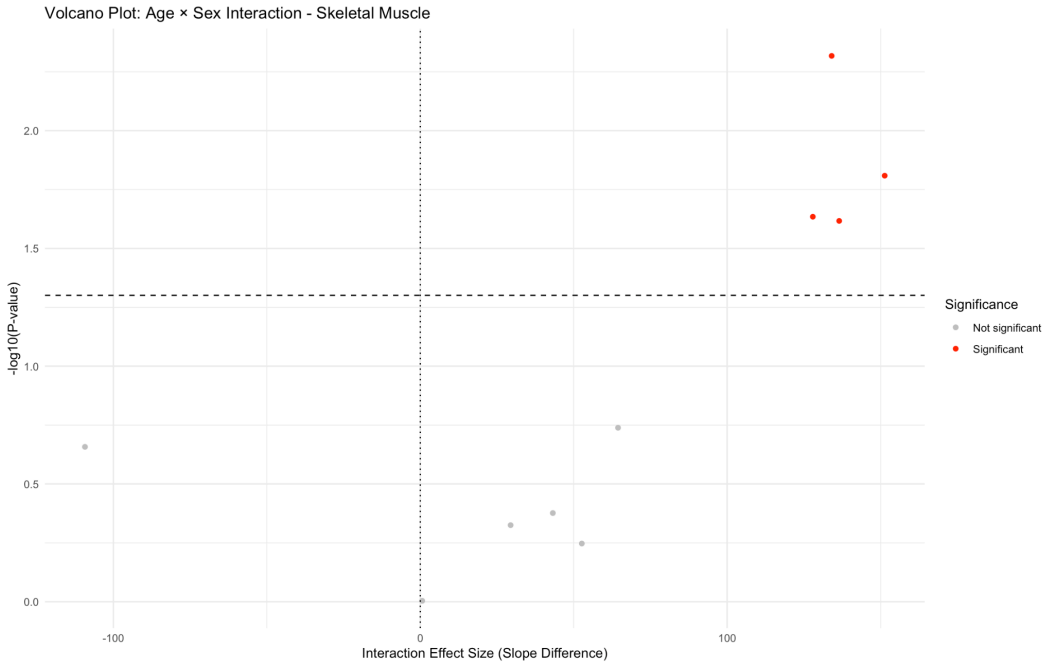

Figure 14. Volcano plot showing the significance of the linear regression slope differences by sex for each of the top ten skeletal muscle genes.

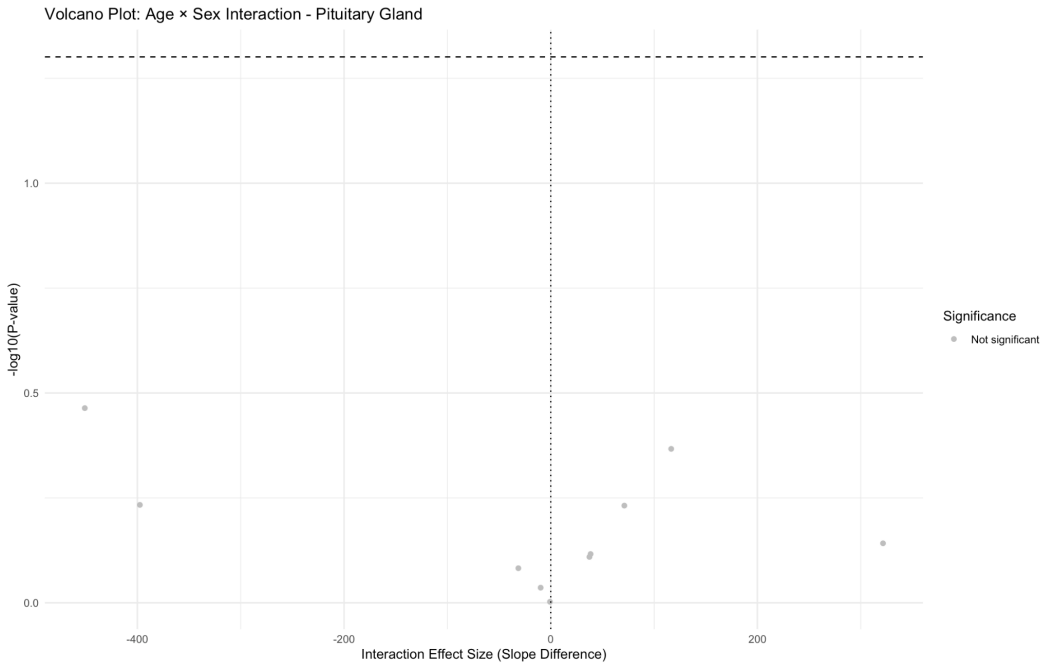

Figure 15. Volcano plot showing the significance of the linear regression slope differences by sex for each of the top ten pituitary gland genes.

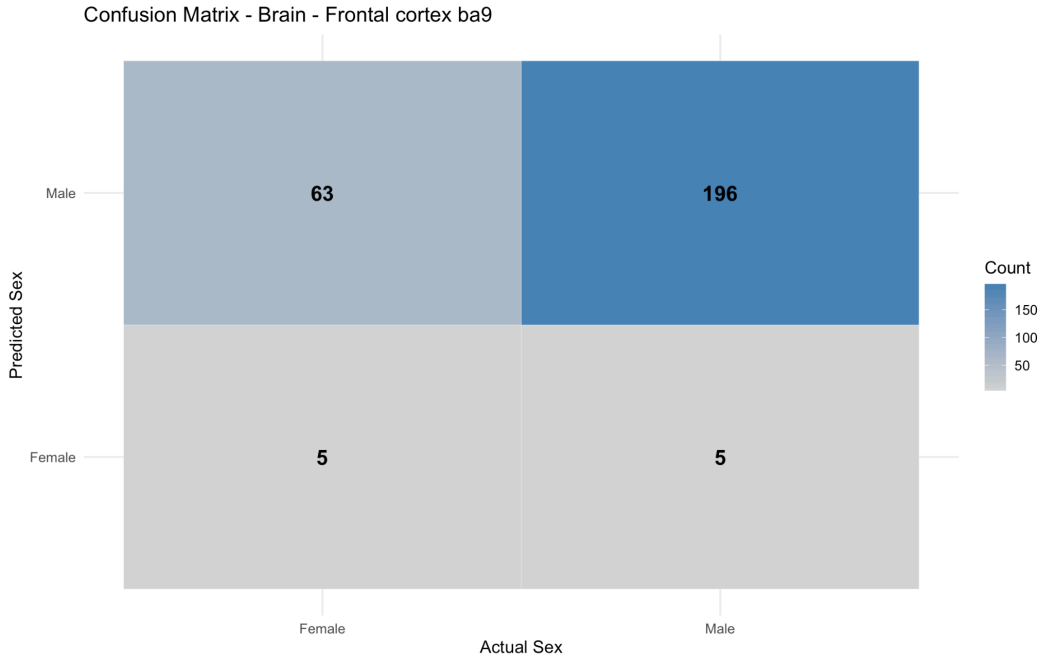

Figure 16. Confusion matrix showing the results of a logistic regression model classifying brain - frontal cortex BA9 tissue samples into male and female.

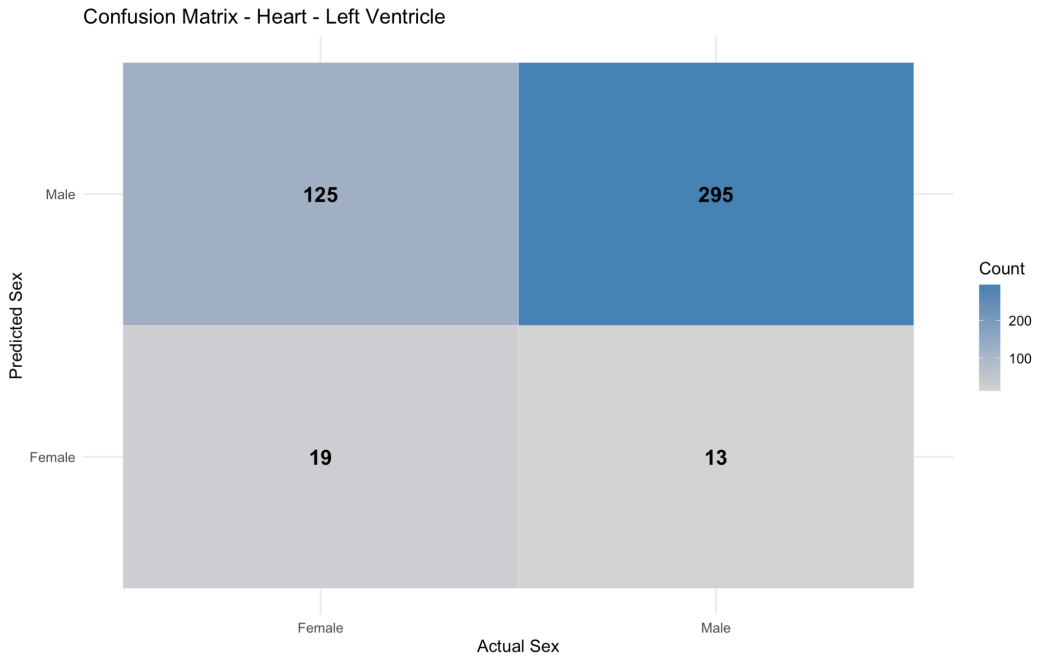

Figure 17. Confusion matrix showing the results of a logistic regression model classifying heart - left ventricle tissue samples into male and female.

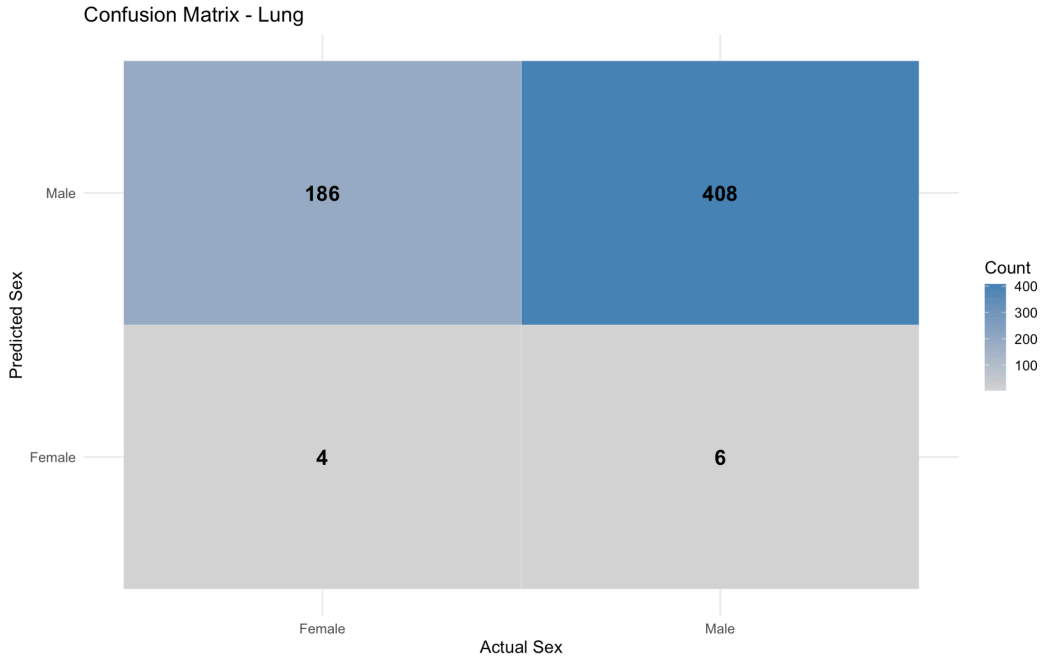

Figure 18. Confusion matrix showing the results of a logistic regression model classifying lung tissue samples into male and female.

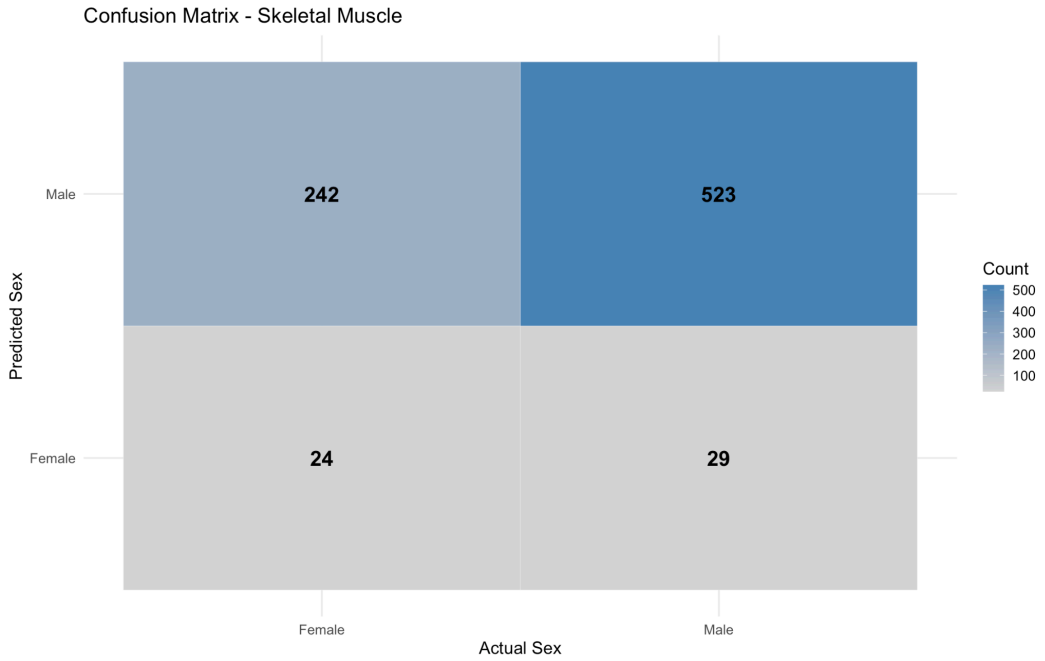

Figure 19.Confusion matrix showing the results of a logistic regression model classifying skeletal muscle tissue samples into male and female.

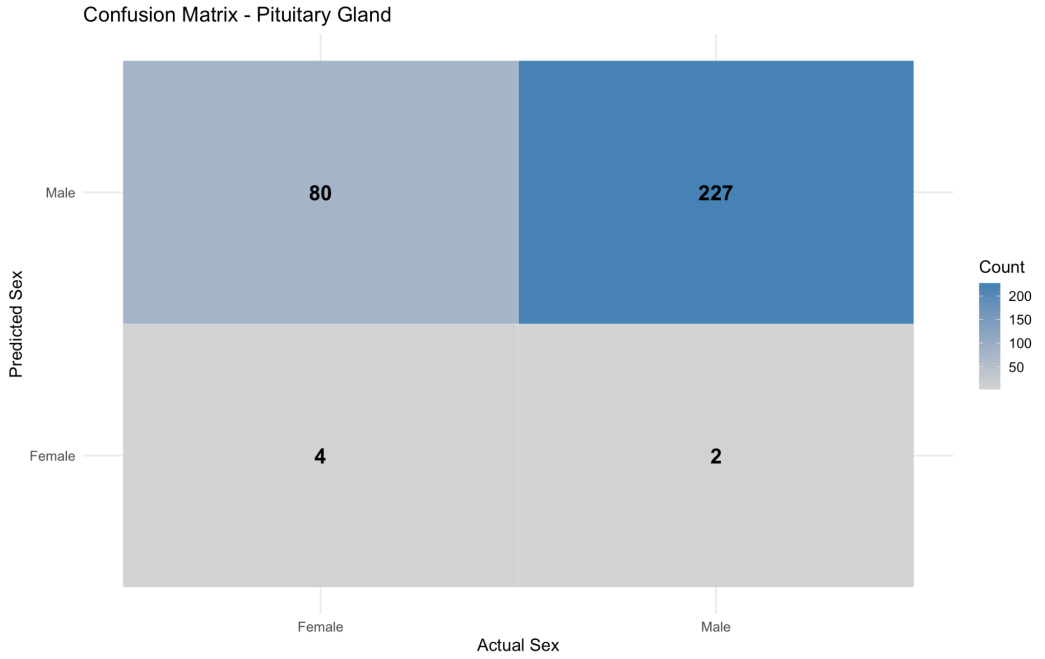

Figure 20. Confusion matrix showing the results of a logistic regression model classifying pituitary gland tissue samples into male and female.

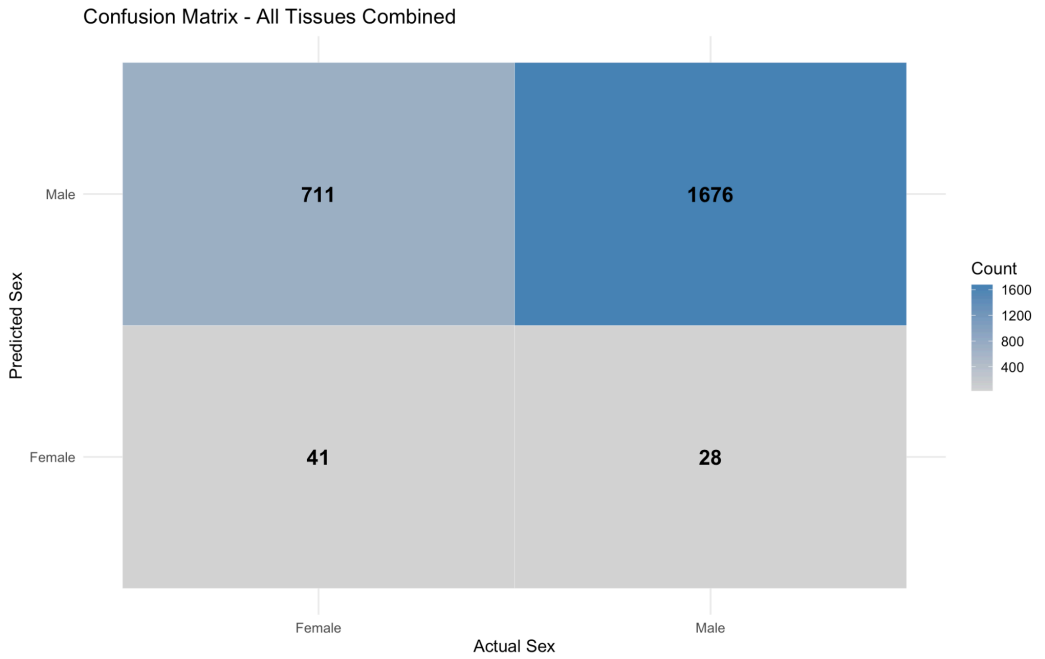

Figure 21. Confusion matrix showing the results of a logistic regression model classifying all tissue samples into male and female.

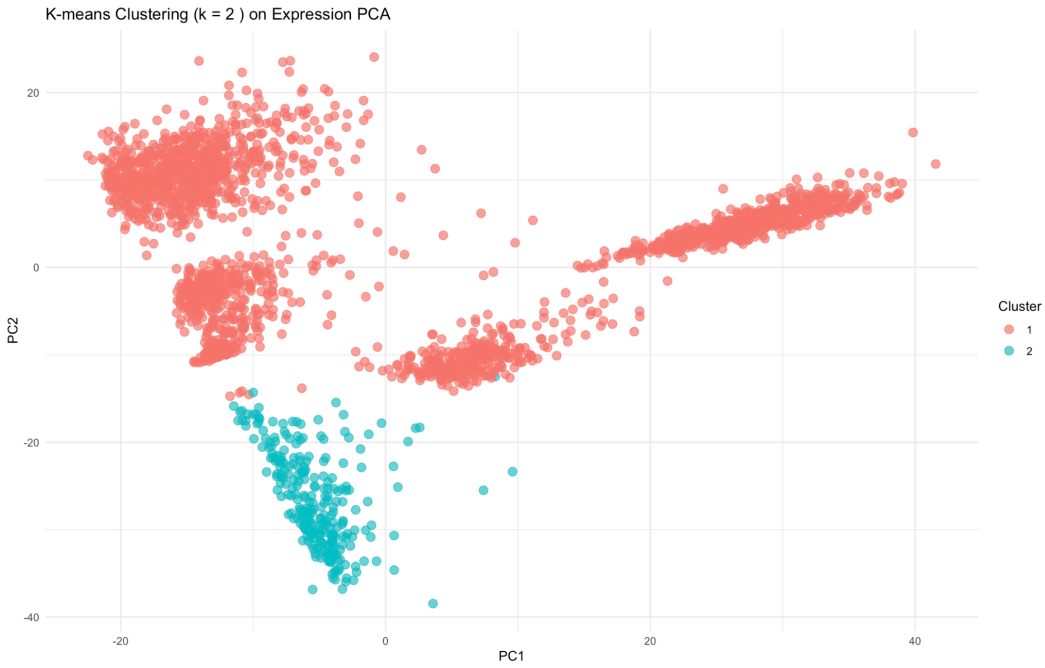

Figure 22. Displays the first two principal components of gene expression from all tissue samples. Each plot is a sample, separated by k-means into two clusters indicated by color.

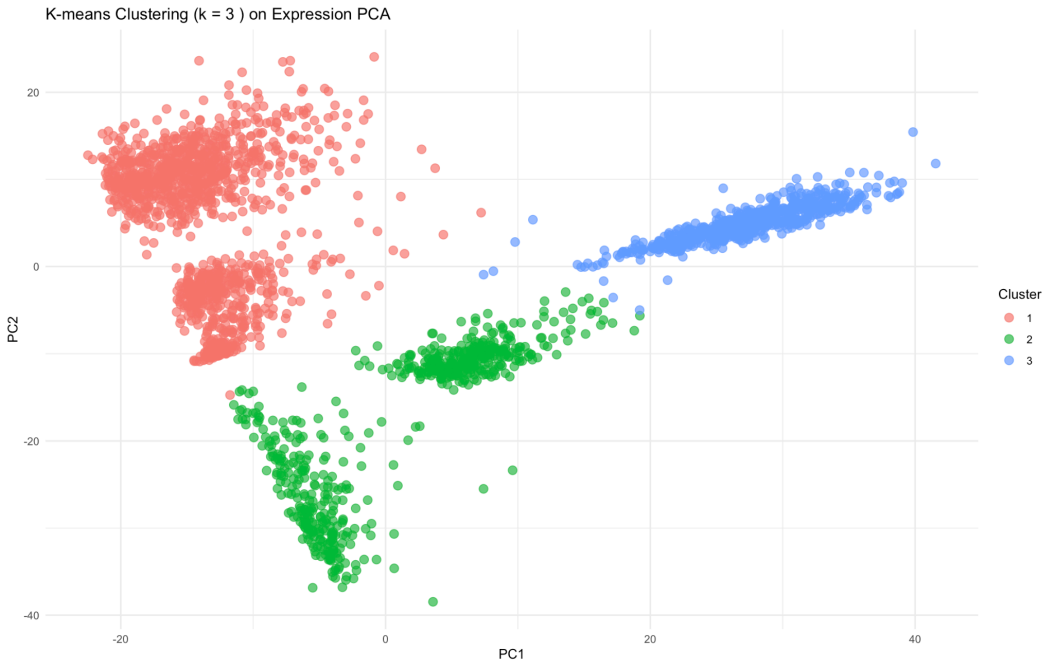

Figure 23. Displays the first two principal components of gene expression from all tissue samples. Each plot is a sample, separated by k-means into three clusters indicated by color.

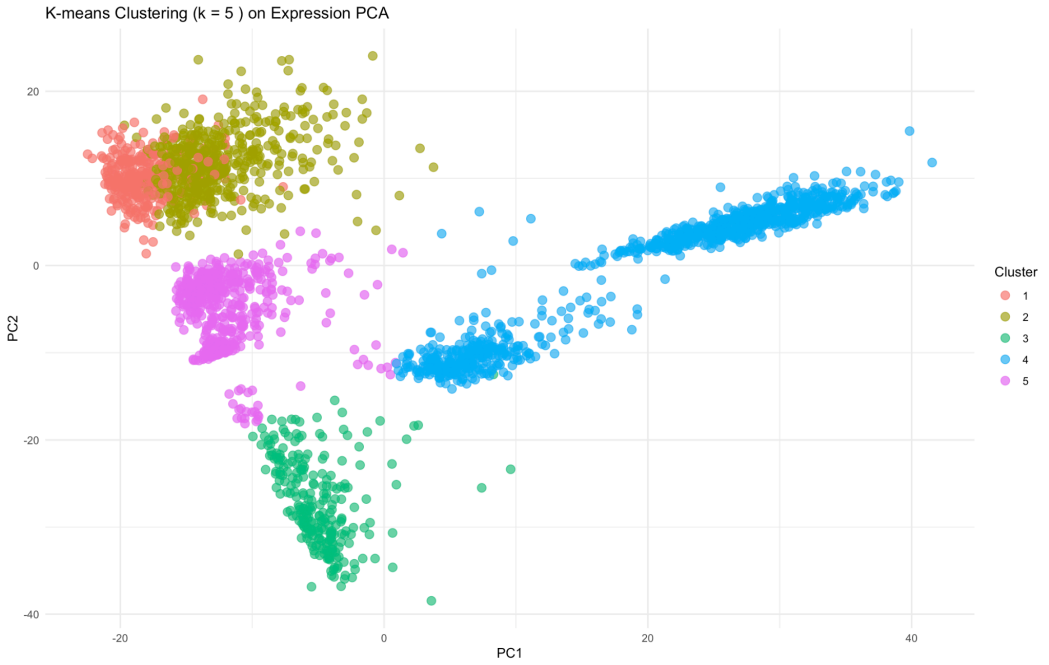

Figure 24. Displays the first two principal components of gene expression from all tissue samples. Each plot is a sample, separated by k-means into five clusters indicated by color.

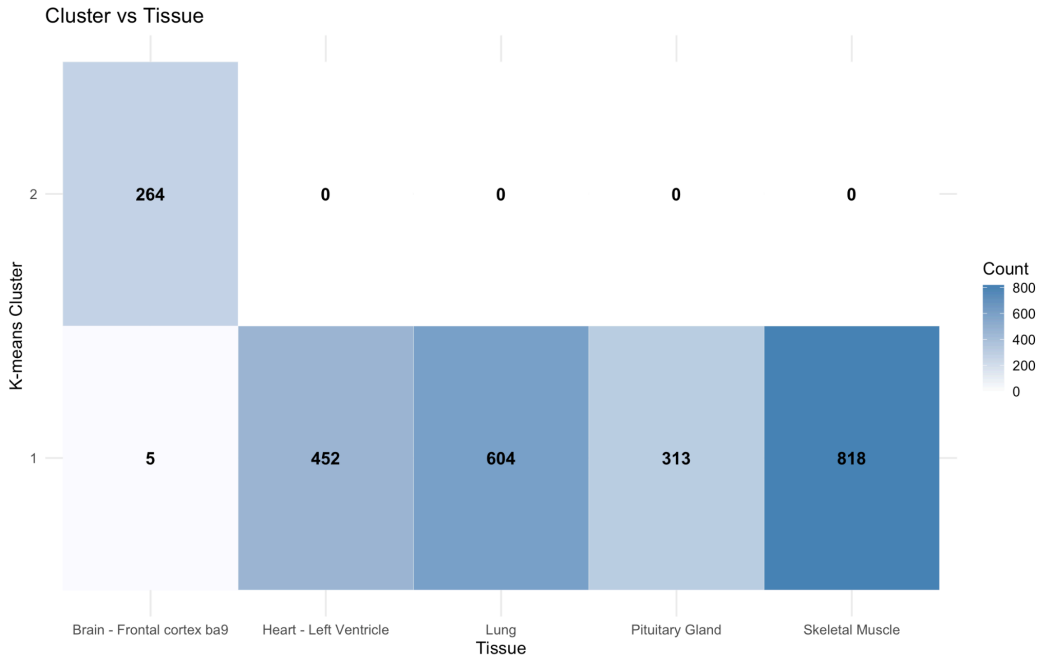

Figure 25. Confusion matrix showing the accuracy of the PCA k-means 2-cluster model at separating the samples into tissue groups.

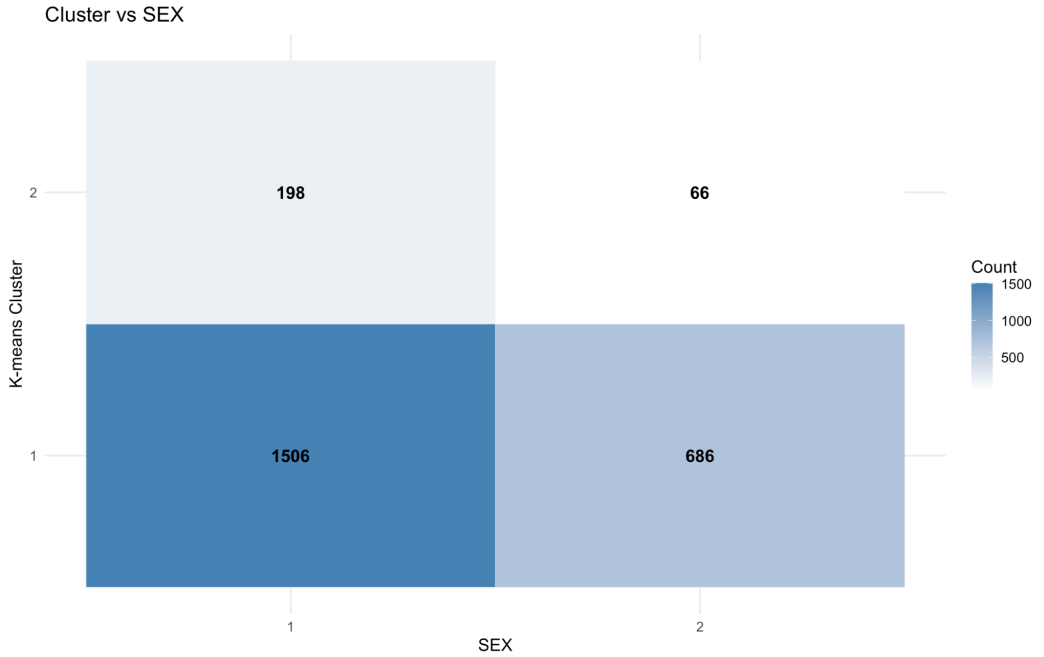

Figure 26. Confusion matrix showing the accuracy of the PCA k-means 2-cluster model at separating the samples into sex groups.

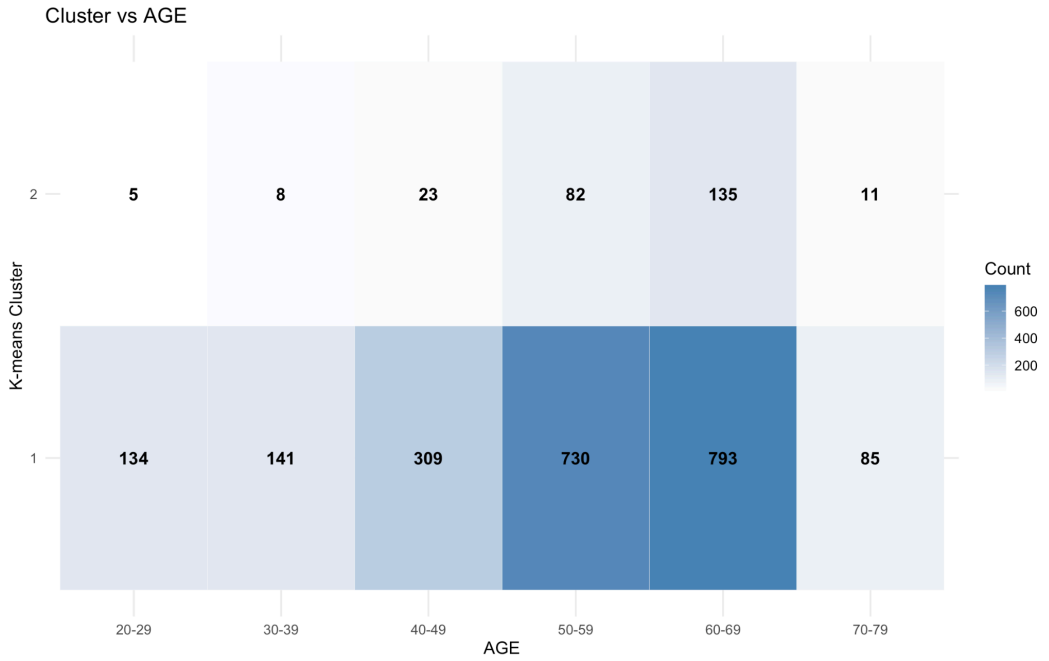

Figure 27. Confusion matrix showing the accuracy of the PCA k-means 2-cluster model at separating the samples into age groups.

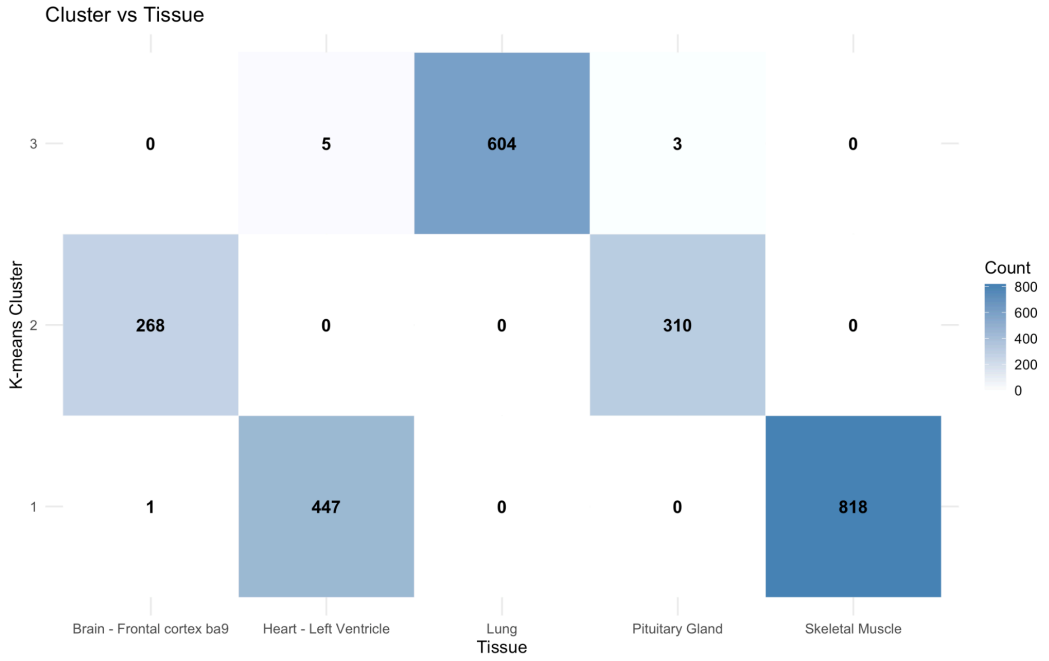

Figure 28. Confusion matrix showing the accuracy of the PCA k-means 3-cluster model at separating the samples into tissue groups.

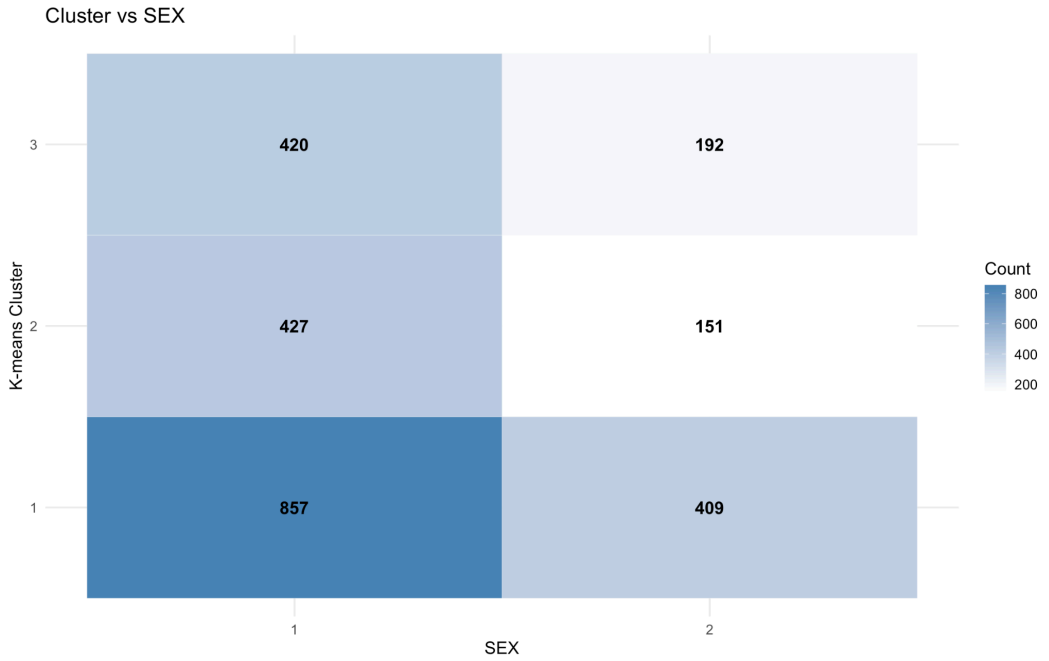

Figure 29. Confusion matrix showing the accuracy of the PCA k-means 3-cluster model at separating the samples into sex groups.

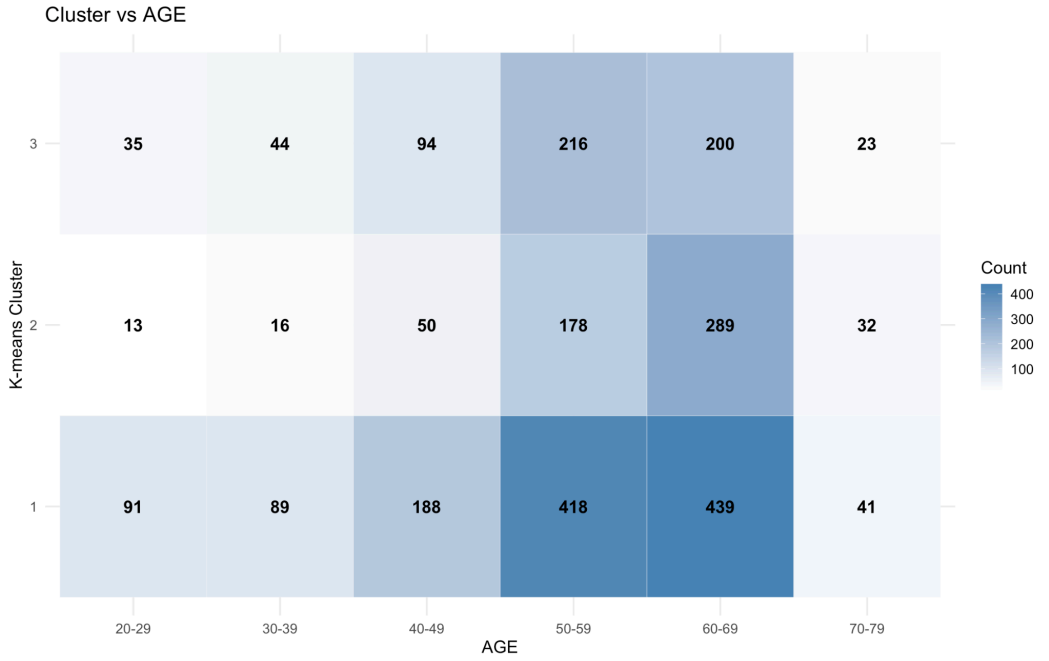

Figure 30. Confusion matrix showing the accuracy of the PCA k-means 3-cluster model at separating the samples into age groups.

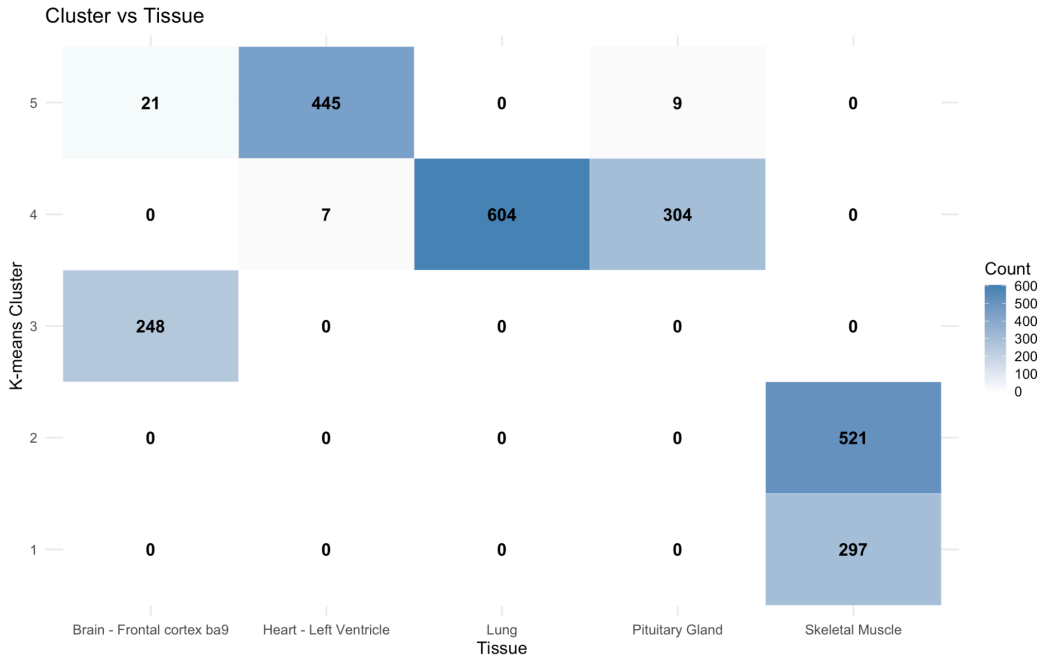

Figure 31. Confusion matrix showing the accuracy of the PCA k-means 5-cluster model at separating the samples into tissue groups.

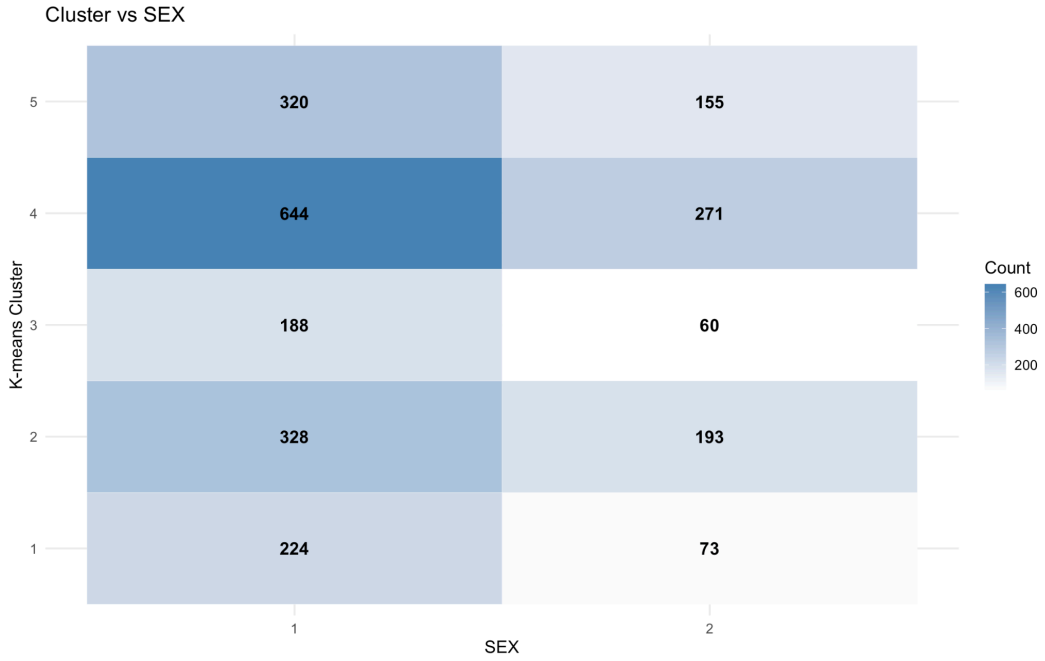

Figure 32. Confusion matrix showing the accuracy of the PCA k-means 5-cluster model at separating the samples into sex groups.

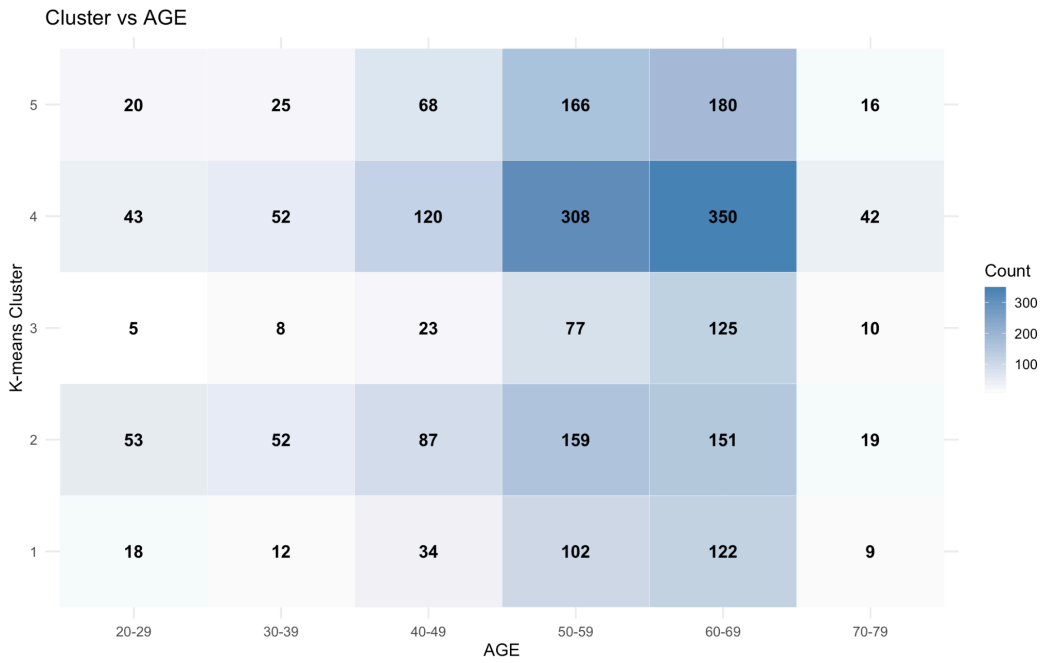

Figure 33. Confusion matrix showing the accuracy of the PCA k-means 5-cluster model at separating the samples into age groups.