BE 700 A1: AI in Systems Biology

“Tell me and I forget, teach me and I may remember, involve me and I learn”

~ Benjamin Franklin

Final Project Report

05/10/2022

Submitted to:

Professor Simon Kasif

Team Members:

Giulia Boccardo

Nikita Tomar

Shrishtee Kandoi

All

Abstract

Jin et al., (2011) focuses on insulin resistance in skeletal muscle, a key phenotype associated with Type 2 Diabetes (T2D) whose molecular mechanisms are still unknown. According to the authors, "Gene expression analysis can be used to identify signatures or early defects related to different phenotypes. Alterations in insulin signaling, mitochondrial dysfunction, endoplasmic reticulum stress, oxidative stress, and inflammation have all been implicated in the induction of cellular insulin resistance" [1]. Lu et al., (2014) examines postmortem neuropathological normal brain samples from the frontal cortical regions of young, middle-aged, and elderly patients in order to gain a better understanding of the molecular processes of aging in the frontal cortical area of the brain [2]. In this project, we aim to examine and integrate datasets from both studies in order to gain new biological insights and information about systems biology of the diseases.

Background

As previously stated, the first research paper focuses on insulin resistance in skeletal muscle; a key phenotype associated with type 2 diabetes for which the molecular mechanisms are still unclear. The second paper focuses on better understanding molecular processes of aging in the frontal cortical area of the brain. To gain new biological insights, heatmaps, PCA analysis, UMap analysis, gene-to-gene and gene-to-phenotype correlations, box plots, Pearson and Spearman correlation analysis, predictive neural networks (top 10 age correlated gene expression as inputs; age group as output), decision trees, David analysis, and GEO2R were conducted.

Methods

Dataset extraction

Gene expression data from the two datasets from the previously mentioned papers was retrieved. The human and phenotype from the first paper (Insulin Resistance and Diabetes in Muscle[GSE25462]) was loaded into R from Blackboard. The data from the second paper (Brain Aging [GDS5204]) was retrieved from the GEO website and the expression matrix was extracted. The phenotype data for the second paper was also extracted.

Heatmaps of all human data

To generate the heatmap of all human expression data, all genes and all samples, the ComplexHeatmap Library was used. In gene expression analysis, heatmaps are utilized for data visualization purposes. In heatmaps, the data is displayed in a grid where each row represents a gene and each column represents a sample. The color and intensity of the boxes is used to represent changes (not absolute values) of gene expression. The heatmap corresponding to paper 1 is displayed in Figure 1 and the heatmap corresponding to paper 2 is displayed in Figure 2.

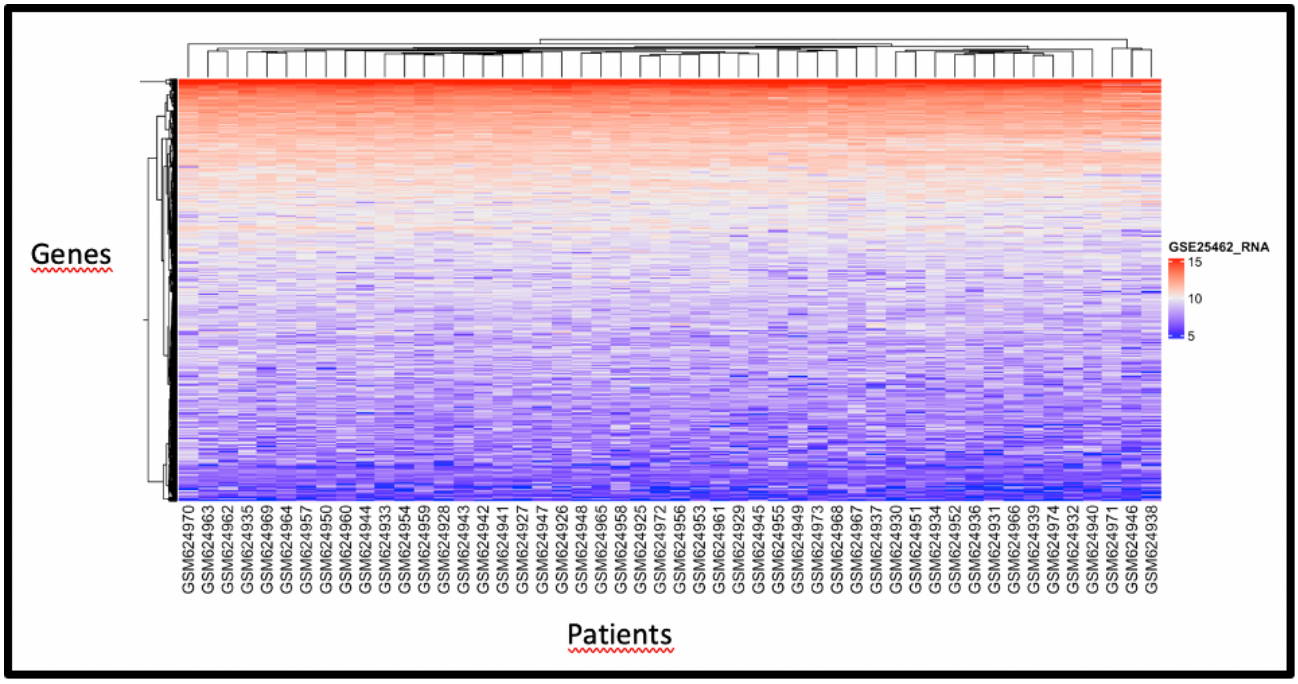


Figure 1: Heatmap of all Human Data for Insulin Resistance and Diabetes in Muscle

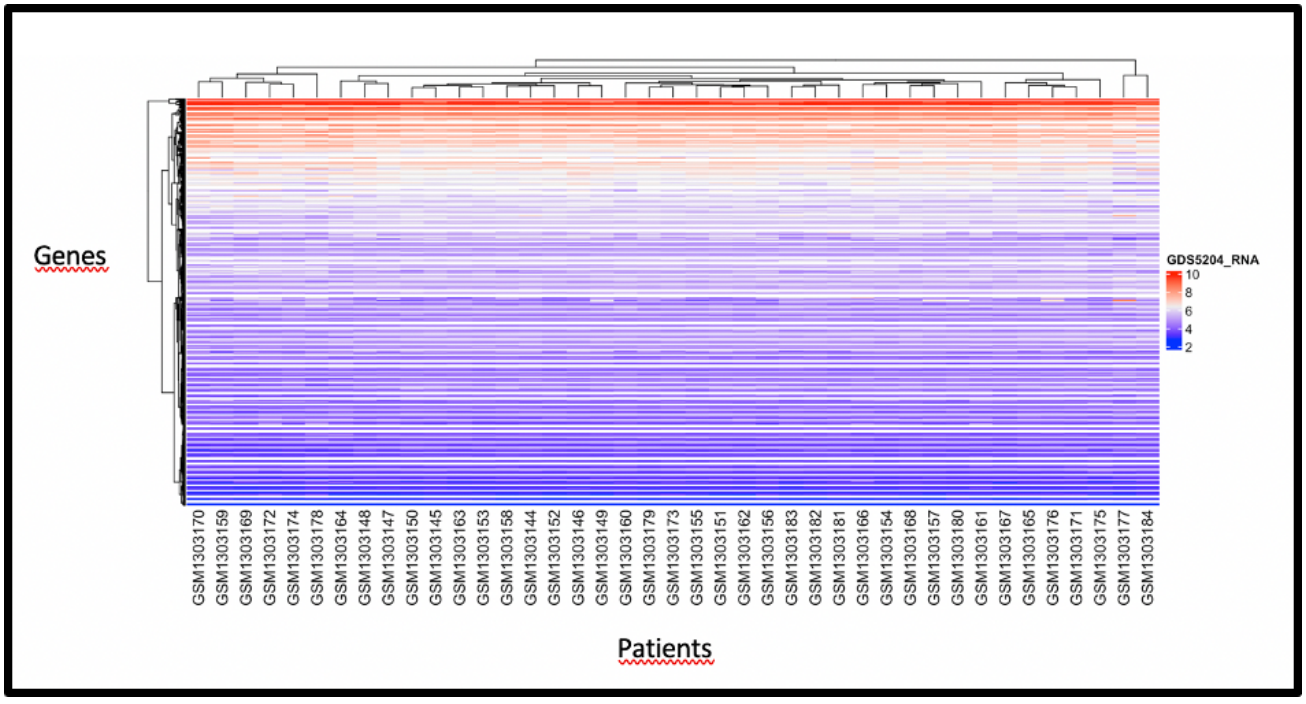
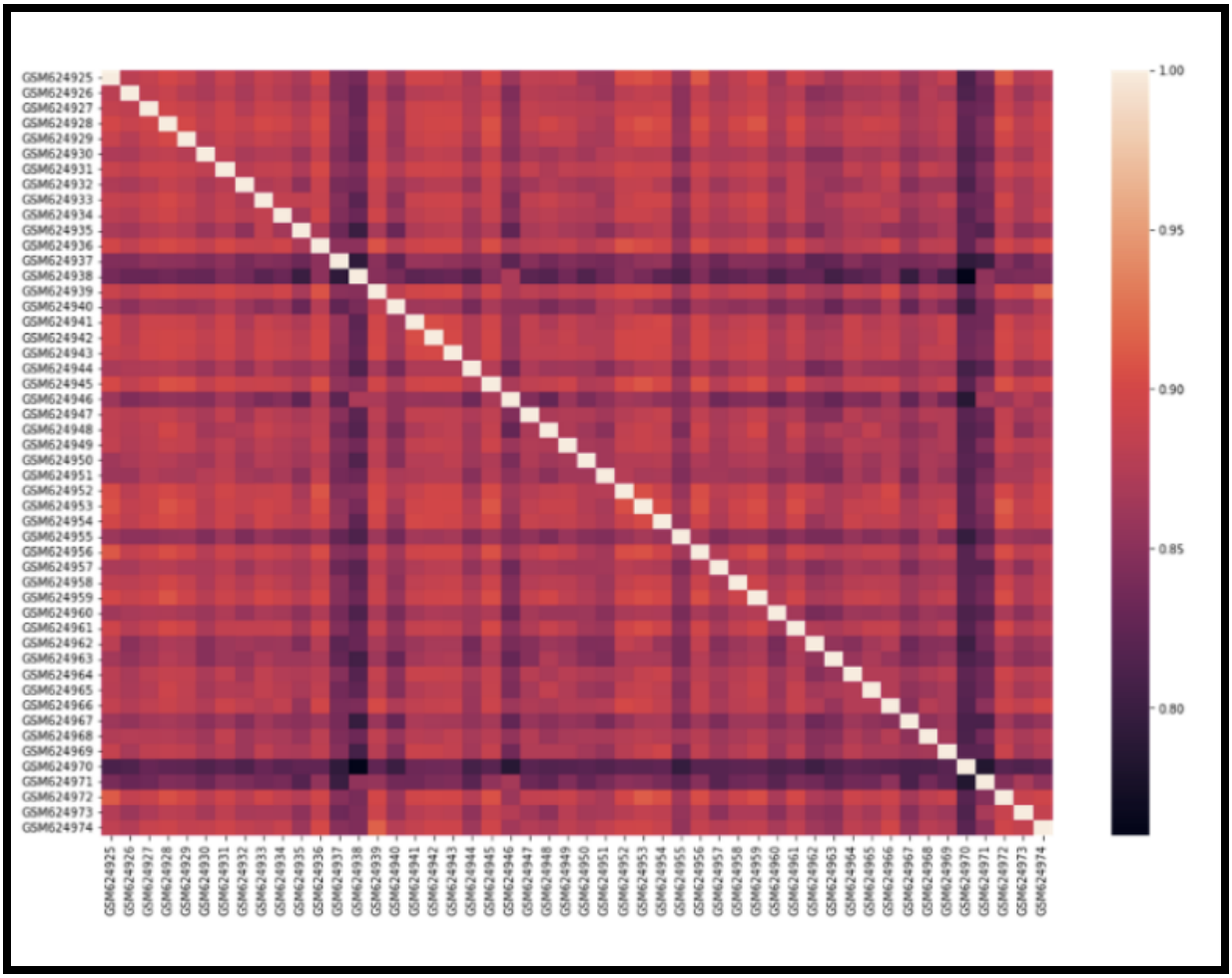


Figure 2: Heatmap of all Human Data for Brain Aging

Gene-to-gene correlations using all data

Gene-to-gene correlations for all data for both papers was conducted. The correlations were obtained using the corr() function in python and plotted using the sns.heatmap function from the sns library. A genetic correlation in multivariate quantitative genetics is the proportion of variance that two traits share due to genetic causes, or the correlation between genetic influences on one trait and genetic influences on another trait, estimating the degree of pleiotropy or causal overlap. The correlation plot in Figure 3 shows the correlations between genes from the insulin resistance data frame whereas the plot in Figure 4 shows the correlations between genes from the Brain Aging data frame .

The corr() method was used to calculate the relationship and find the pairwise correlation between each column in our data set. Any "NA" values were automatically excluded. Any non-numeric data type columns in the dataframe were ignored. Note: The correlation of a variable with itself is 1 which is why we can see a diagonal with the same colors.


GSM624925
GSM624926
GSM624927
GSM624928
GSM624929
GSM624930
GSM624931
GSM624932
GSM624933
GSM624934
GSM624935
GSM624936
GSM624937
GSM624938
GSM624939
GSM624940
GSM624941
GSM624942
GSM624943
GSM624944
GSM624945
GSM624946
GSM624947
GSM624948
GSM624949
GSM624950
GSM624951
GSM624952
GSM624953
GSM624954
GSM624955
GSM624956
GSM624957
GSM624958
GSM624959
GSM624960
GSM624961
GSM624962
GSM624963
GSM624964
GSM624965
GSM624966
GSM624967
GSM624968
GSM624969
GSM624970
GSM624971
GSM624972
GSM624973
GSM624974
1.00
0.95
0.90
0.85
0.80

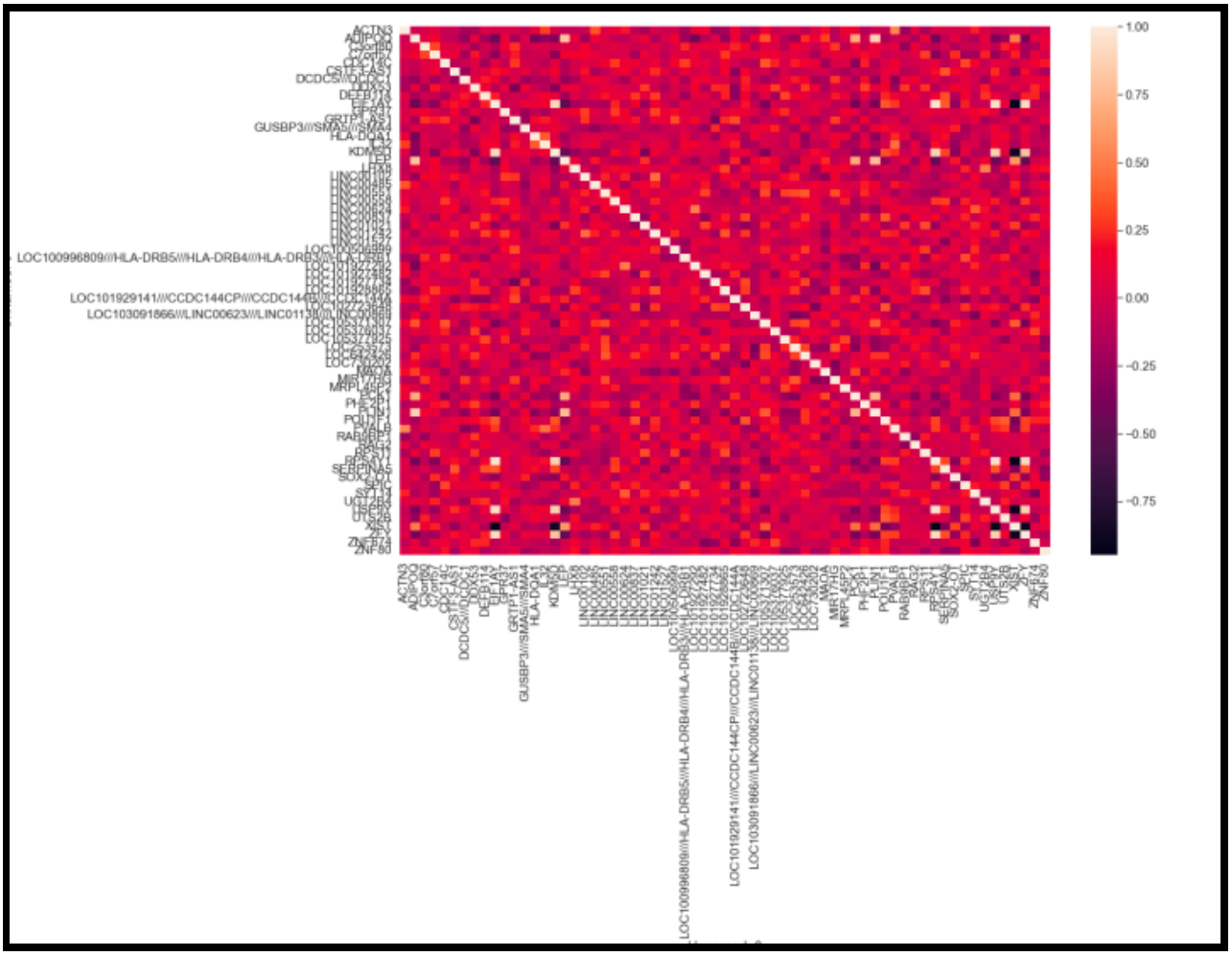
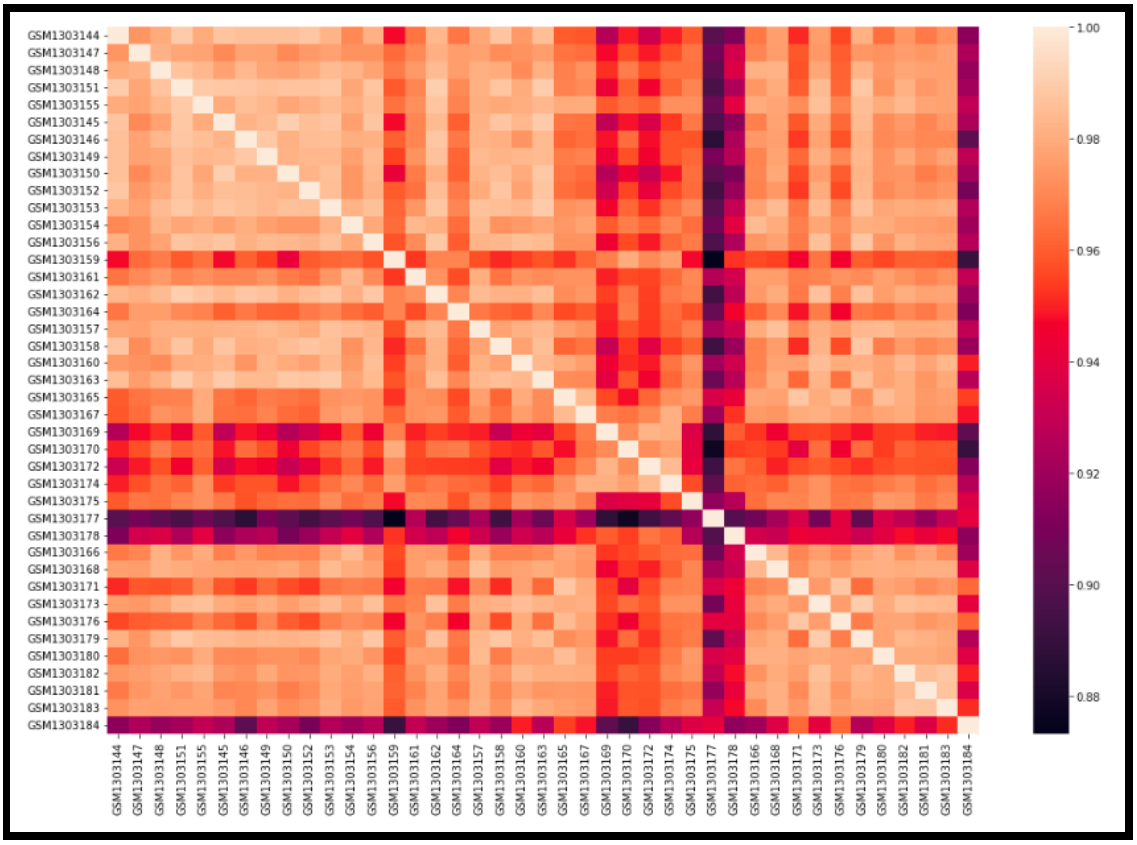


Figure 3: Correlation Analysis on Insulin resistance and Diabetes data



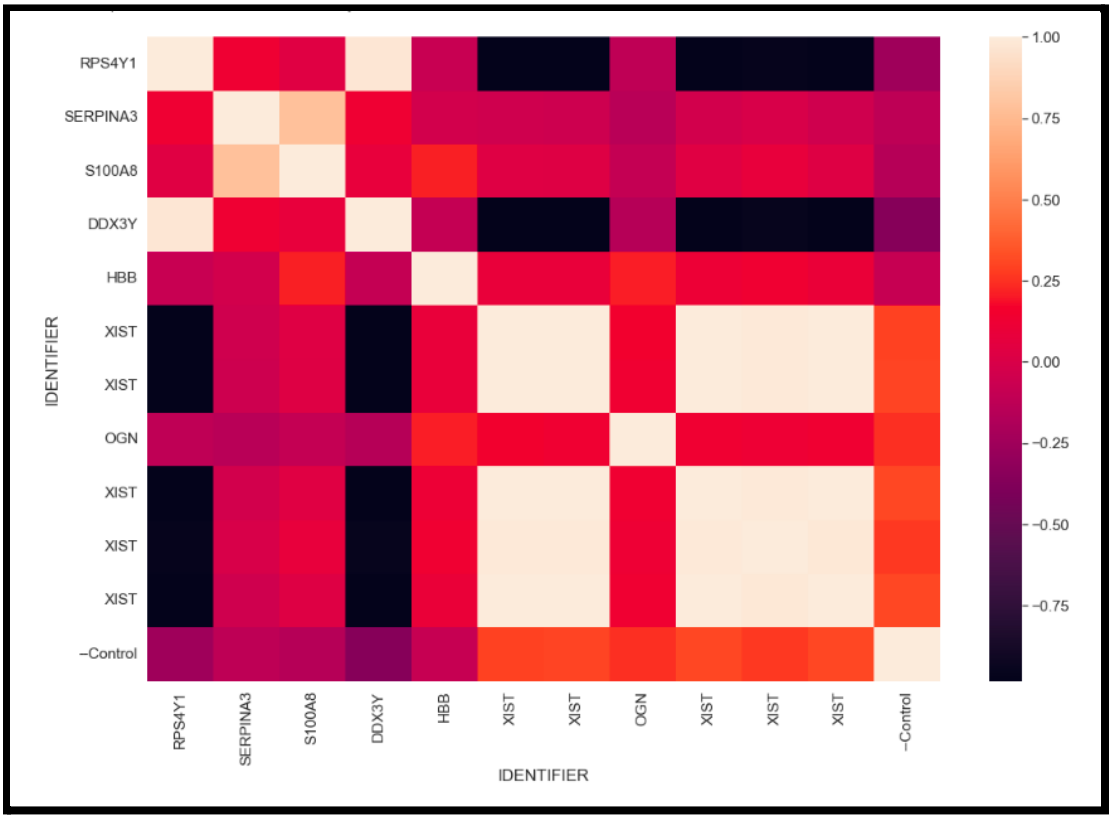


Figure 4: Correlation Analysis on Brain data

K-Means Clustering

The extracted data was then clustered by gene expression and by patients by implementing the k-means clustering algorithm. Clustering allows us to identify which observations are alike, and potentially categorize them together. K-means clustering is the simplest and most commonly used clustering method for splitting a dataset into a set of k groups. It is an iterative algorithm that aims to partition the dataset into K pre-defined individual non-overlapping clusters where each data point belongs to only one group.

Specifically, the k-means algorithm chooses k cluster centers to coincide with k randomly-chosen patterns or k randomly defined points inside the hypervolume containing the pattern set. Each pattern is subsequently assigned to the closest cluster center. Next, the cluster centers are recomputed using the new cluster memberships. If a convergence criterion is not met (e.g. no/minimal pattern reassignment to new clusters or minimal decrease in squared error), steps 2-4 (the last two steps) are repeated again.

The maximum number of iterations and number of clusters was specified. The number of clusters was determined by utilizing the elbow curve displayed in Figures 5 and 6. The elbow curve is utilized to define the optimal number of clusters so that the total intra-cluster variation (known as total within-cluster variation) is minimized. Optimally, one observes a sharp drop at a certain cluster and the curve then plateaus, however, for this particular clustering, that trend was not observed, as the graph continues to go down (gradual decrease) as the number of clusters increase. It is important to note that the instructions specified one could use two or more clusters.

The data frame was transposed, rows and columns were switched, as clustering was conducted by patients, not by genes. The different clusters were then plotted. Samples in cluster 1 in red, cluster 2 in yellow, cluster 3 in blue. This is displayed in Figure 8 and 9. K-means clustering conducted by gene expression is displayed in Figure 7.

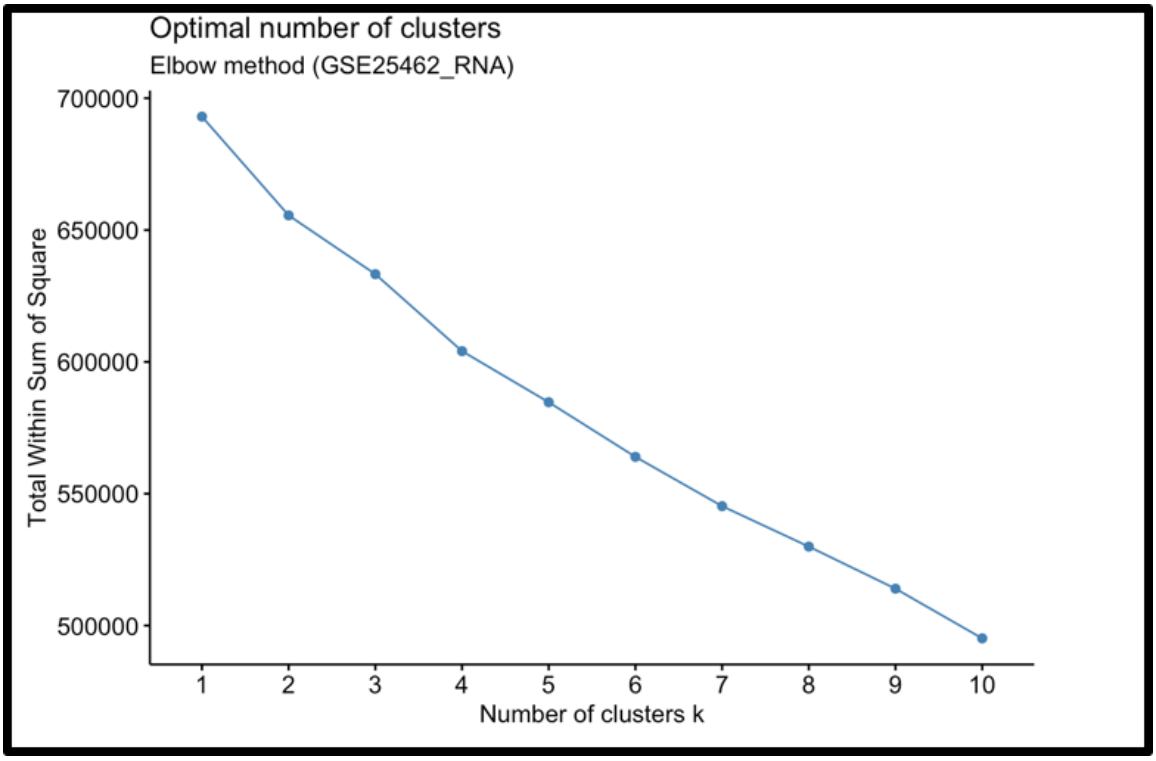


Figure 5: Elbow Curve for Insulin Resistance and Diabetes in Muscle Dataframe

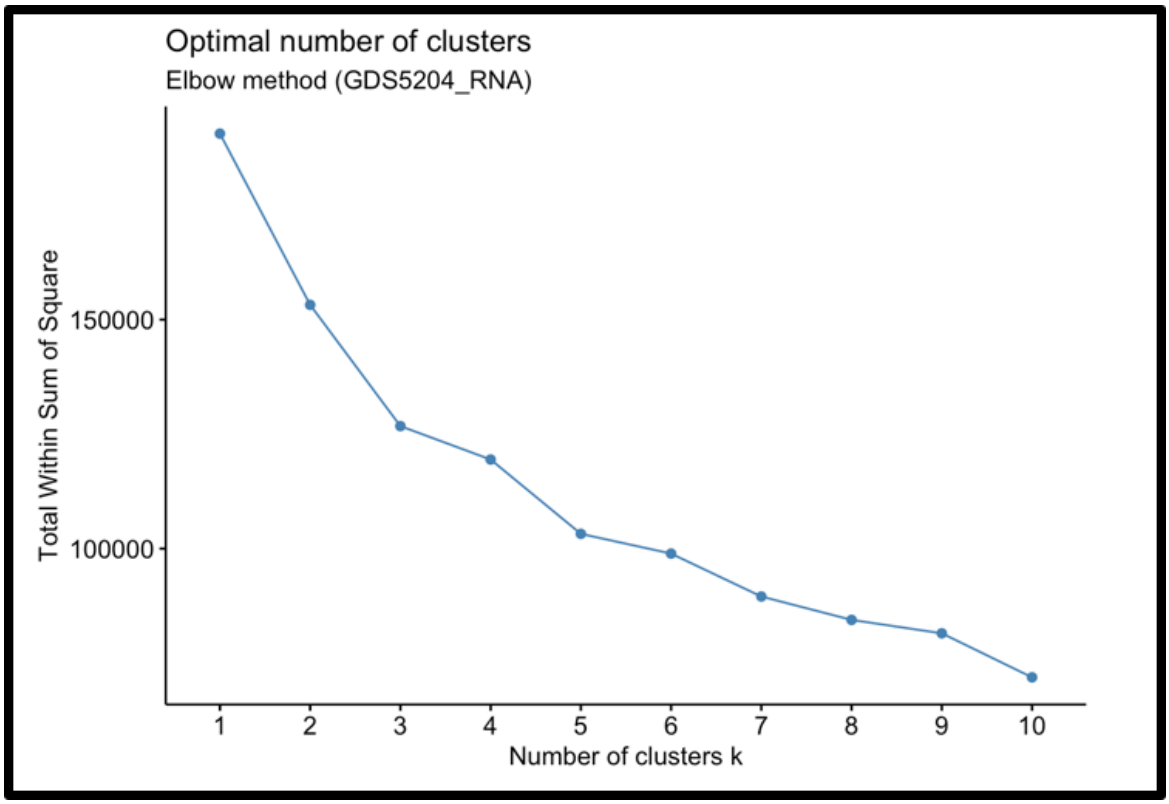


Figure 6: Elbow Curve for Brain Aging Dataframe

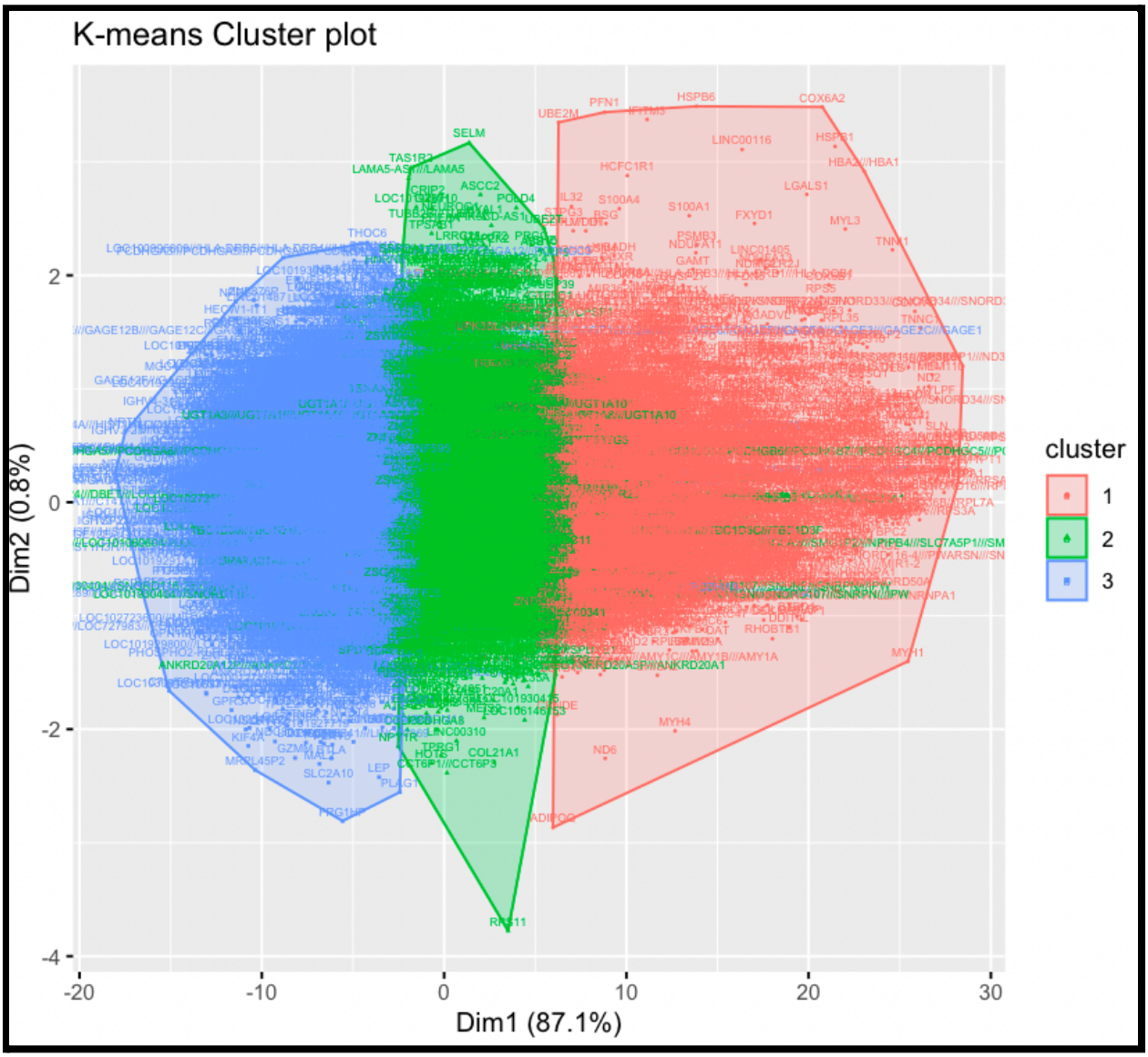


Figure 7: K-means Clustering with k=3 for Human Data for Insulin resistance and Diabetes in Muscle dataset

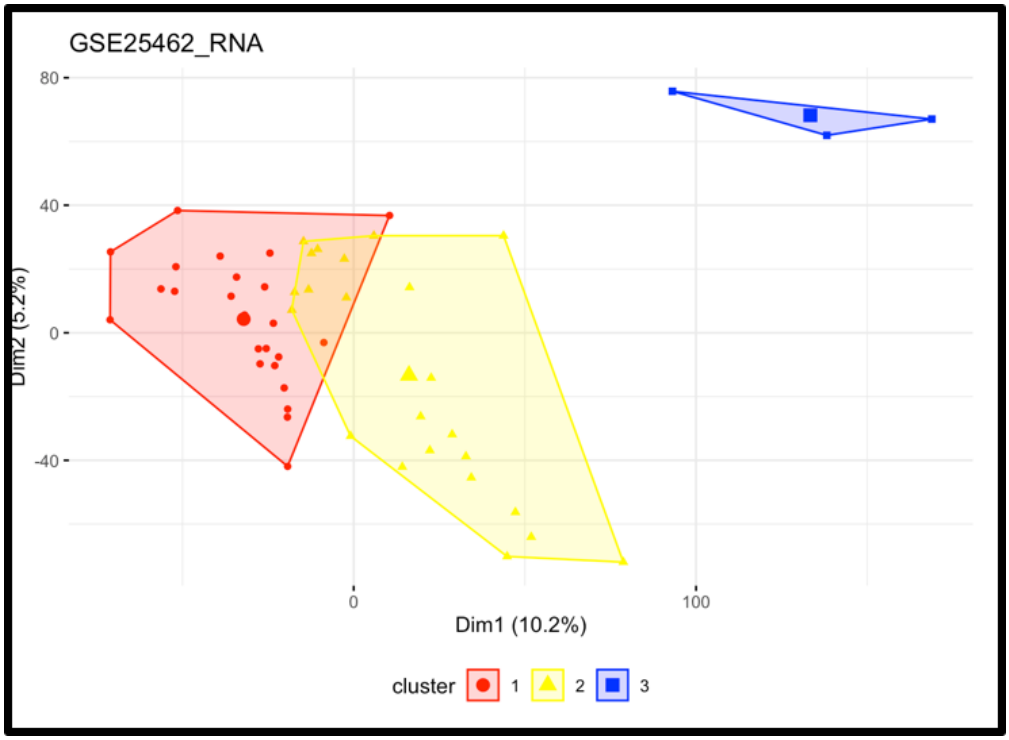


Figure 8: K-Means Clustering with K=3 for Human Data for Insulin Resistance and Diabetes in Muscle Dataframe

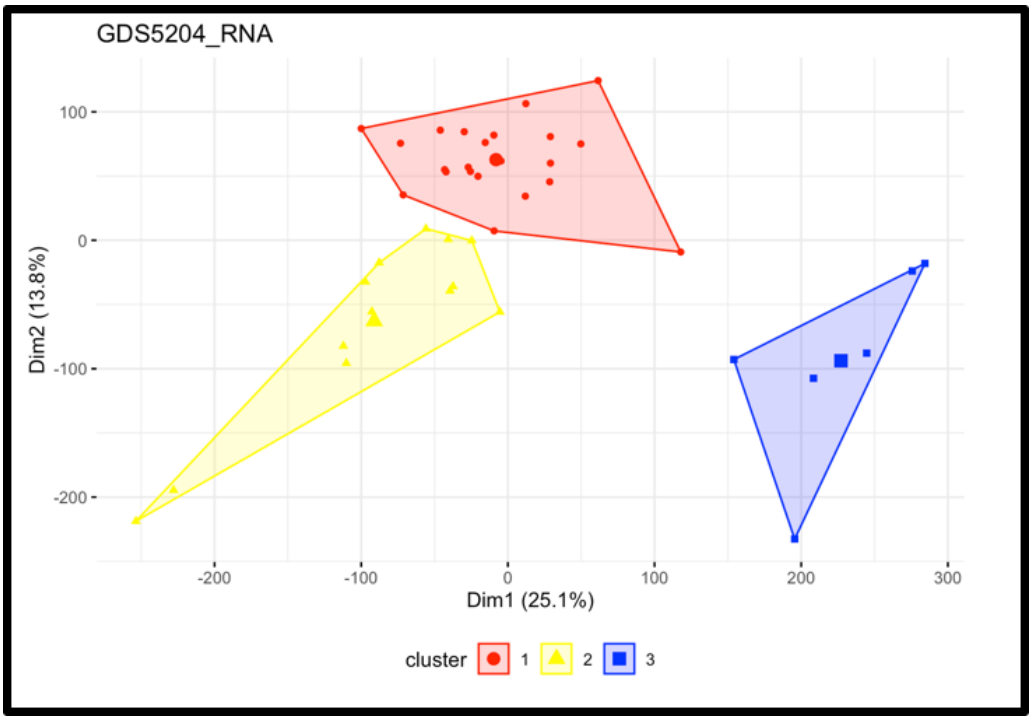


Figure 9: K-Means Clustering with K=3 for Brain Aging Dataframe

Principal Component Analysis

Principal component analysis (PCA) is a technique used to reduce the dimensionality of datasets, increasing interpretability while minimizing information loss. It accomplishes this by generating new uncorrelated variables that gradually maximize variance.

Preserving as much variability as possible means identifying new variables that are linear functions of those in the original dataset, maximize variance sequentially, and are uncorrelated with one another. Finding such new variables, known as principal components (PCs), is essentially the same as solving an eigenvalue/eigenvector problem.

The standard scalar normalizes the data, which is necessary before performing PCA. In the case of PCA, we first find the mean-centered Empirical Covariance Matrix (normalized). The top k(reduced dimension) Eigenvalues of the empirical covariance matrix are then determined. These Eigenvalues correspond to the direction of maximum variance in the data, and the corresponding eigenvectors are perpendicular to each other.

We fit the data matrix and convert it to PCA space. This is the variable pca-data. The percent explained variance is then calculated to determine how much of the variance in the data is explained by each PC. The cumulative proportion is the accumulated amount of explained variance. For example, if we used the first 10 components we would be able to account for >95% of total variance in the data.

For standardized scaling, fitting, transforming, and decomposing, the **sklearn** library in python was utilized. Finally, the data points in PC1, PC2, and PC3 were plotted in 3D space, where each datapoint represents a sample/subject. This analysis is displayed in Figures 10-15.

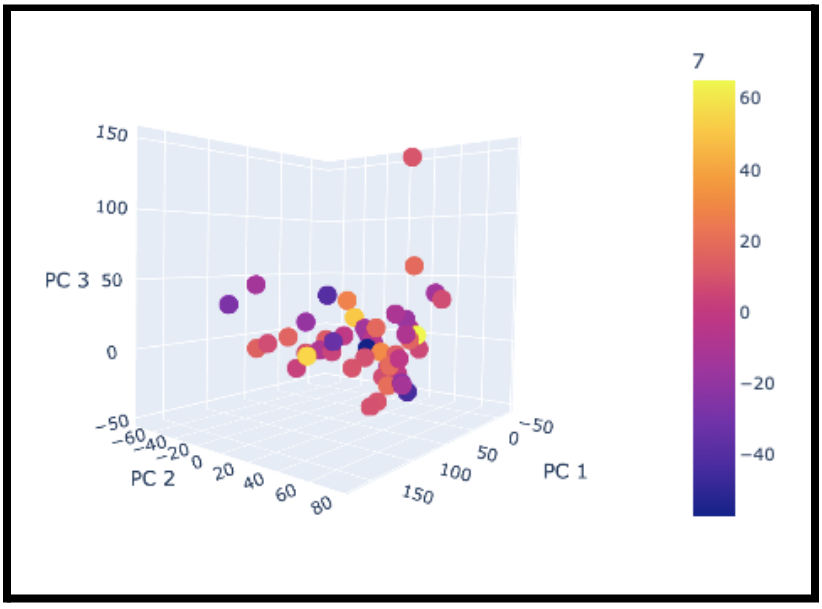


Figure 10: Principal Component Analysis of 3 PC's

K-means on PCA reduced data

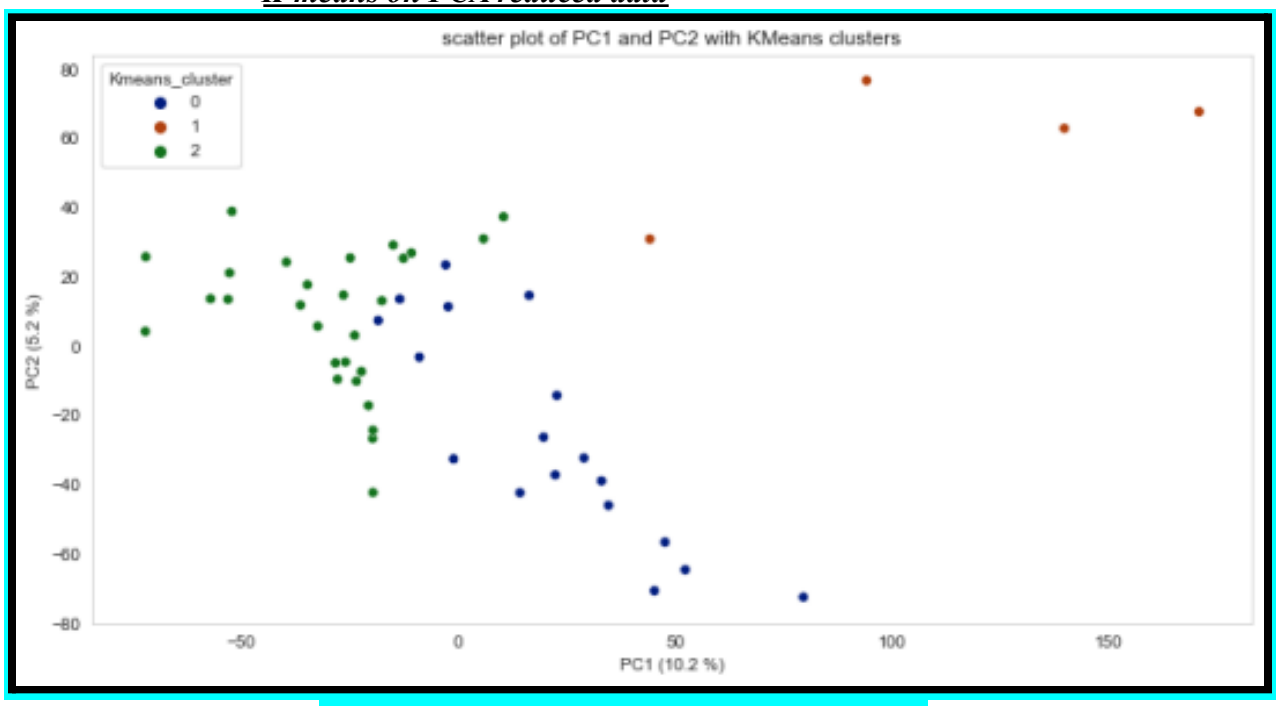


Figure 11: PC1 vs PC2 with K-means cluster

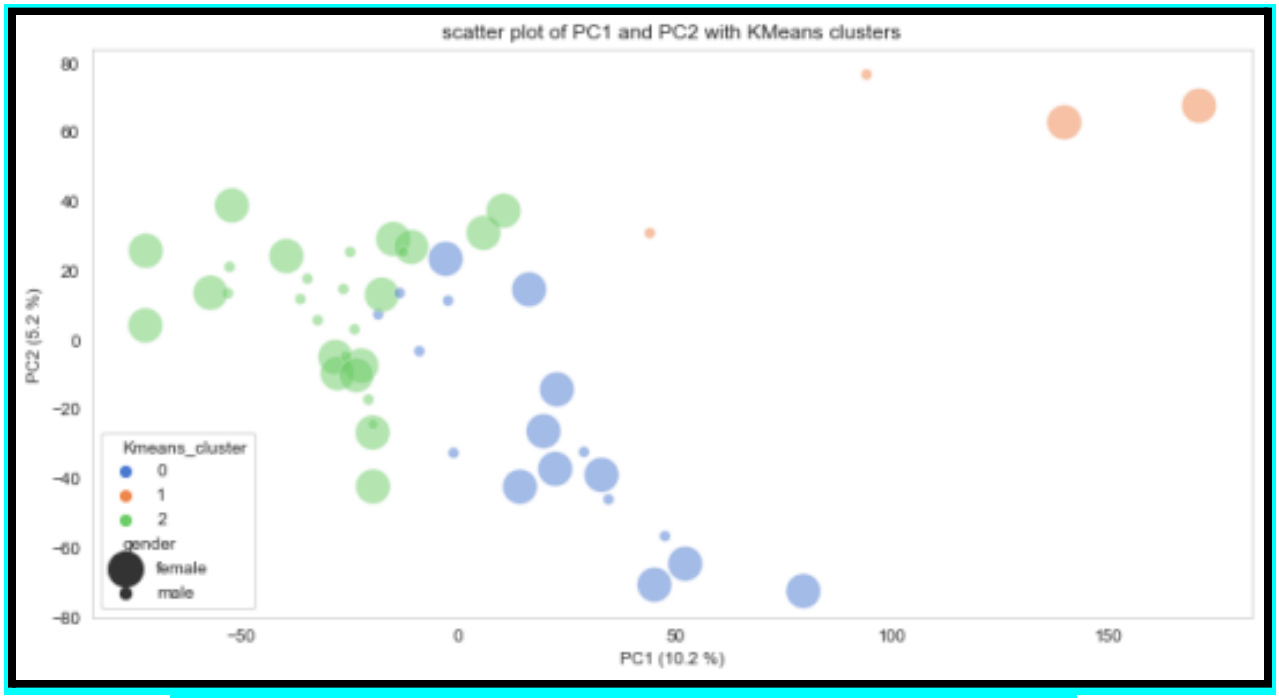


Figure 12: PC1 vs PC2 with K-means cluster (Gender enrichment)

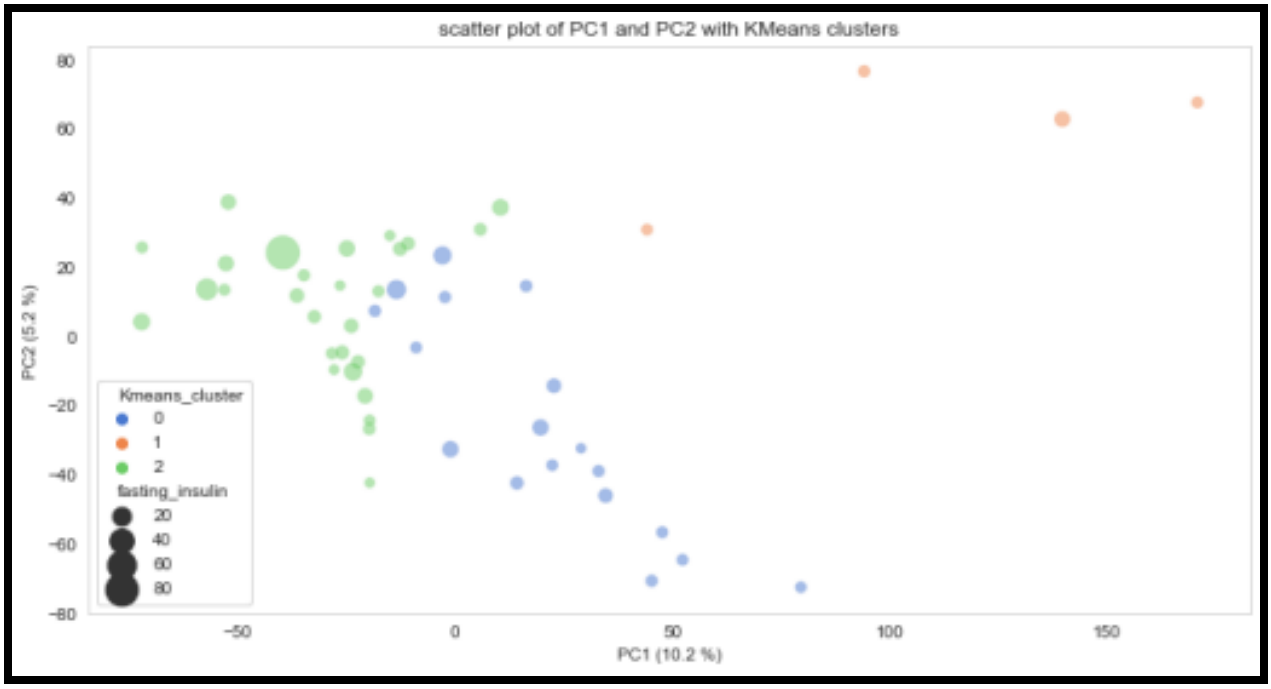


Figure 13: PC1 vs PC2 with K-means cluster (fasting_insulin enrichment)

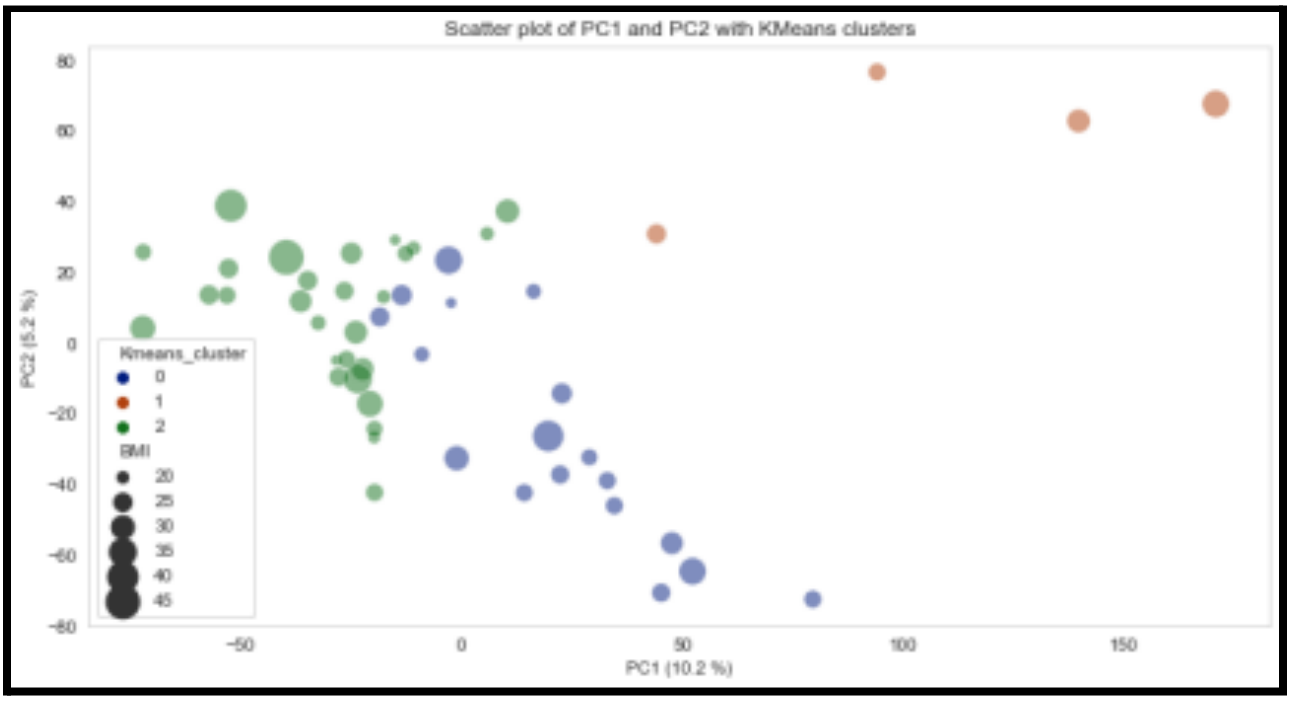


Figure 14: PC1 vs PC2 with K-means cluster (BMI)

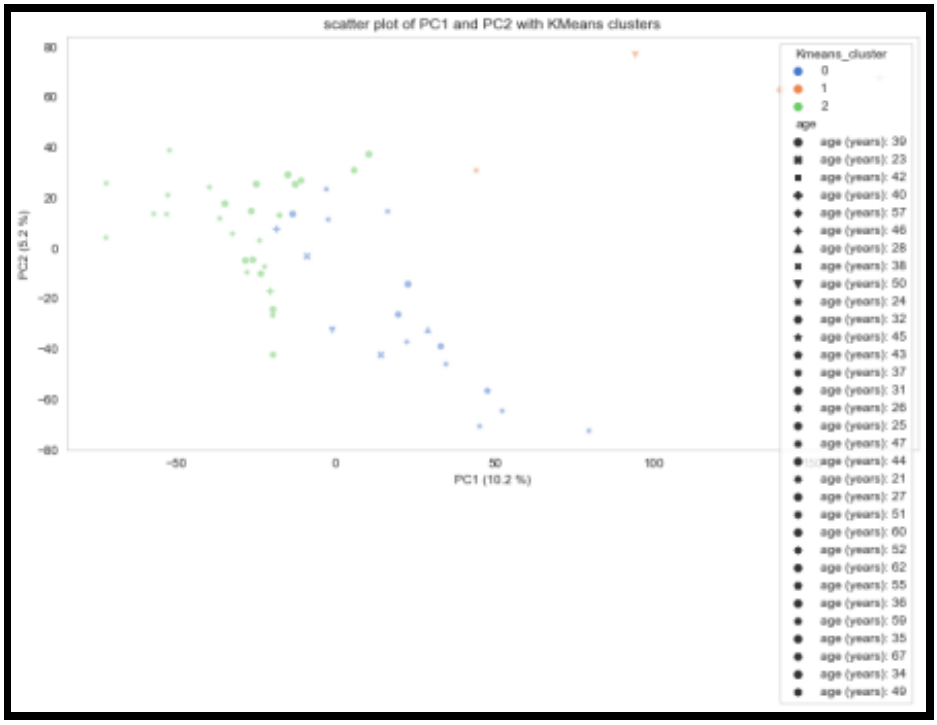


Figure 15: PC1 vs PC2 with K-means cluster (Age)

Regression models to understand gene phenotype enrichment:

Regression models were used to analyze whether the genes in both datasets were enriched in certain phenotypes such as gender, age, and BMI. A regression model is used to analyze whether the variable (young-adults/middle-aged/old adults, male/female) is predictive of being in a particular cluster. In other words, it is useful to analyze the importance of each variable in predicting what cluster each sample belongs to.

To determine if the previously plotted clusters are enriched in certain phenotypes, all the variables utilized for the regression model needed to be set to the proper classification. Cluster was set to an as.factor variable, gender was set to an as.factor variable, and body mass, age, and fasting glucose were set to numeric variables.

Four phenotypes, gender, BMI, age, and fasting glucose were selected and the table summary function was used to specify the columns of the table as clusters. Next, for all of the continuous variables, the mean, standard deviation, median, minimum, and maximum were calculated. The percentage was calculated for the categorical variables. In the table summary, (Table 1 in the Results Section,) the p-value and q-values were added as additional columns. Furthermore, the total number of patients were added as a column. Ultimately, the statistics label was included to specify which test was used to calculate the p and q-values. The Fisher's Exact

test [1] was used to calculate the gender p-value and the Kruskal-Wallis rank sum test [2] was used to calculate the p-value for the BMI, age, and fasting glucose.

[1] *Fisher's Exact Test:* Fisher's exact test is a statistical significance test used in contingency table analysis. The test is useful for categorical data resulting from classifying objects in two ways; it is used to investigate the significance of the association (contingency) between the two types of classification.

[2] *Kruskal-Wallis rank sum test:* The Kruskal-Wallis rank test is a non-parametric alternative to the one-way ANOVA test that extends the two-samples Wilcoxon test when there are more than two groups. It is advised when the assumptions of the one-way ANOVA test are not met.

Concerning the table summary, the p-value for gender is calculating the p-value of gender independent of the other variables. The p-value is testing the single hypothesis of whether the gender is equally or not equally distributed among the different clusters. On the other hand, the q-value is based on a multi-hypothesis analysis testing whether the gender among the other variables included in the table remains significant (if we can still reject the null hypothesis).

To visually represent whether male and female patients cluster, a plot of patients clustered by gender was displayed for both data frames. A continuous representation of the age was also plotted using the same axis dimensions as the gender clustering plot. As seen in Figures 26 and 27 in the Results Section, the color represents the continuous variable age where blue represents low age and red represents high age.

Correlation plots for the top 10 age related genes for both datasets

The genes correlated to age were identified by using the cor.test function in R. The p-score and t-score of each of the genes in the dataset was calculated with age and the genes corresponding to the lowest 10 p-value were displayed. This step was repeated for the second data frame. The correlation score can be a positive or negative number between 0 and 1. For the first dataset, UNC13C was the gene with the highest correlation with age, with a corresponding p-value of 2.278 *10^-6. This clear positive correlation is displayed in Figure 3.

To calculate which of these correlation values were significant and which were non-significant, a table was created displaying the t-test score and p-value for each gene correlated to age. The p-values and associated genes were arranged in ascending order (from lowest to highest). All significant values (with p value less than 0.05 were stored). Next, the top 5 positively correlated genes and top 5 negatively correlated genes (top 5 significant genes with

the best correlation with age and the top 5 significant genes with best inverse correlation with age) were selected based on t-value and used to plot a correlation matrix heatmap.

Gene-to-phenotype correlations for the top 10 genes was also conducted for both papers. The RNA expression of the top 10 genes from the 50 patients was extracted and stored. A correlation matrix was used. Firstly, a column was added to the data frame containing expression of the top 10 genes that will display the age. A correlation matrix was then done using the resulting 11 column data frame (10 genes and the age) for both datasets. This is displayed in Figures 28 and 29 of the Results Section

Box plots for the top 10 age related genes for both datasets to show significantly different genes

The box plot depicts the distribution of expression values for each sample in the data set. For the different samples to be comparable, these distributions must be similar. If this is not the case, the data should be normalized using techniques such as Quantile Normalisation for one-color systems and Scale Normalisation for two-color systems.

We took the top ten genes correlated to age and performed distribution analysis for three groups (Young Adults, Middle-aged Adults and Old Adults). The top ten genes for Insulin Resistance and Diabetes Muscle Dataset were C2orf88, COMMD7, DLAT, DSCAM.AS1, MFF, MRPL15, PCDH9, SCN4B, SLIT2 and UNC13C. The black lines in the boxplot represent medians and interquartile ranges of relative gene expression and p-scores respectively. Whiskers represent minimum and maximum 1.5 interquartile range and dots are outliers.

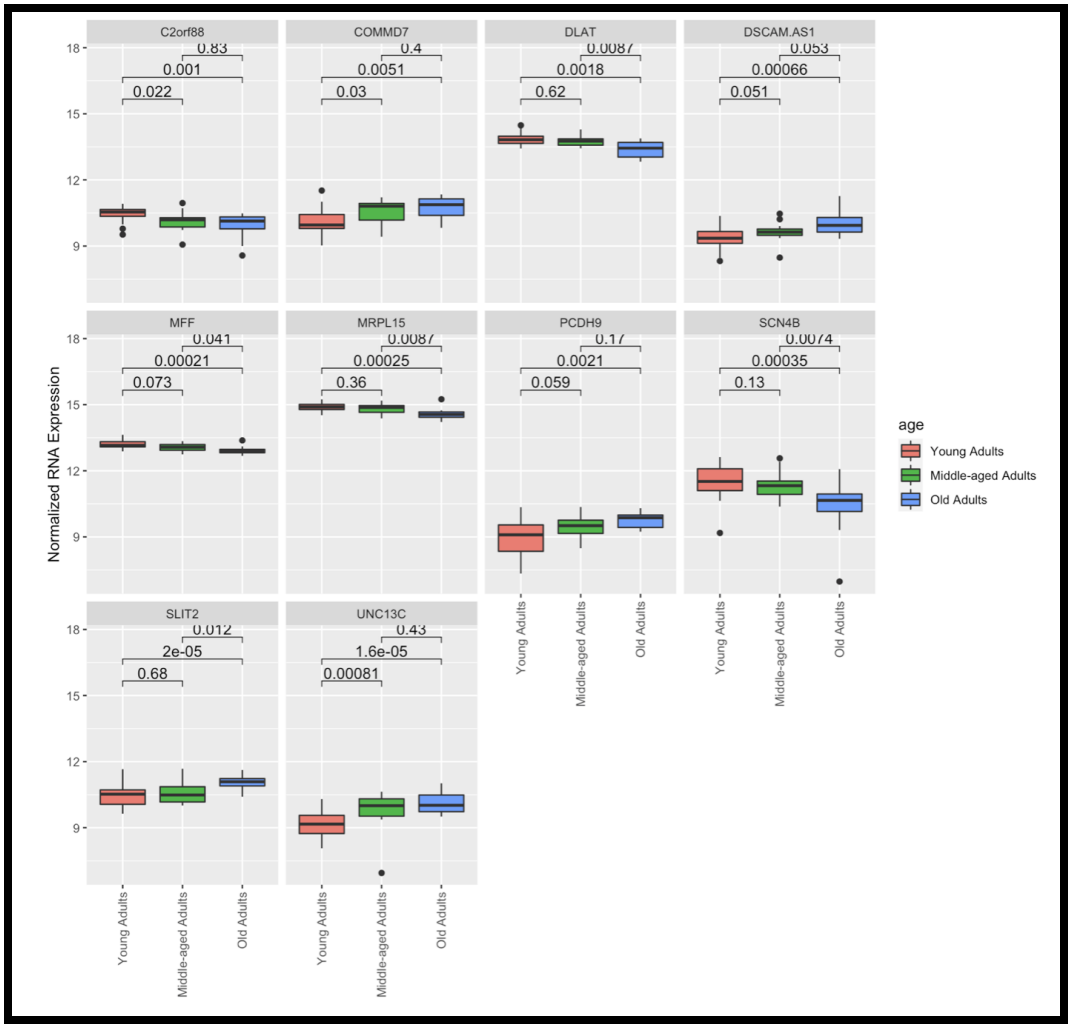


Figure 16: Box plots for the top 10 age related genes for both datasets to show significantly different Genes for Insulin Resistance and Diabetes in Muscle Dataset

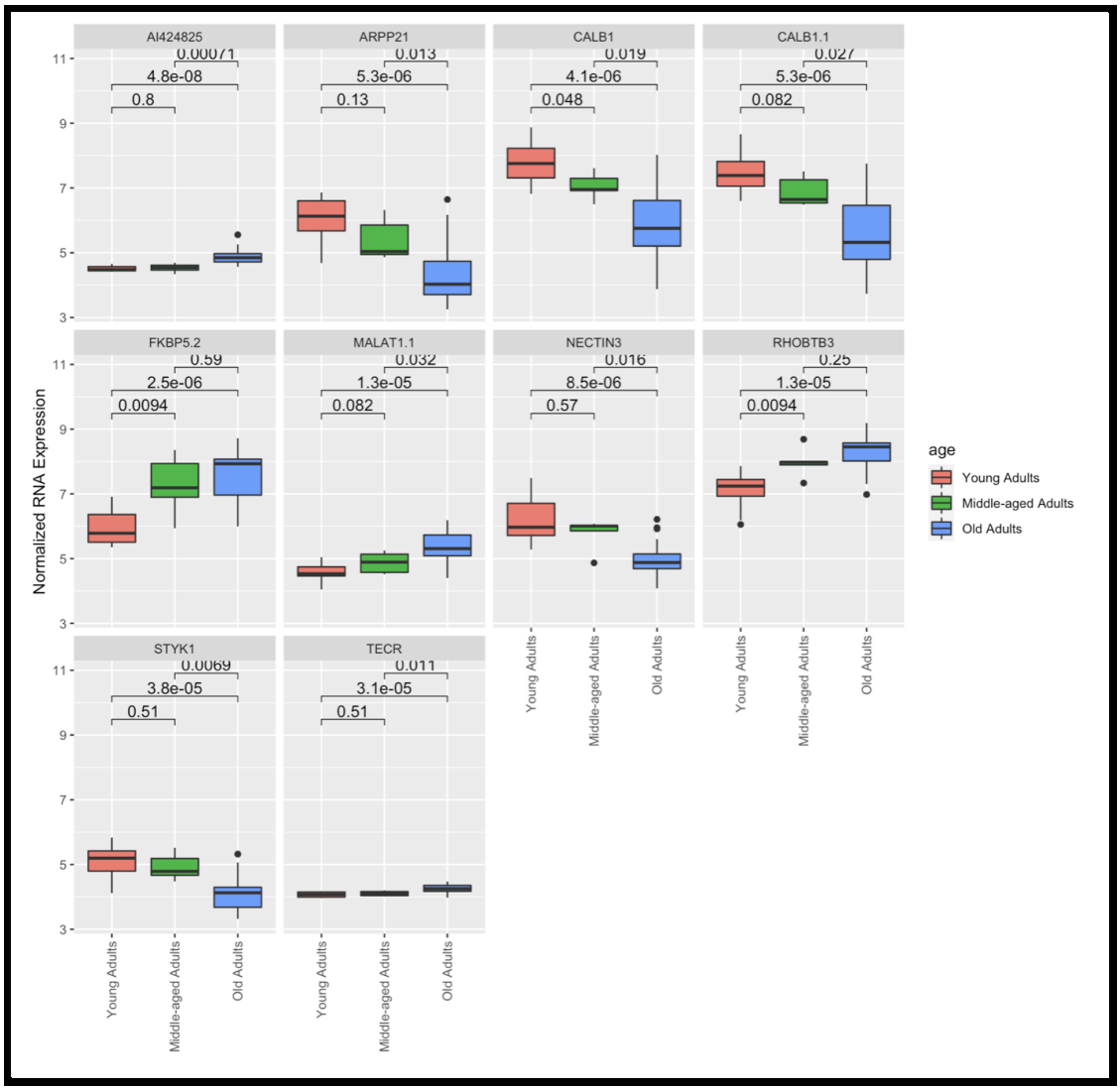


Figure 17: Box plots for the top 10 age related genes for both datasets to show significantly different Genes for Insulin Resistance and Diabetes in Muscle Dataset

Decision Trees Methods

Supervised learning is an approach to creating artificial intelligence (AI), where a computer algorithm is trained on input data that has been labeled for a particular output. Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. The steps to build the BMI and hemoglobin decision trees are outlined below:

1. Finding correlation of genes to phenotypes hemoglobin and BMI:

   The gene expression dataset of Insulin Resistance and Diabetes in Muscle was transposed

to get genes in the columns and samples in rows. Correlation of this data-frame was done with the hemoglobin data-frame and the Body Mass Index data-frame, two single column matrices obtained from the phenotype data-frame. The corrwith function of pandas was used to find the correlation coefficients. **Figure 18 A and 18 B** shows the distribution of the correlation coefficients in all of the genes for both phenotypes.

2. Filtering top genes that correlate highly with hemoglobin and BMI phenotype:
The correlation data-frame was sorted according to its absolute values and top 10 or 20 rows were selected which correspond to the top genes.

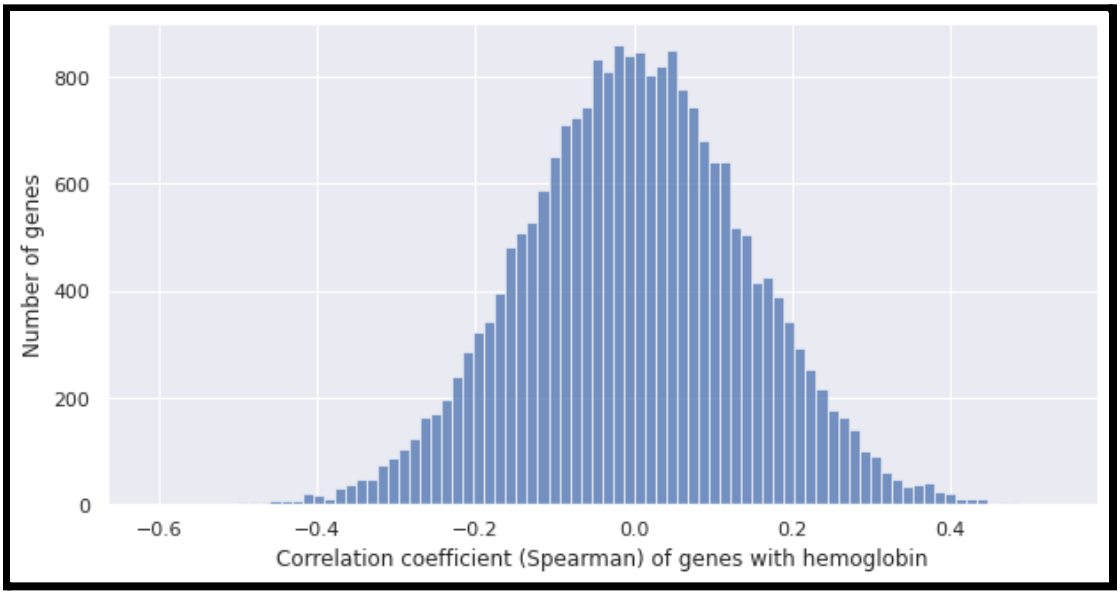


Figure 18 A: Histogram plot of correlation of genes with phenotype – hemoglobin

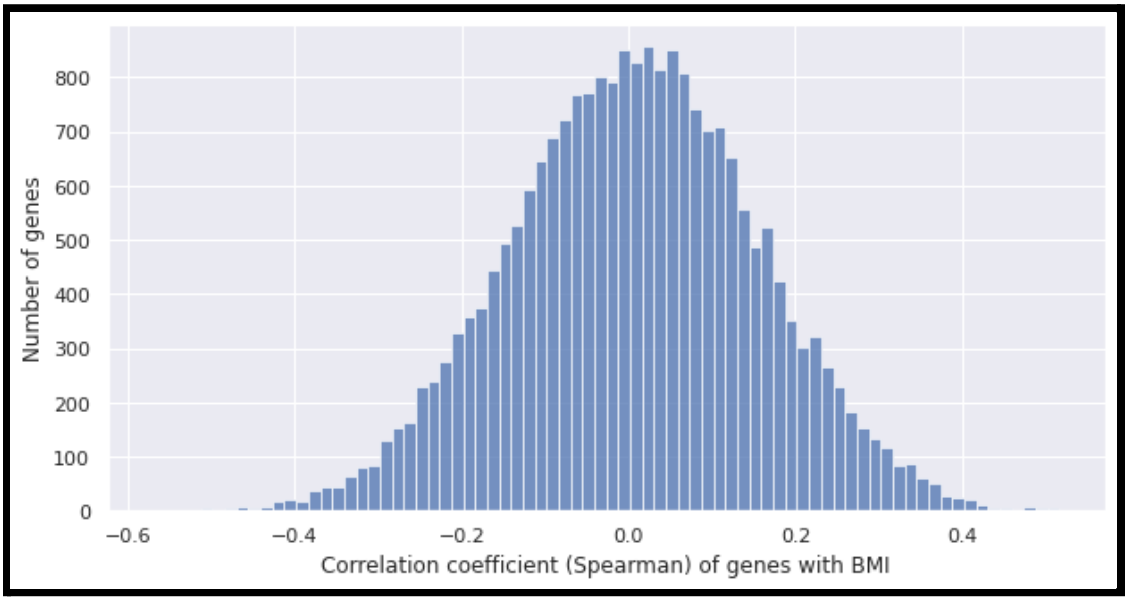


Figure 18 B: Histogram plot of correlation of genes with phenotype – BMI

3. Classifying the phenotype data into categorical labels:

A histogram plot in **Figure 19 A and 19 B** shows the distribution of the numerical values of phenotypes with the counts. This gives us an idea to categorize the distinct numerical values to three or four levels. This was done for both the phenotypes. Hemoglobin data was classified into three classes while BMI was classified into four classes.

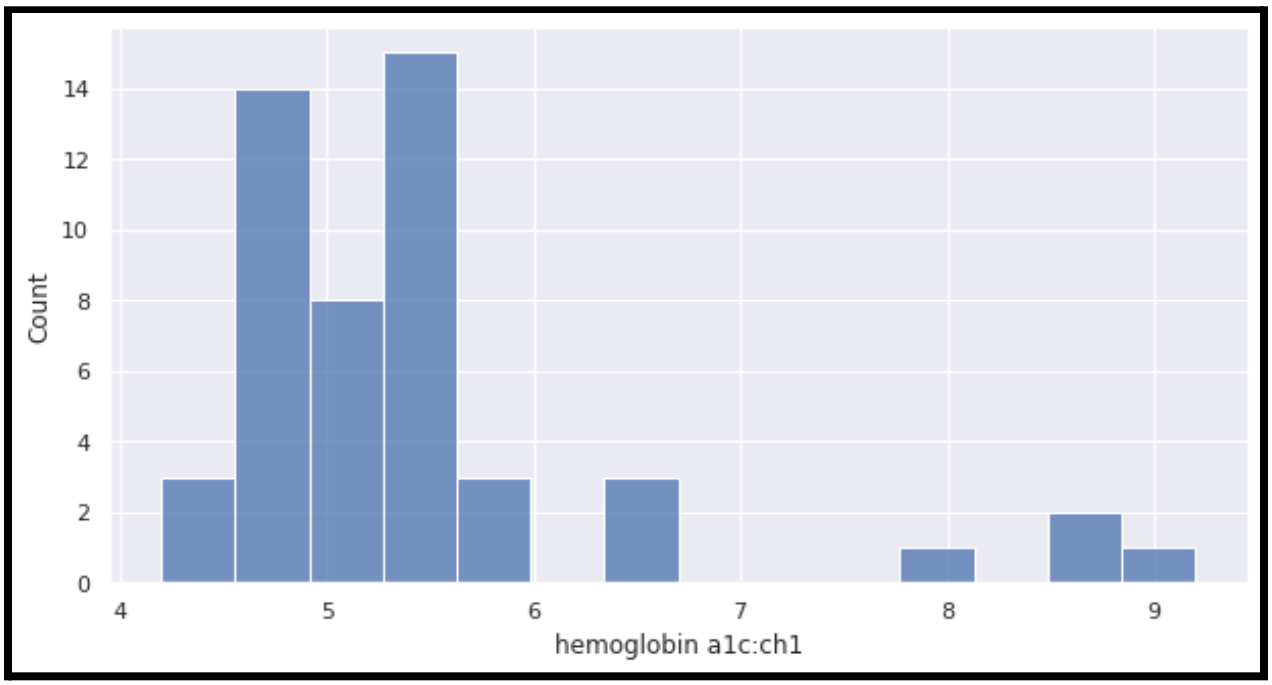


Figure 19 A: Histogram to show the distribution of hemoglobin in the samples

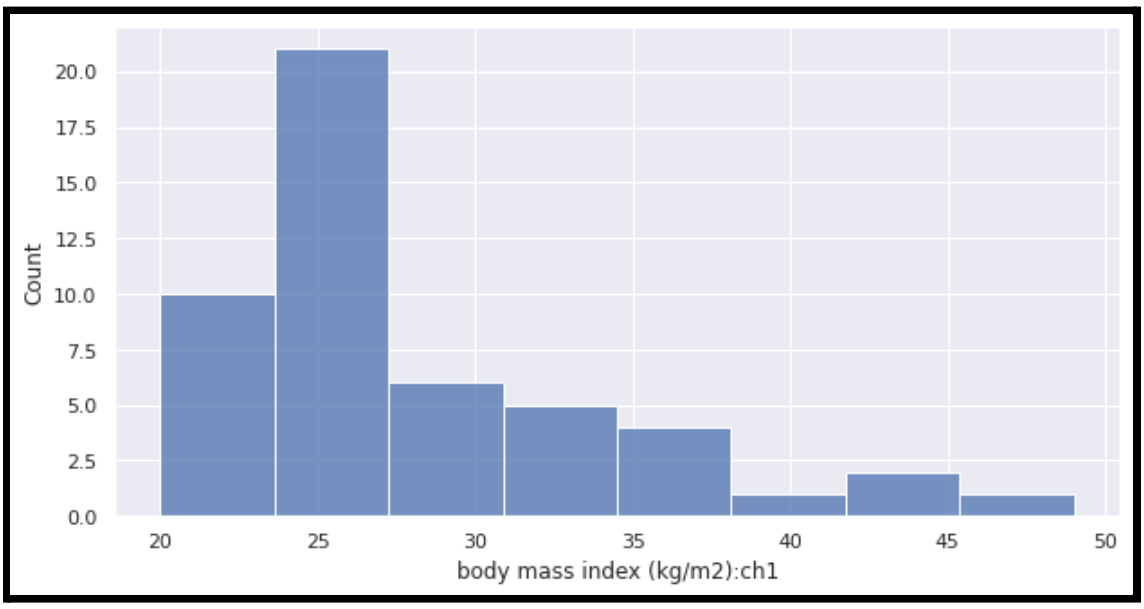


Figure 19 B: Histogram to show the distribution of BMI in the samples

4. Partitioning dataset and Supervised Learning:

After labeling the phenotype data for all samples we divided the dataset into two parts. The first part taking up about 85% was set aside to train and the rest was to test the model. The supervised learning algorithm of Decision Trees was used to create a prediction model. **Figure 20** shows the decision tree plot generated after training. DecisionTreeClassifier function from the tree library of sklearn was used for this purpose.

5. Prediction and confusion matrix:

The decision tree model was used to predict the labels of the test dataset and the predictions were compared with the actual values. A confusion matrix was generated using confusion_matrix function of the sklean.metrics package.

The correlation of genes with the BMI and hemoglobin phenotypes was calculated, the top 10 most highly correlated genes were found. Furthermore, the gene expression data for those genes was used to train a decision tree model for the selected phenotypes and accurately predict the test phenotype labels. This is displayed in Figure 20.

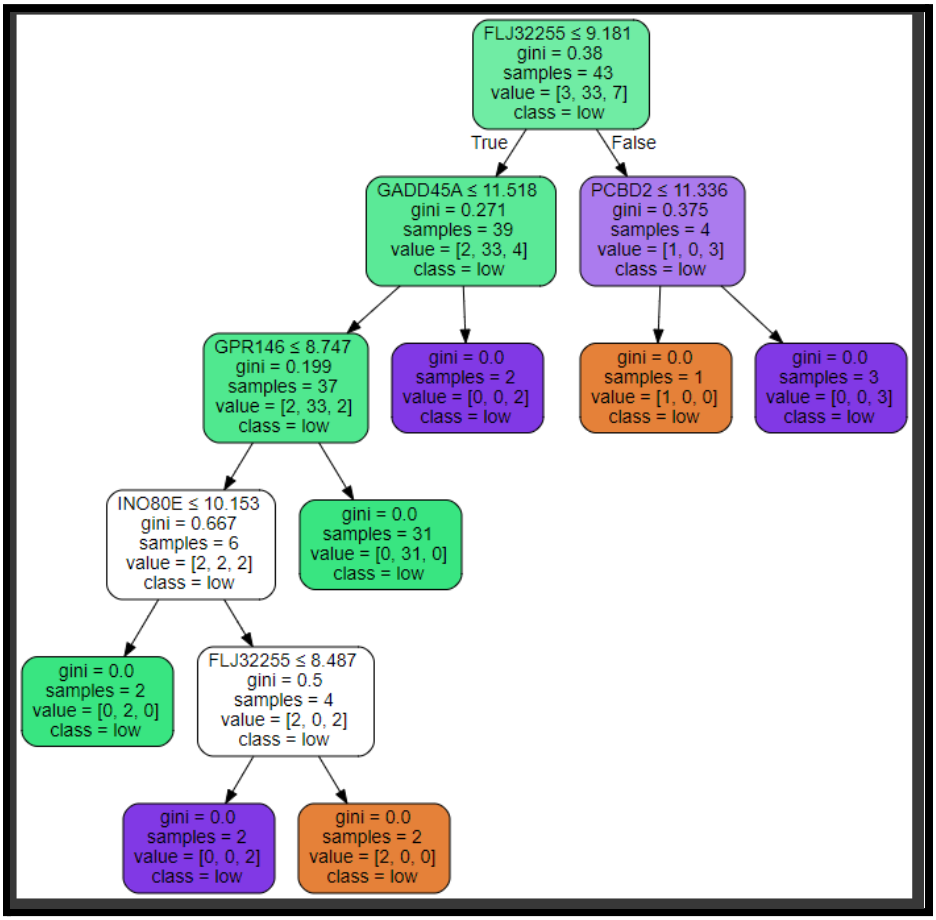


Figure 20 A: Decision Tree plot for the supervised learning of top gene expression to classify into different hemoglobin levels of low medium and high

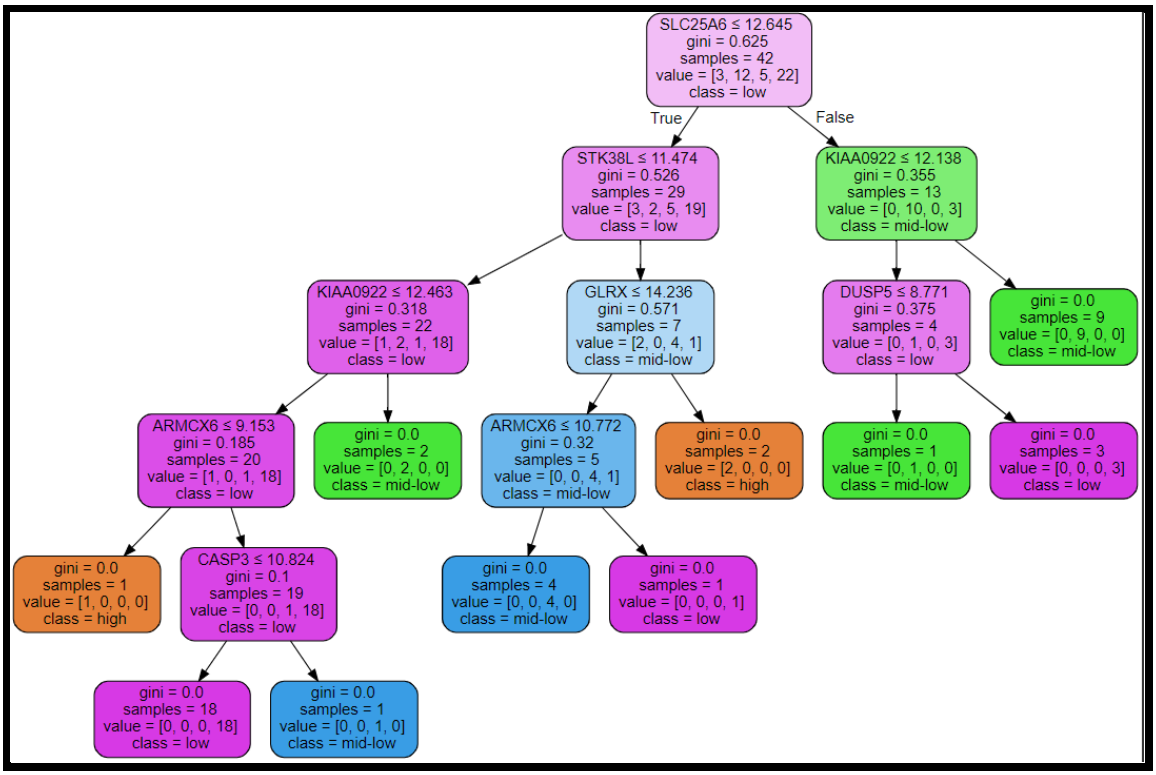


Figure 20 B: Decision Tree plot for the supervised learning of top gene expression to classify into different BMI levels of low, mid-low, mid-high and high.

Next, another regression decision tree was built for both datasets. The genes that were differentially expressed (top 10 genes correlated with age) were used to predict the age. In other words, we are trying to use the top 10 genes to predict the age group (as a classifier).

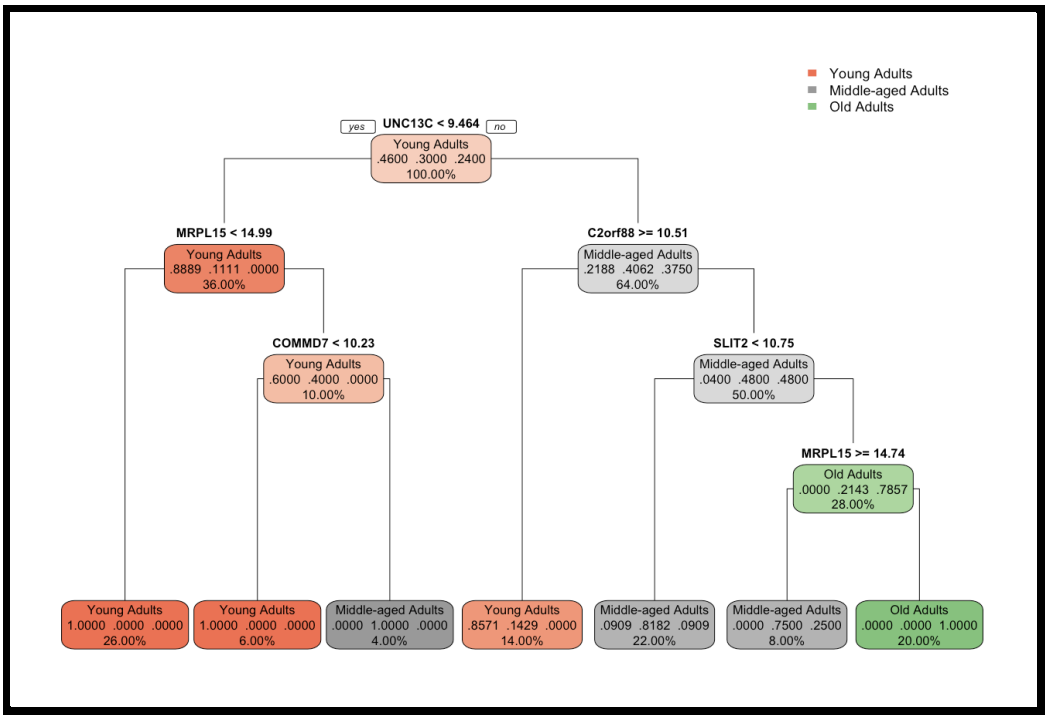


Figure 21 A: T-test on the top 10 age correlated genes for Insulin Resistance and Diabetes in Muscle Dataset

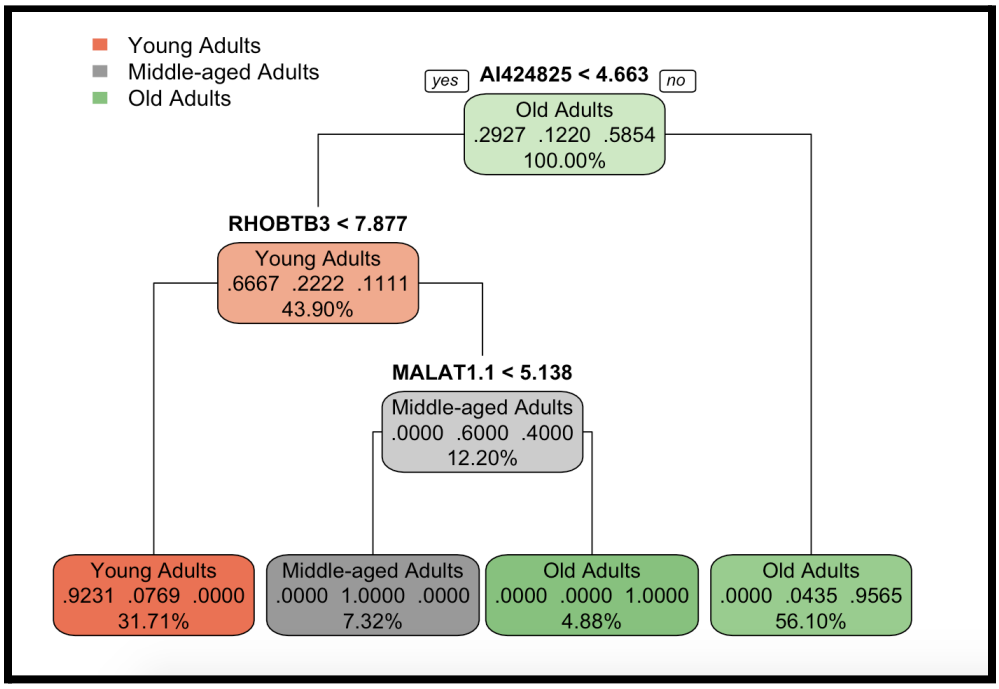


Figure 21 B : T-test on the top 10 age correlated genes for Insulin Resistance and Diabetes in Muscle Dataset

Neural Networks

Neural networks were formed with three different thresholds to **predict** which category the 5 most positively correlated and 5 most negatively correlated with age genes belonged to (Young Adults, Middle-aged adults, or Old adults). The threshold is a numeric value specifying the threshold for the partial derivatives of the error function as the stopping criteria.

Pearson and Spearman Analysis

There are two types of methods of correlation – Pearson method and Spearman method.

•Pearson correlation: Pearson correlation evaluates the linear relationship between two continuous variables

•Spearman correlation: Spearman correlation evaluates the monotonic relationship.

Correlation Matrix

The gene expression data for Insulin Resistance and Diabetes in Muscle and the gene expression data for brain aging was taken and a correlation matrix was calculated using the corr() functionality of pandas library in python for both the datasets. Both methods – Spearman and Pearson were used to calculate the correlation coefficients. **Figure 4 A & B** in the **Appendix** Section shows the first 5 rows from each of the coefficient matrices for the datasets of diabetes and brain aging respectively. It was found that both the methods of finding coefficients produced the exact same values on the first dataset and produced different values on the second dataset. As the two methods produced either same or comparable correlation values the person correlation matrix was used for further analysis. Code for this can be found in corr.py

Producing networks

The correlation matrix was then used to create a graph object using the networkx library of python. This graph object was used for further analyses of the networks in the correlation matrix. Circular layout plots were plotted for the nodes in the graph filtered for different values of coefficient threshold. Code for this can be found in networks.py

Degree distribution

The degree values were accessed from the filtered graph object using the degree functionality of the network library. A distribution plot was plotted to analyze the degrees for different values of coefficient threshold. Code for this can be found in degree.py

GEO2R analysis

GEO2R is an interactive web tool that allows users to compare two or more groups of Samples in a GEO Series in order to identify genes that are differentially expressed across experimental conditions. The results are presented in the form of a table of genes sorted by significance. GEO2R compares original submitter-supplied processed data tables using the

Bioconductor project's GEOquery and limma R packages. Bioconductor is an open source software project based on the R programming language that provides tools for high-throughput genomic data analysis.

We extracted the dataset with GEO accession number GSE25462 and performed GEO2R to compare three groups of Samples with different Family histories (10 subjects with type 2 diabetes, 25 subjects with a family history of type 2 diabetes (one or both parents), and 15 subjects with no family history of type 2 diabetes) in order to identify genes that are differentially expressed across experimental conditions. The Results were presented as a table of genes ordered by significance (p-value <0.05).

Similarly, the dataset with GEO accession number GSE53890 was extracted and GEO2R analysis was performed to compare four groups of Samples with different age groups (12 young (<40yr), 9 middle aged (40-70yr), 16 normal aged (70-94yr), and 4 extremely aged (95-106yr)). The results were obtained as a table of genes ordered by significance (p-value <0.05). The web based GEO throws up only genes that are deemed statistically significant within that entire dataset.

These genes were then copied and pasted into DAVID for functional annotation analysis. The Database for Annotation, Visualization and Integrated Discovery (DAVID) provides a comprehensive set of functional annotation tools for investigators to understand the biological meaning behind large lists of genes. The tool compares the internal relationships of the clustered terms to the typical linear, redundant term report, in which similar annotation terms may be distributed among hundreds or thousands of other terms. Furthermore, to fully utilize the well-known KEGG and BioCarta pathways, the new DAVID Pathway Viewer, a feature of the DAVID Functional Annotation Tool, can display genes from a user's list on pathway maps to facilitate biological interpretation in a network context.

The list of genes were then analyzed and Functional Annotation Clustering was selected. For the first dataset, we discovered two pathways relating to KEGG (namely, AGE-RAGE signaling pathway and Diabetic Cardiomyopathy pathway) that could be of some relevance.

Annotation Cluster 85	Enrichment Score: 0.29			Count	P_Value	Benjamini
KEGG_PATHWAY	VEGF signaling pathway	RT		4	3.6E-1	1.0E0
KEGG_PATHWAY	Non-alcoholic fatty liver disease	RT		7	4.9E-1	1.0E0
KEGG_PATHWAY	T cell receptor signaling pathway	RT		5	5.2E-1	1.0E0
KEGG_PATHWAY	AGE-RAGE signaling pathway in diabetic complications	RT		4	7.1E-1	1.0E0

Annotation Cluster 107	Enrichment Score: 0.15			Count	P_Value	Benjamini
KEGG_PATHWAY	Sphingolipid signaling pathway	RT		5	6.3E-1	1.0E0
KEGG_PATHWAY	PD-L1 expression and PD-1 checkpoint pathway in cancer	RT		4	6.3E-1	1.0E0
KEGG_PATHWAY	Diabetic cardiomyopathy	RT		7	7.5E-1	1.0E0
KEGG_PATHWAY	Chemical carcinogenesis - reactive oxygen species	RT		7	8.2E-1	1.0E0

Figure 20: DAVID annotation clustering results for Insulin resistance and Diabetes Paper

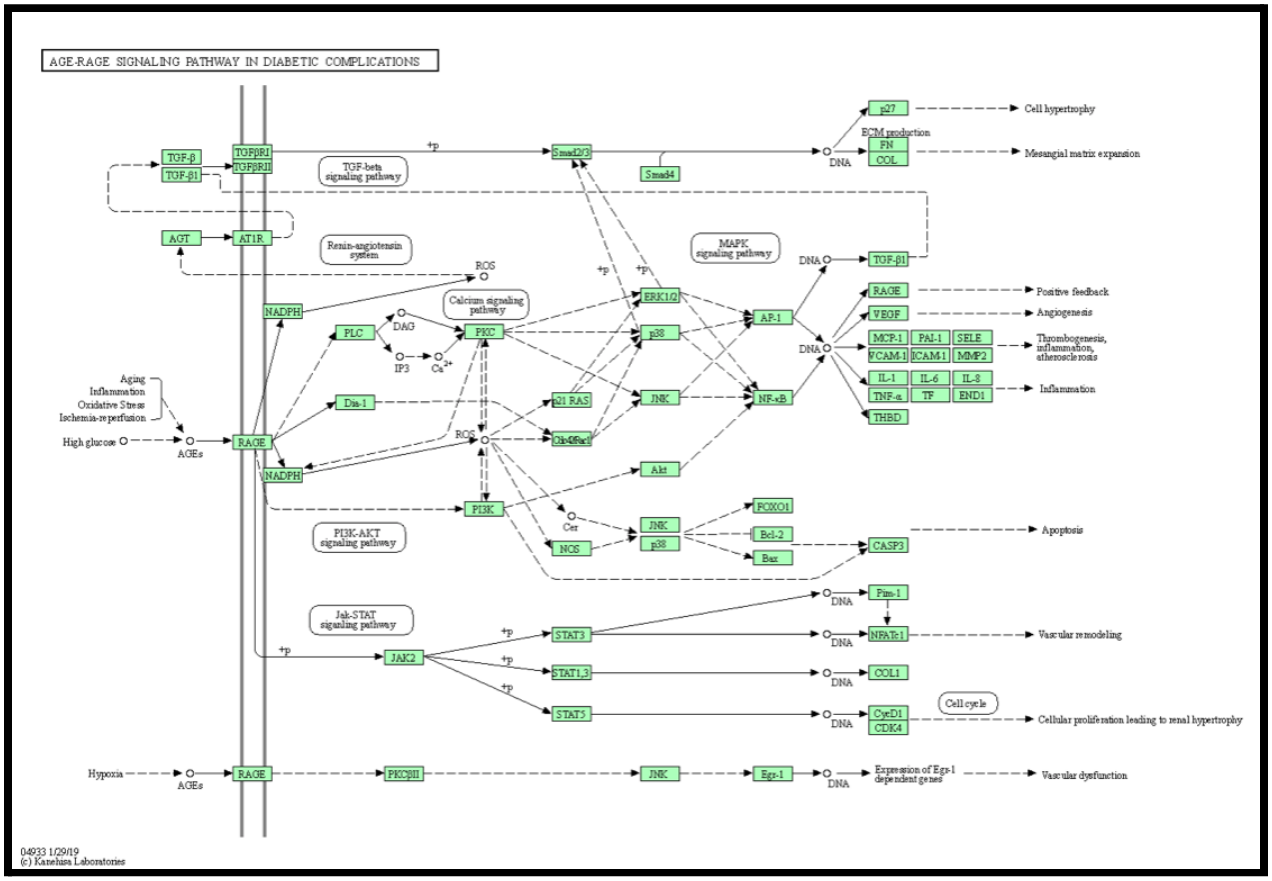


Figure 21: AGE-RAGE signaling pathway in Diabetic complications

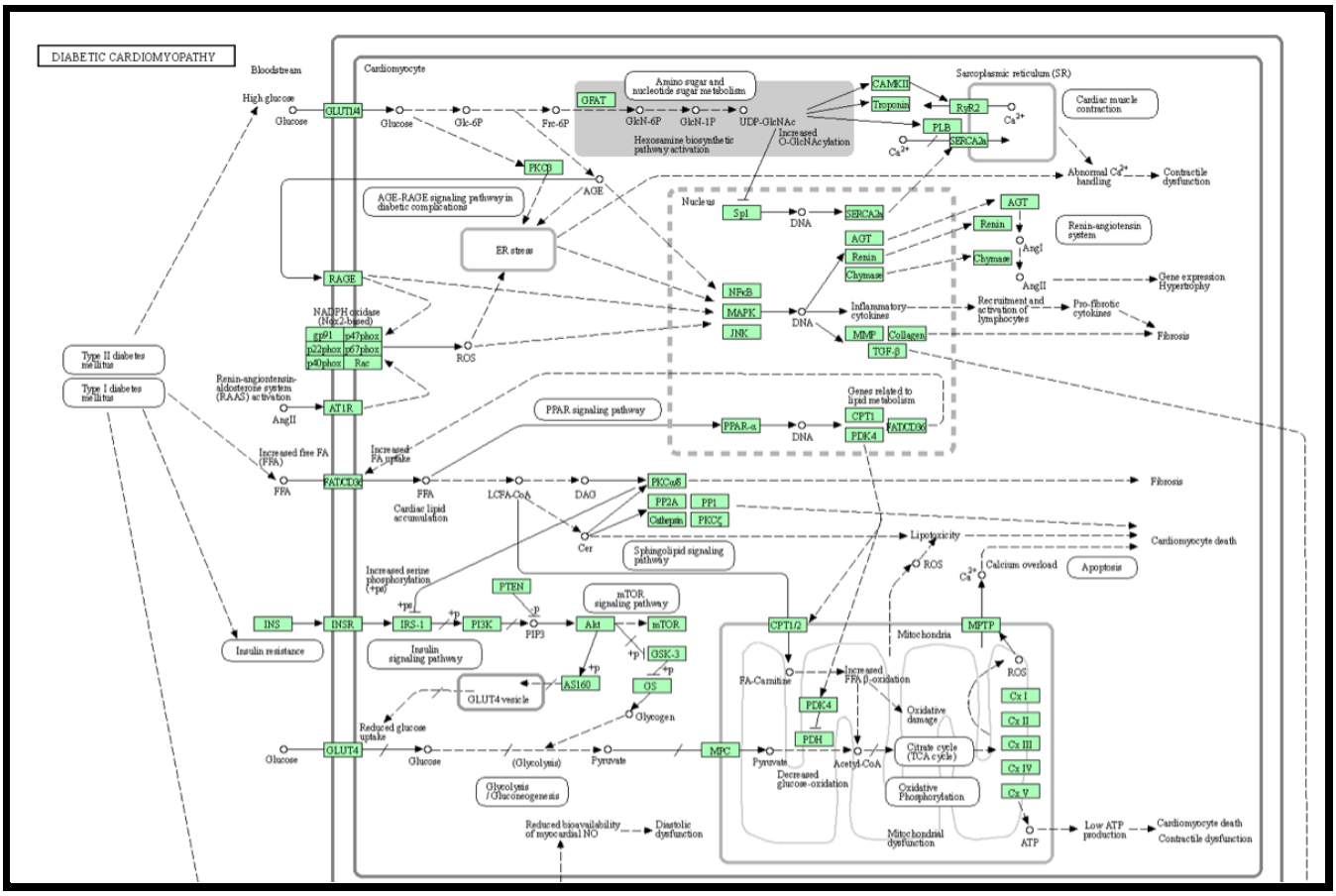


Figure 22: Diabetic Cardiomyopathy pathway

Annotation Cluster 3	Enrichment Score: 4.71			Count	P_Value	Benjamini
KEGG_PATHWAY	Alzheimer disease	RT		42	9.8E-7	3.0E-4
KEGG_PATHWAY	Huntington disease	RT		35	3.6E-6	5.6E-4
KEGG_PATHWAY	Pathways of neurodegeneration - multiple diseases	RT		46	8.0E-6	8.1E-4
KEGG_PATHWAY	Prion disease	RT		30	4.5E-5	3.1E-3
KEGG_PATHWAY	Parkinson disease	RT		29	7.1E-5	3.6E-3
KEGG_PATHWAY	Amyotrophic lateral sclerosis	RT		33	6.3E-4	1.8E-2
Annotation Cluster 4	Enrichment Score: 4.63			Count	P_Value	Benjamini

Figure 23: DAVID annotation clustering results for Alzheimer's and Aging Paper

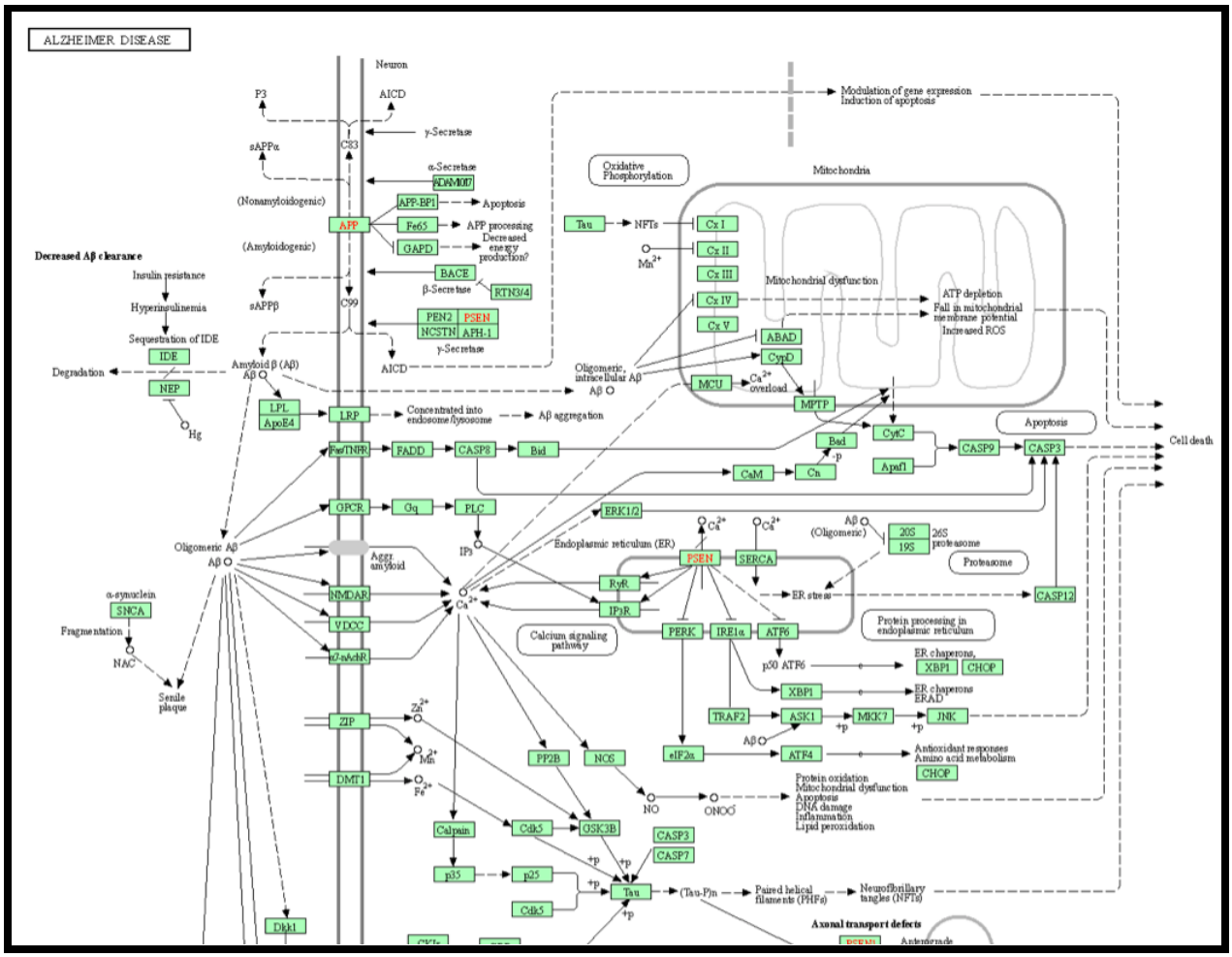


Figure 24: Alzheimer's Disease Pathway

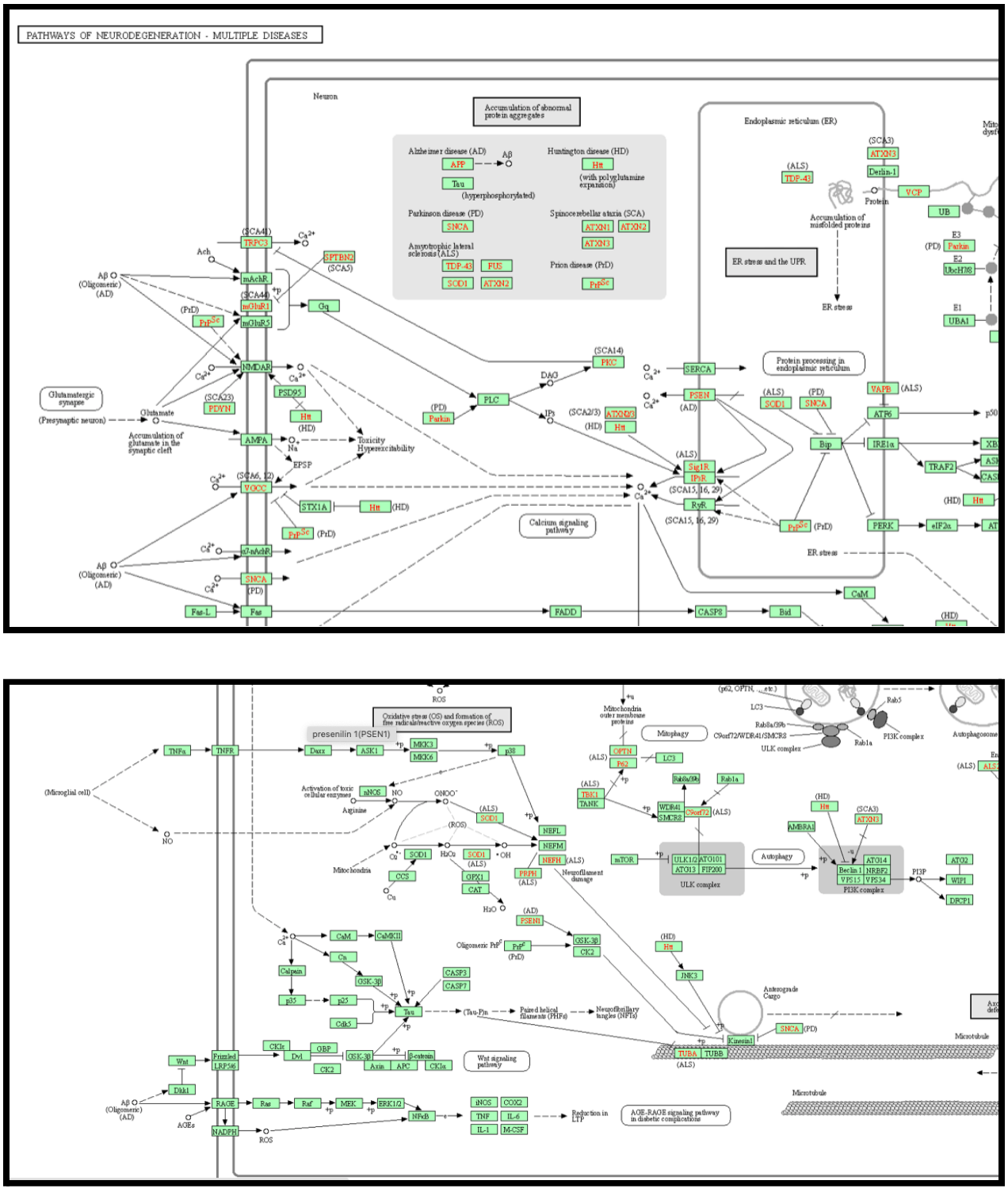


Figure 25: Pathways of Neurodegeneration - Multiple Diseases

Results:

Heatmaps of all human data Results

Heatmaps serve as a preliminary analysis to visualize the gene expression data. The generated heat maps are useful for identifying genes that are commonly regulated, or biological signatures associated with a particular condition (for e.g, in our case, insulin resistance and Alzheimer's disease related to aging). In the heatmaps in Figures 1,2 (methods section) the x-axis represents the 50 patients (samples) and the y-axis represents gene expression. In terms of their biological significance, red represents upregulation of the genes (high gene expression of the specific gene), and blue represents gene downregulation (low gene expression of the specific gene). Concerning the heatmap for the second data frame corresponding to the second research paper, the white horizontal lines observed in the Figure 2 heatmap are caused by the very low expressed genes. This is due to the fact that the second dataset contained genes that had very low expression and were not filtered out.

PCA clustering results

Figures 11-15 series is the result of the 3-means clustering with specific groups highlighted. BMI was separated into three categories, greater than 30 (obese), between 25 and 29.9 (overweight), and less than 24.9 (healthy weight until 18.5). There seemed to be no obvious clustering of BMIs, rather more overlapping in general. A similar conclusion can be found when overlaying genders for each subject, shown in figure 12**.** Overall clusters were not enriched with either gender.

Because there are so many genes, we chose those with a high absolute value. As can be seen in the image below, the cluster number is 0. The coefficient threshold was set to 0.002. A total of 687 genes were obtained here that are above the cutoff. Females were observed to be more enriched in all three clusters as compared to males in figure 12. As for BMI, higher levels of body mass index are more clustered in cluster 2 than in 1 and 3. Similar patterns were observed with fasting_insulin.

Table Summary and Phenotype Clustering Results

Table 1 presents the data to establish whether each phenotype is significant or non-significant for the first dataframe. All four phenotypes are found insignificant with p and q values all greater than 0.05.

Characteristic	Overall, N = 50	1, N = 24	2, N = 23	3, N = 3	p-value[1]	q-value[2]
gender.ch1, n/N (%)					0.7	0.9
female	28/50 (56%)	12/24 (50%)	14/23 (61%)	2/3 (67%)		
male	22/50 (44%)	12/24 (50%)	9/23 (39%)	1/3 (33%)		
body.mass.index..kg.m2..ch1, Mean+/-SD; Median (Range)	28.1+/-6.3; 26.0 (20.0 - 49.0)	29.0+/-7.1; 27.0 (20.0 - 49.0)	26.8+/-5.5; 25.0 (20.0 - 42.0)	30.7+/-5.5; 31.0 (25.0 - 36.0)	0.3	0.9
age..years..ch1, Mean+/-SD; Median (Range)	40.2+/-12.5; 41.0 (21.0 - 67.0)	39.1+/-13.6; 39.0 (21.0 - 67.0)	40.7+/-12.3; 40.0 (23.0 - 62.0)	46.3+/-3.2; 45.0 (44.0 - 50.0)	0.6	0.9
fasting.glucose..iv0gavg..ch1, Mean+/-SD; Median (Range)	100.5+/-28.9; 93.5 (77.0 - 222.0)	97.5+/-21.5; 92.5 (77.0 - 185.0)	104.3+/-36.6; 94.0 (81.0 - 222.0)	94.3+/-1.5; 94.0 (93.0 - 96.0)	0.9	0.9

[1]Fisher's exact test; Kruskal-Wallis rank sum test

[2]False discovery rate correction for multiple testing

Table 1: Table Summary of potentially enriched phenotypes for Human Data for Insulin Resistance and Diabetes in Muscle Dataframe

Characteristic	Overall, N = 41	1, N = 22	2, N = 13	3, N = 6	p-value[1]	q-value[2]
age_num, Mean+/-SD; Median (Range)	63.0+/-28.1; 66.0 (24.0 - 106.0)	46.2+/-24.0; 36.5 (24.0 - 93.0)	85.6+/-19.1; 91.0 (44.0 - 106.0)	75.8+/-15.9; 81.3 (44.0 - 87.0)	<0.001	<0.001
Sex:ch1, n/N (%)					0.3	0.3
Female	21/41 (51%)	9/22 (41%)	9/13 (69%)	3/6 (50%)		
Male	20/41 (49%)	13/22 (59%)	4/13 (31%)	3/6 (50%)		

[1]Kruskal-Wallis rank sum test; Fisher's exact test

[2]False discovery rate correction for multiple testing

Table 2: Table Summary of potentially enriched phenotypes for Brain Aging Dataframe

To visually represent whether male and female patients cluster, a plot of patients clustered by gender was displayed for both data frames. A continuous representation of the age was also plotted. One can clearly see that in the Human Data for Insulin Resistance and Diabetes in Muscle data frame there is no correlation between the age and the clusters. For example, both young and old patients are present in cluster 1, and both young and old patients are present in cluster 2 etc.. Concerning the Brain Aging data frame, male and female patients did not cluster either. However, a correlation between clusters and age was indeed observed. Cluster 1 contained young patients, cluster 2 contained patients with old age, and cluster 3 contained patients with medium age.

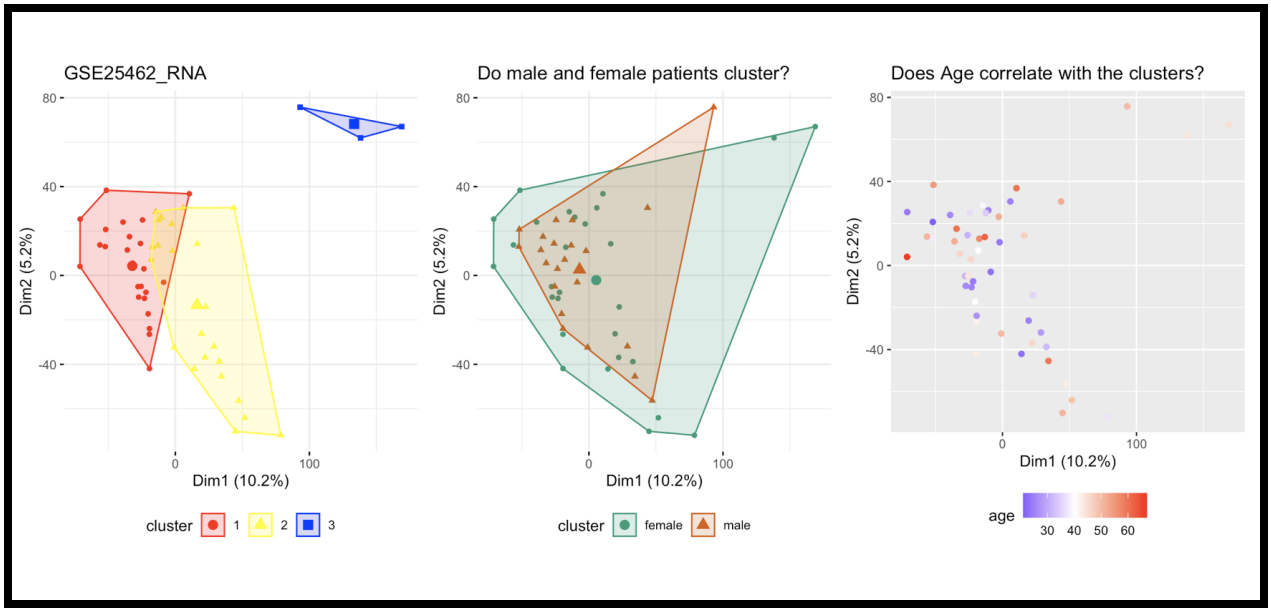


Figure 26: Categorical representations of the four clusters, categorical representation of gender, and continuous representation of age for Human Data for Insulin Resistance and Diabetes in Muscle Dataframe

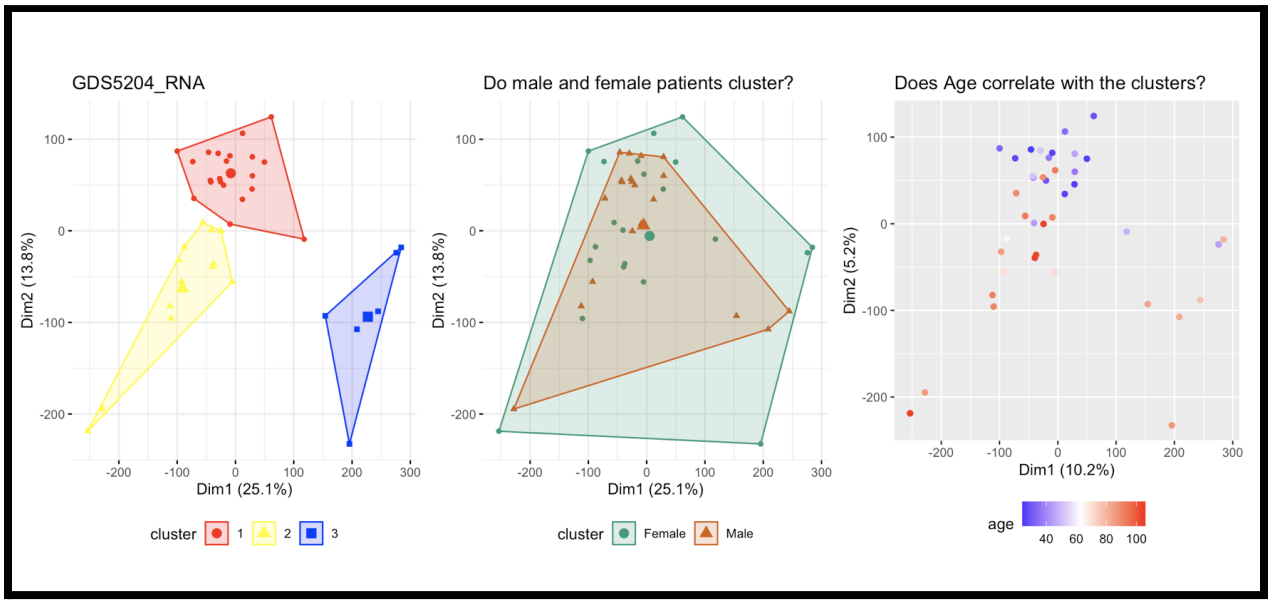


Figure 27: Categorical representations of the four clusters, categorical representation of gender, and continuous representation of age for Brain Aging Dataframe

Top 10 genes correlated with age correlation Results

Concerning the correlation of the age versus the RNA expression of the top 10 genes for the first data frame (Figure 28), one can see by looking at all of the rows and the age column, that the genes most positively correlated with age are represented as relatively dark blue dots (first five rows), and the genes most negatively correlated with age are represented as relatively dark red dots (last five rows) as seen in Figure 10. Furthermore, one can notice that the genes

positively correlated with age are also positively correlated with each other, and that the genes negatively correlated with age are also positively correlated with each other. Furthermore, when looking at the correlation between genes that are positively correlated with age and genes that are negatively correlated with age, these genes are negatively correlated with each other. The same correlation trends apply to the second Brain Aging data frame (Figure 29). However, for this data frame we were able to reach higher t-value and more negative t-values compared to the first data frame as shown in Figures 11 and 14.

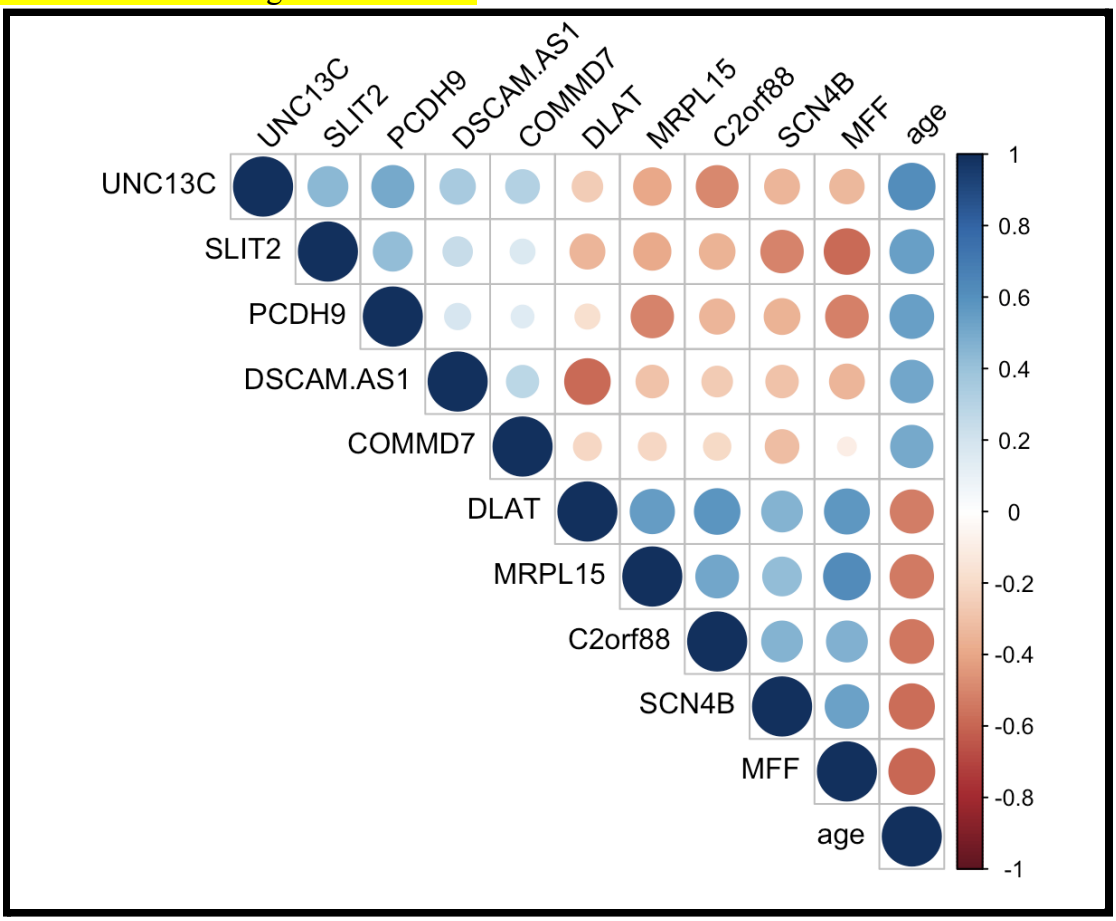


Figure 28: Correlation plot of top 10 genes correlated to age for Human Data for Insulin Resistance and Diabetes in Muscle Dataframe

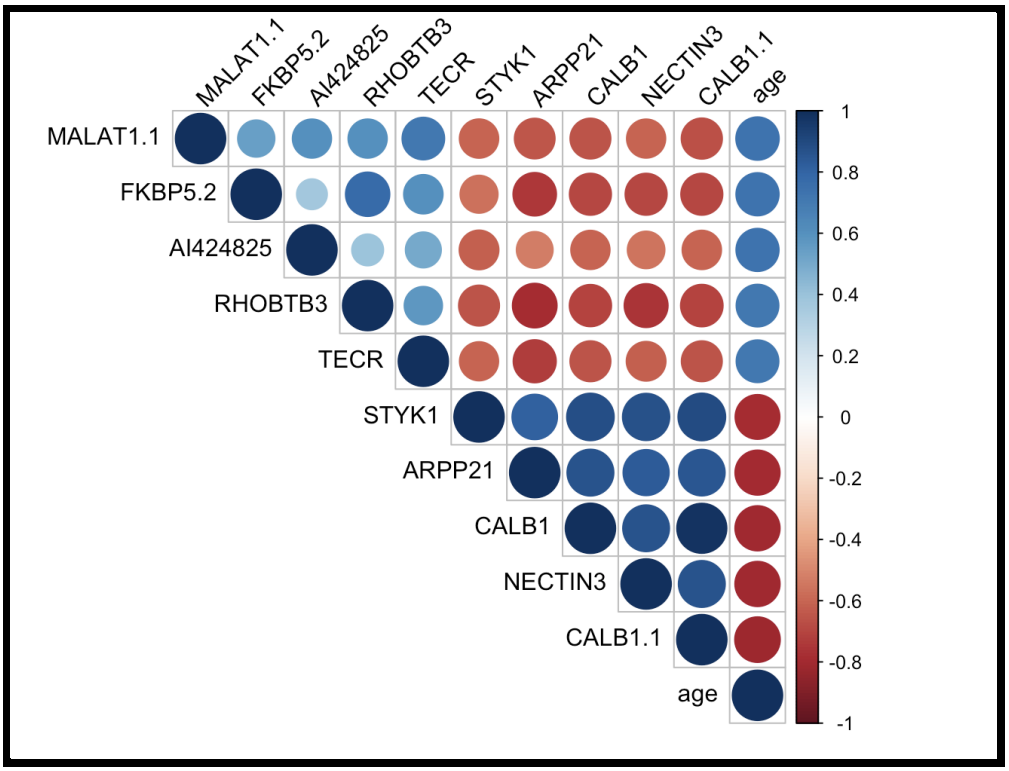


Figure 29: Correlation plot of top 10 genes correlated to age for Brain Aging Dataframe

Decision Tree Results

Concerning the BMI and hemoglobin decision trees, The top genes that highly correlate with the two phenotypes had a roughly equal distribution of positively and negatively correlated genes. Both the sets of top genes had an absolute minimum and maximum values of 0.45, 0.48 and 0.61, 0.63 respectively. **Figure 30** shows a scatter plot of the top genes with their coefficients. Red ones are negatively correlated while green ones are positively correlated.

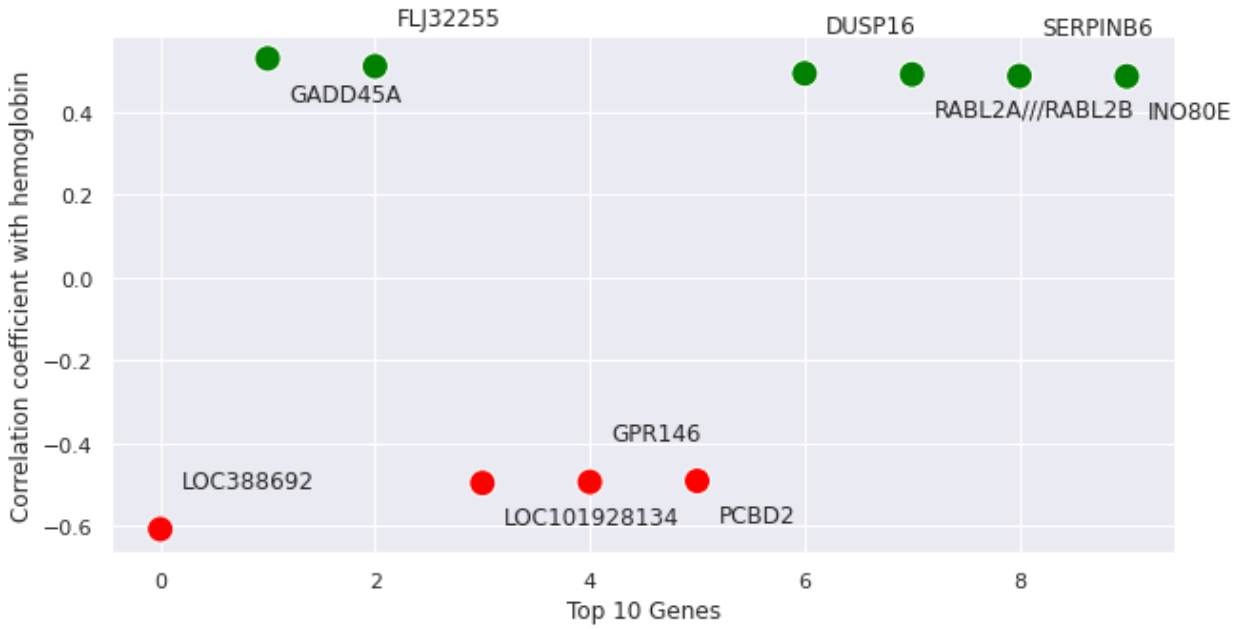


Figure 30 A: Top 10 genes highly correlated with hemoglobin.

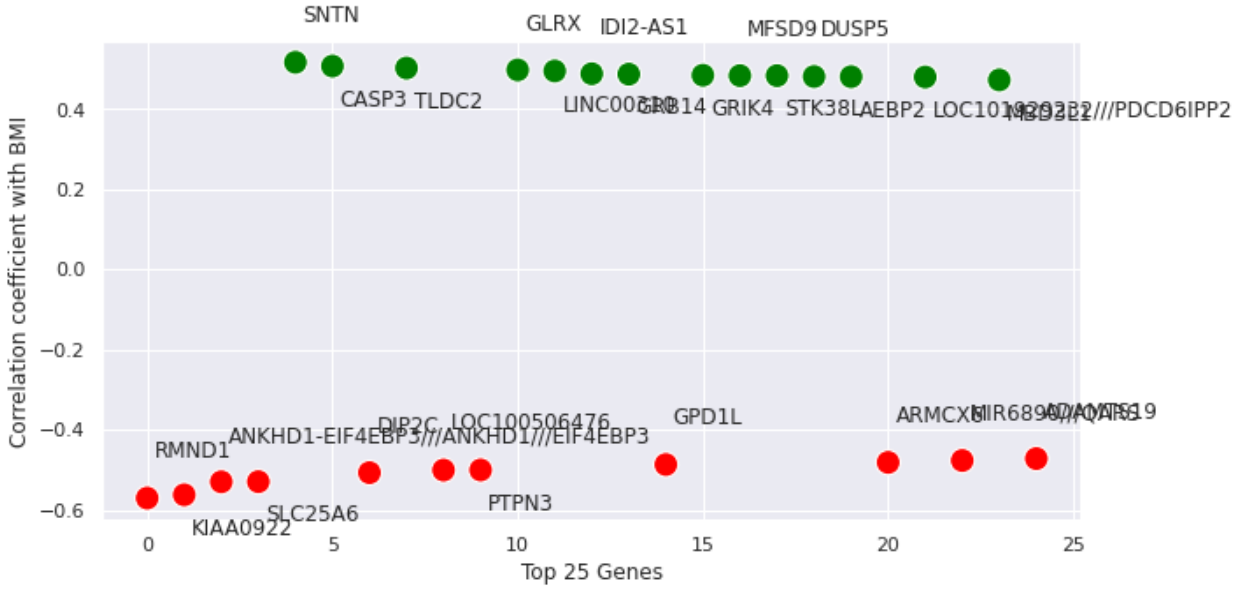


Figure 30 B: Top 10 genes highly correlated with BMI.

The predicted values were compared to the actual values. An accuracy of 71.43% was obtained for the prediction of hemoglobin test data while an accuracy of 75% was obtained for the prediction of BMI test data. **Figure 31** shows the plot of the confusion matrix showing the different labels with their actual and predicted values.

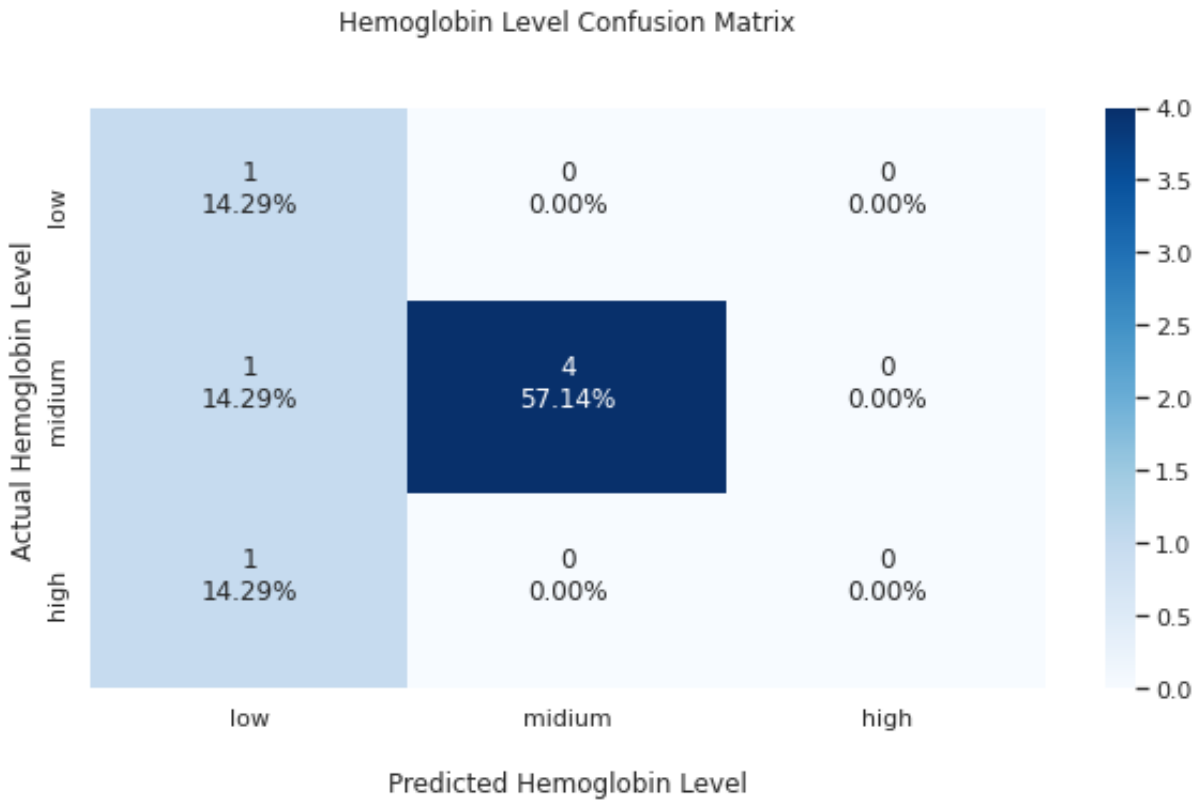


Figure 31 A: Confusion matrix for the prediction vs actual values of hemoglobin levels

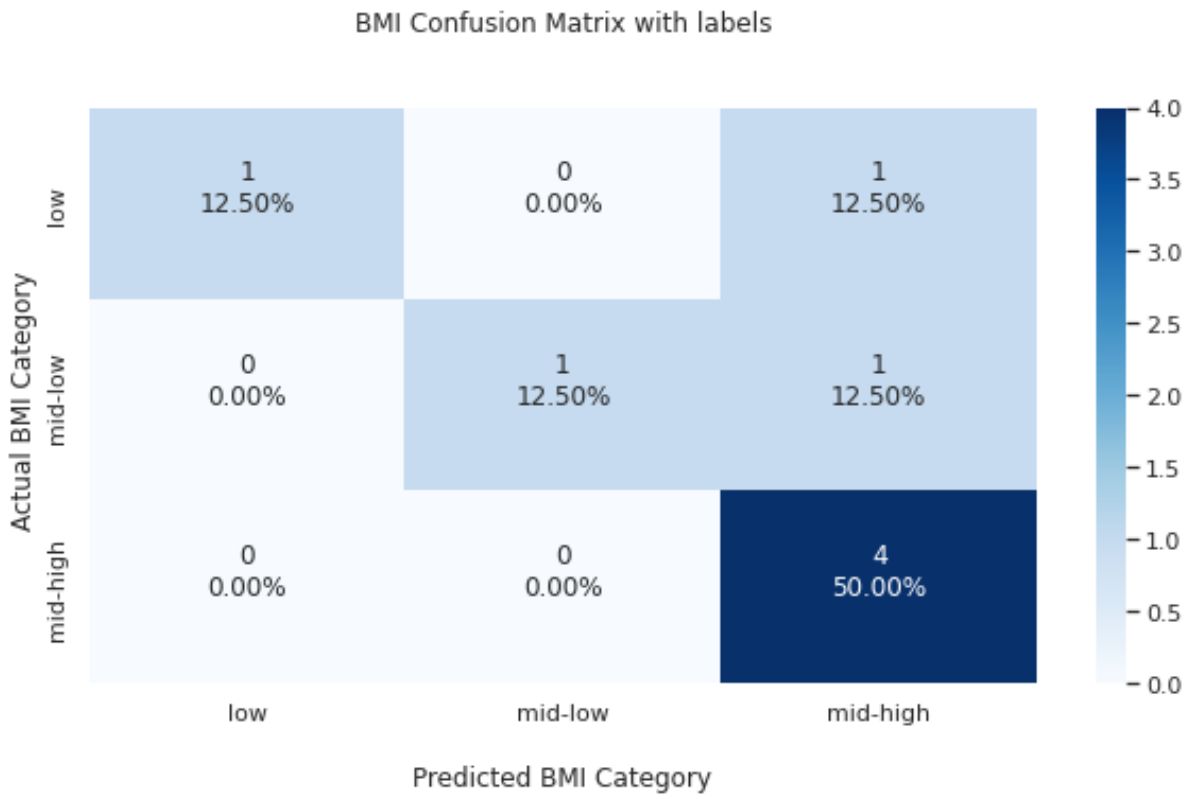


Figure 31 B: Confusion matrix for the prediction vs actual values of BMI levels

Concerning the age decision tree results: UNC12 C, C2orf88, and SLIT2, were the variables of highest importance, the genes that have the highest ability of predicting the age group. If you sequence people for UNC13C and normalize the expression, and then take the patients that have less than 9.464 normalized expressions, you are 100% predicting their age correctly. Inside every box there are three rows, the first row, the second row, and the third row. The first row represents the predicted class (e.g. young, middle, old). The second row shows you the probability of each class (the probability of being young, the probability of being middle aged, and the probability of being old.) The second row always adds to one. The third row will present the percentage of the total population that is inside the box (e.g. we start at 100% population, but then as we go down we have 36 and 64) and so on. The idea is that if the UNC13 is less than 9.4, then you predict them to be young, that is the most significant variable. If UNC13 is more than 9.4 you go down to the right. Next, if C2rof88 is greater than or equal to 0.51 then you go to the left and predict young again if they are less than 10.51 you go to SLIT2, and so forth. Concerning the decision tree for the second dataset, the variable of highest importance was AI424825. The first dataset had 92% accuracy and the second dataset had 95% accuracy.

Neural Networks Results

Concerning the Neural Networks, as seen in Figures 32-34 the more you increase the threshold, the higher the error and the less the steps. This same trend is observed in the error for Dataset 2, as displayed in Figures 35-37. Whenever we are changing the thresholds, we are affecting the errors and the steps. The networks displayed below depict the impact the top 10 age-correlated genes have to cluster both datasets into three clusters and all the way on the right side we see the labels (young age, middle age, old age). Inputs are gene expression and outputs are age groups.

Results

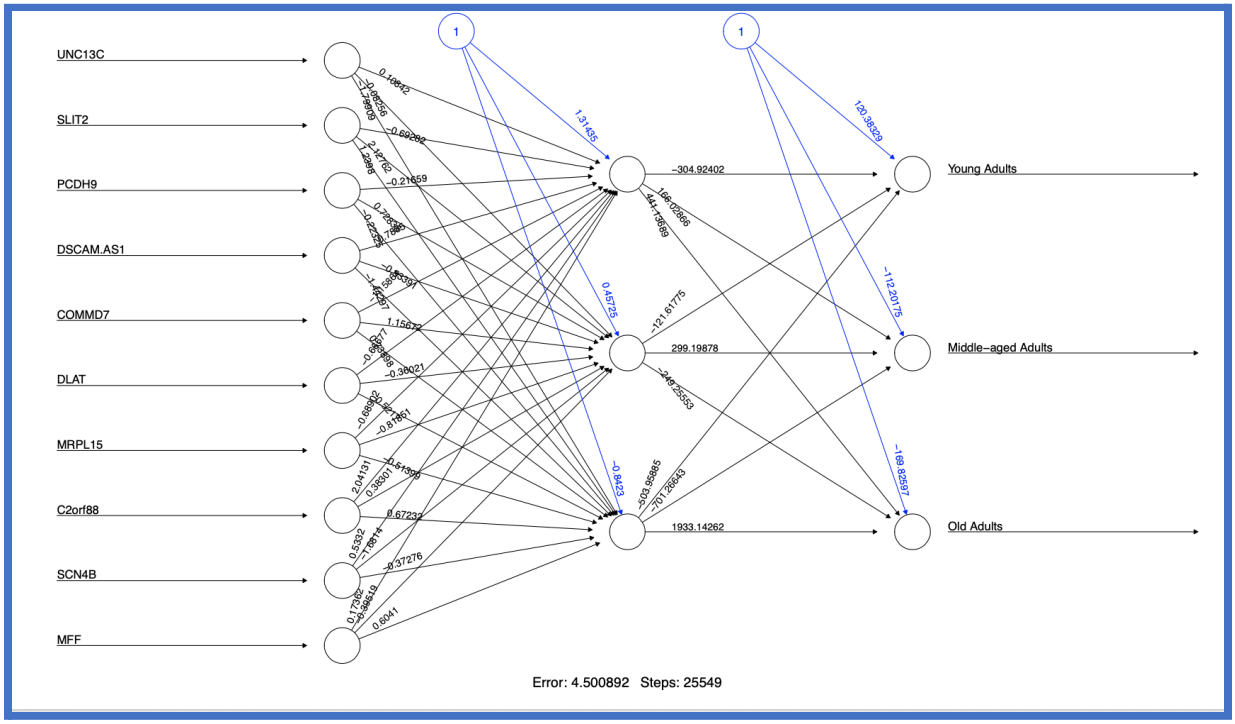


Figure 32: Neural Network Dataset 1 Threshold 0.01

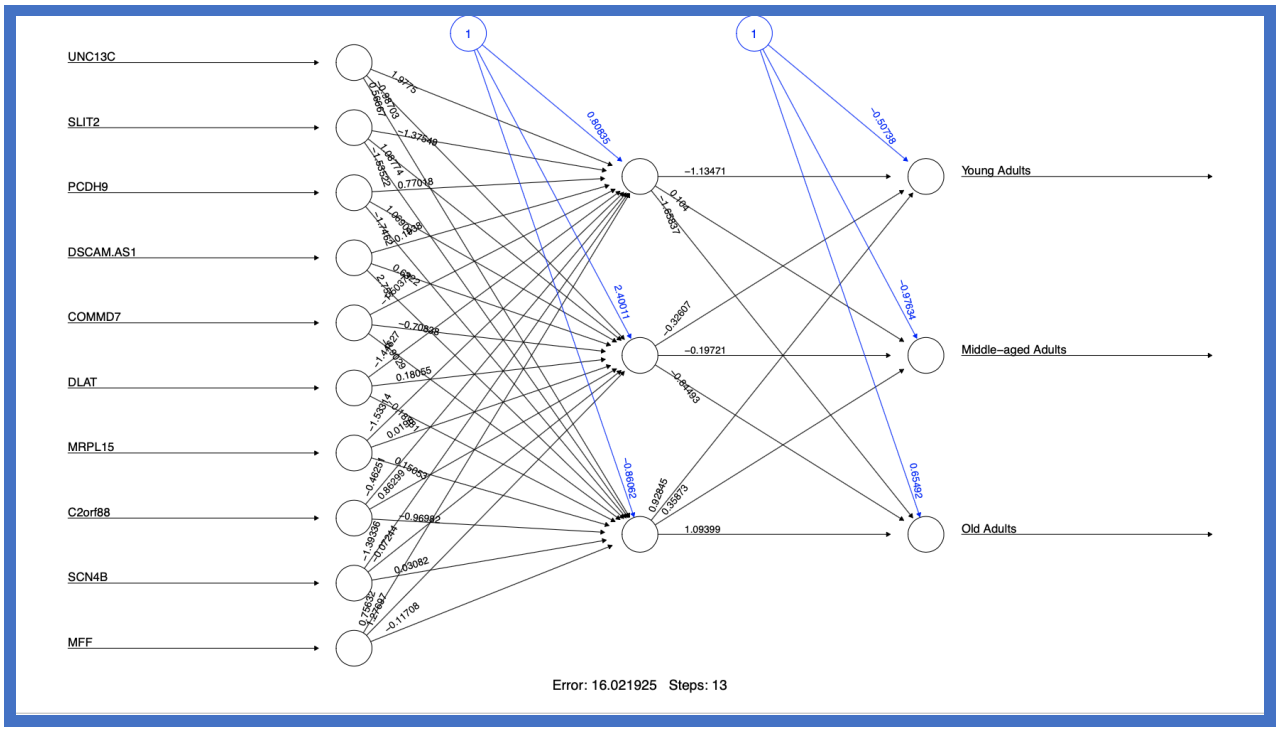


Figure 33: Neural Network Dataset 1 Threshold 0.1

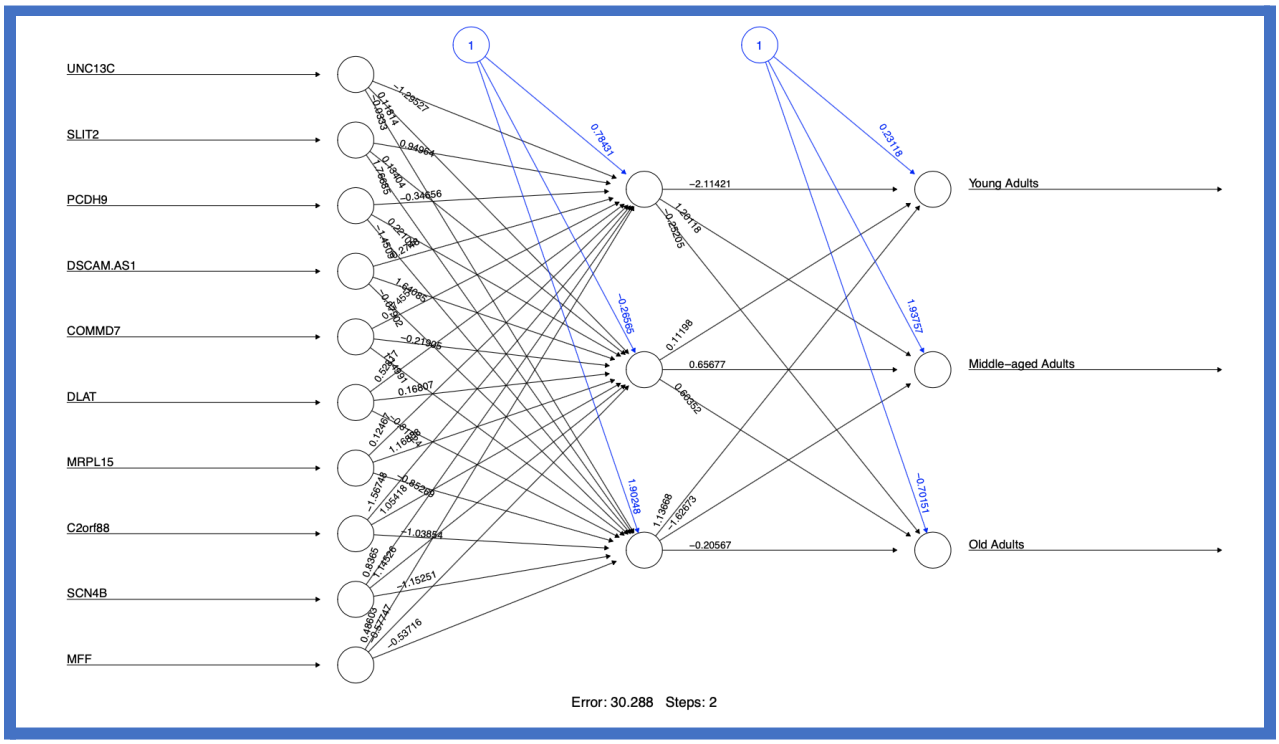


Figure 34: Neural Network Dataset 1 Threshold 1

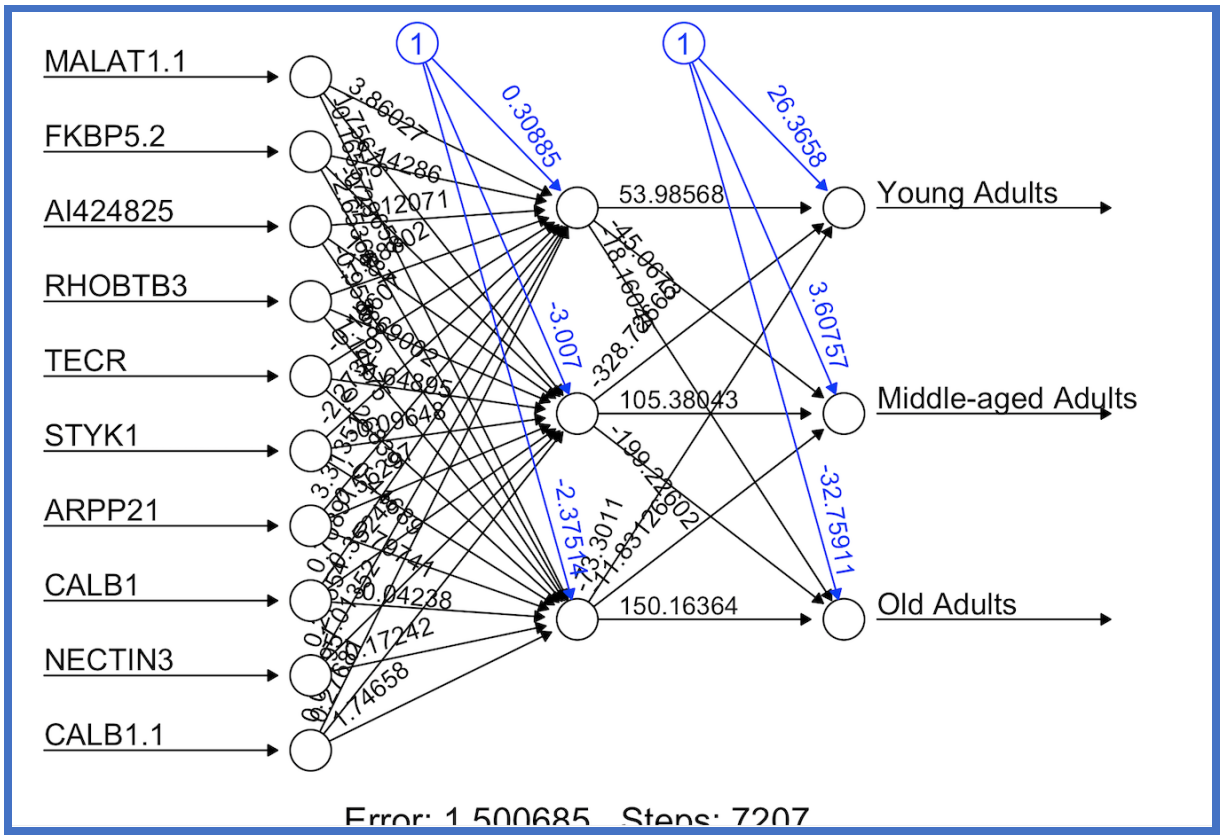


Figure 35: Neural Network Dataset 2 Threshold 0.01

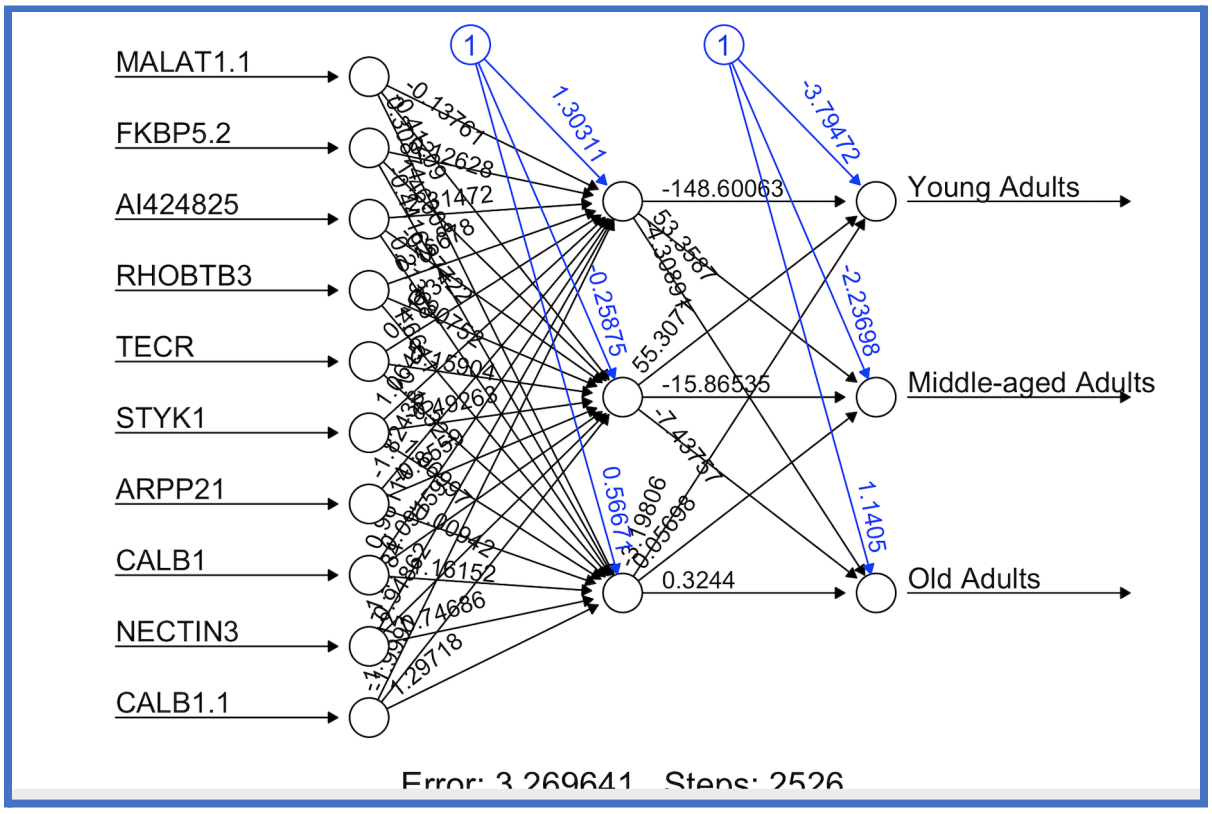


Figure 36: Neural Network Dataset 2 Threshold 0.1

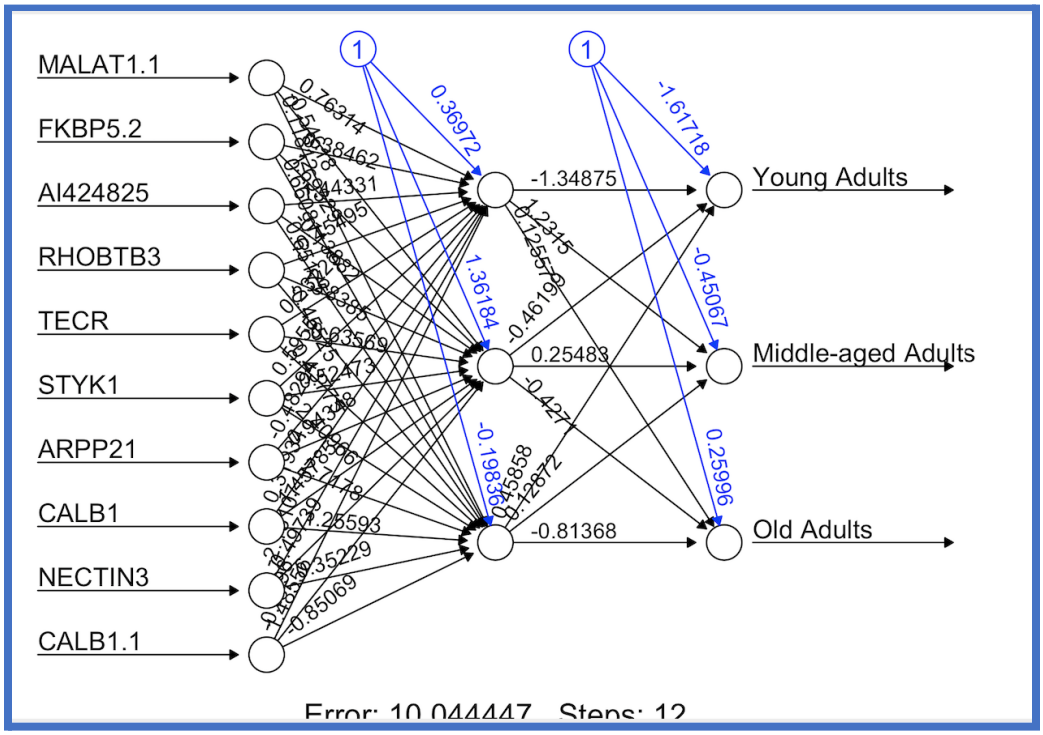


Figure 37: Neural Network Dataset 2 Threshold 1

Pearson and Spearman Results

The correlation matrix plot can be seen in **Figure 38A & 38B** for the two datasets of diabetes and brain aging respectively. Looking at the colors in the plot it was observed that almost all of the coefficients were positive and of a high value of around and above 0.8 for diabetes and above 0.9 for brain aging. Furthermore two plots were plotted for brain aging since the two methods of correlation showed different values. As seen from the plots it can be concluded from the colors that although the values are different they are quite comparable and have almost no deviations between them.

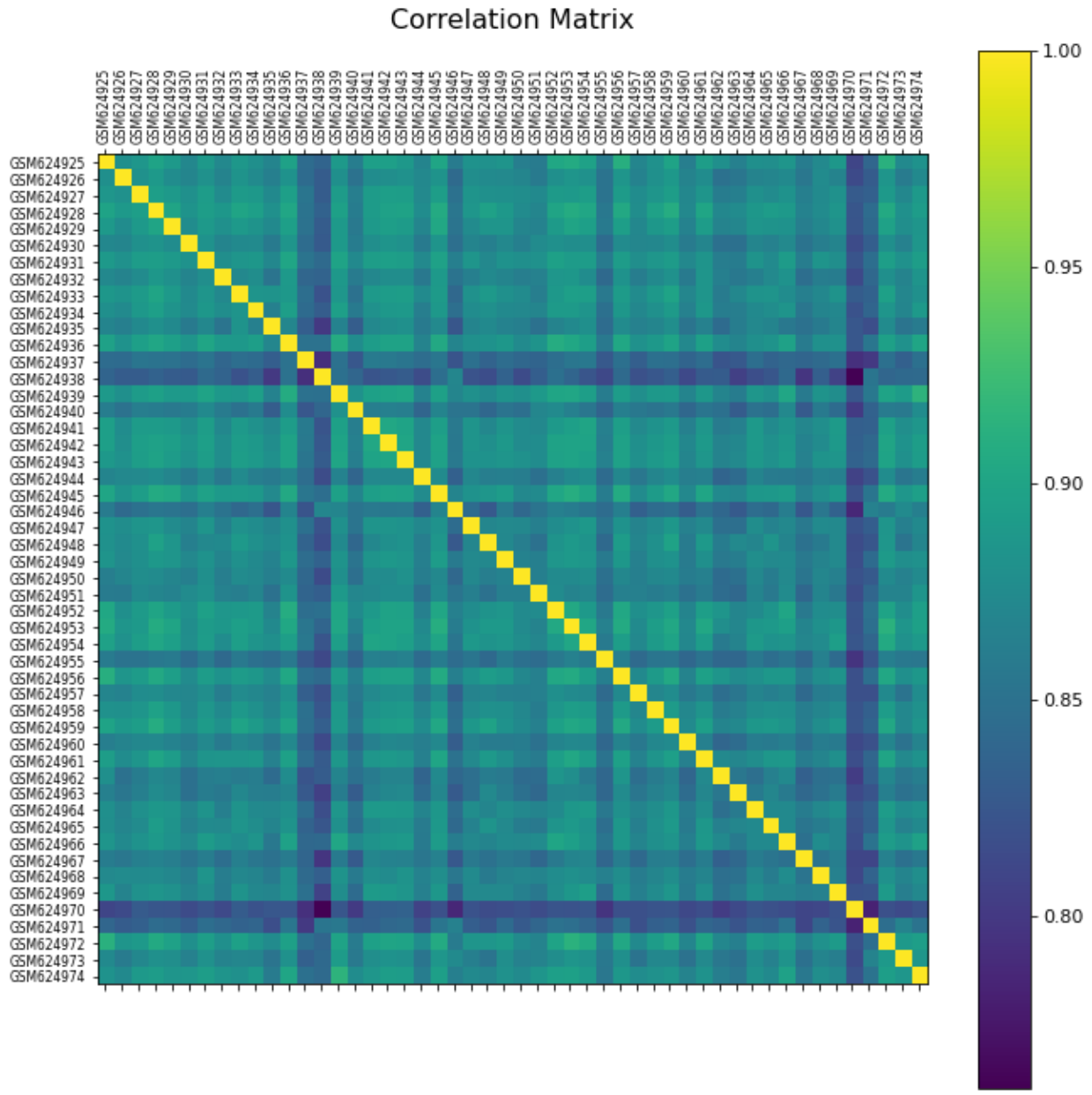


Figure 38A: Correlation Matrix plot for human gene expression data of diabetes in muscles.

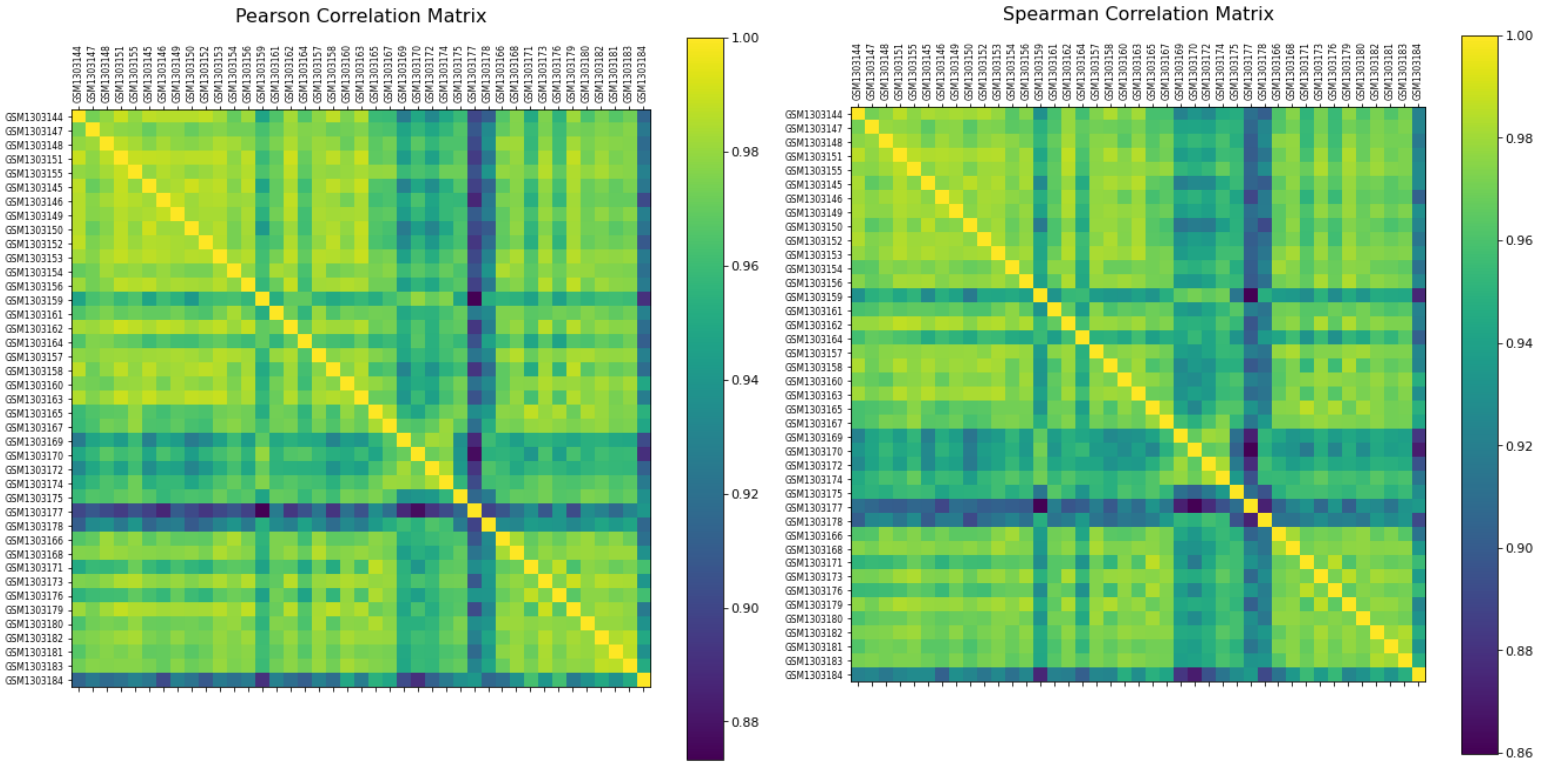


Figure 38B: Correlation Matrix plot for human gene expression data of brain aging.

A series of images depicting the networks in the correlation matrix for different values of coefficient threshold can be seen in **Figure 4 (A, B, C, D, and E) & Figure 6 (A, B, C, and D) in the Appendix** for the two datasets of diabetes and brain aging respectively. Figure 3 and Figure 5 shows the networks for all the values in the coefficient matrix for both the datasets.

The following series of images is a plot of the degree distribution for each of the threshold values taken above. **(Figure 39 & Figure 40 for two datasets respectively)** It can be observed that for higher values of threshold more than 50% of nodes have degree less than 5.

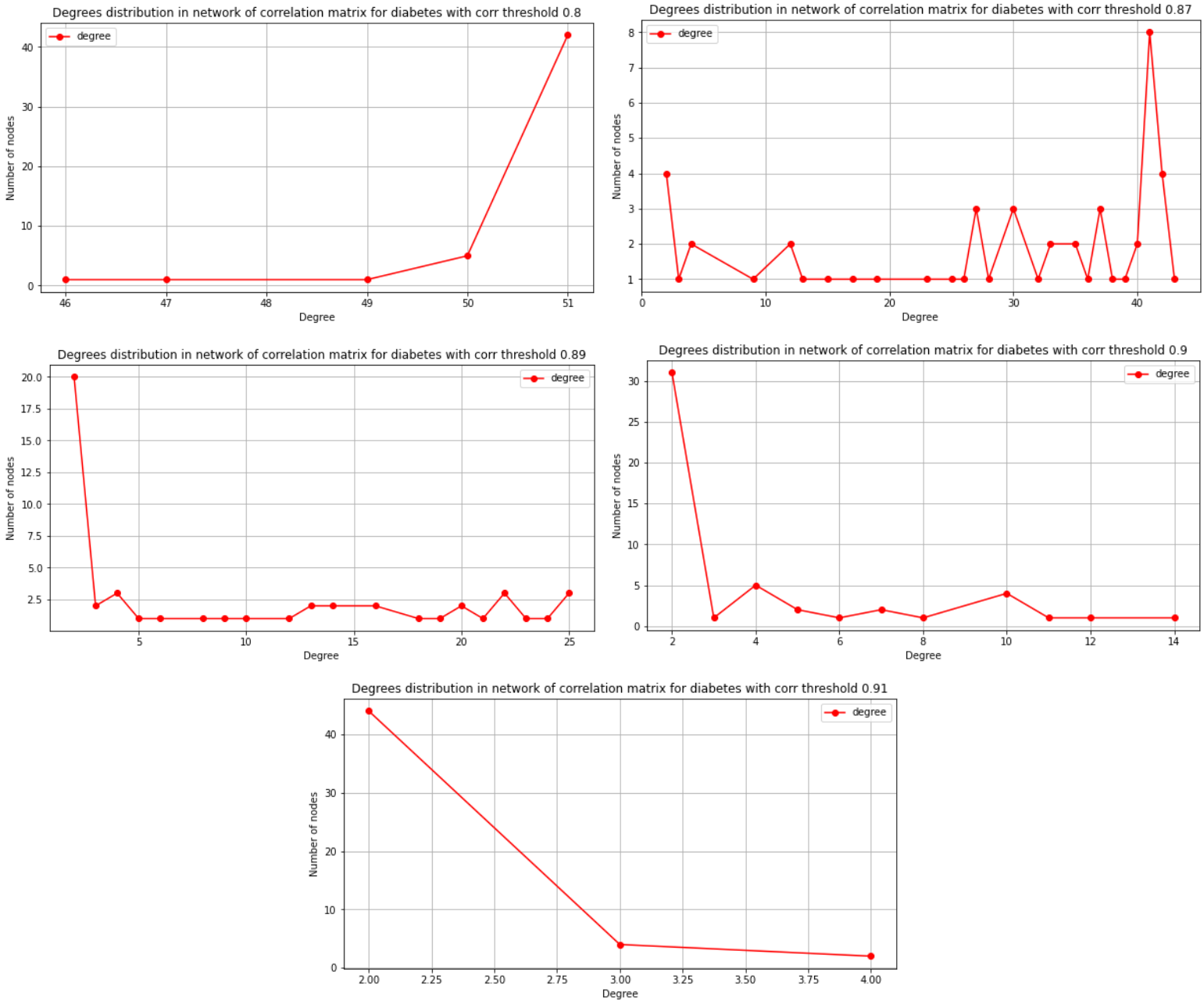


Figure 39: Distribution of degree of correlation for threshold values of 0.8, 0.87, 0.89, 0.9, and 0.91 for the diabetes dataset.

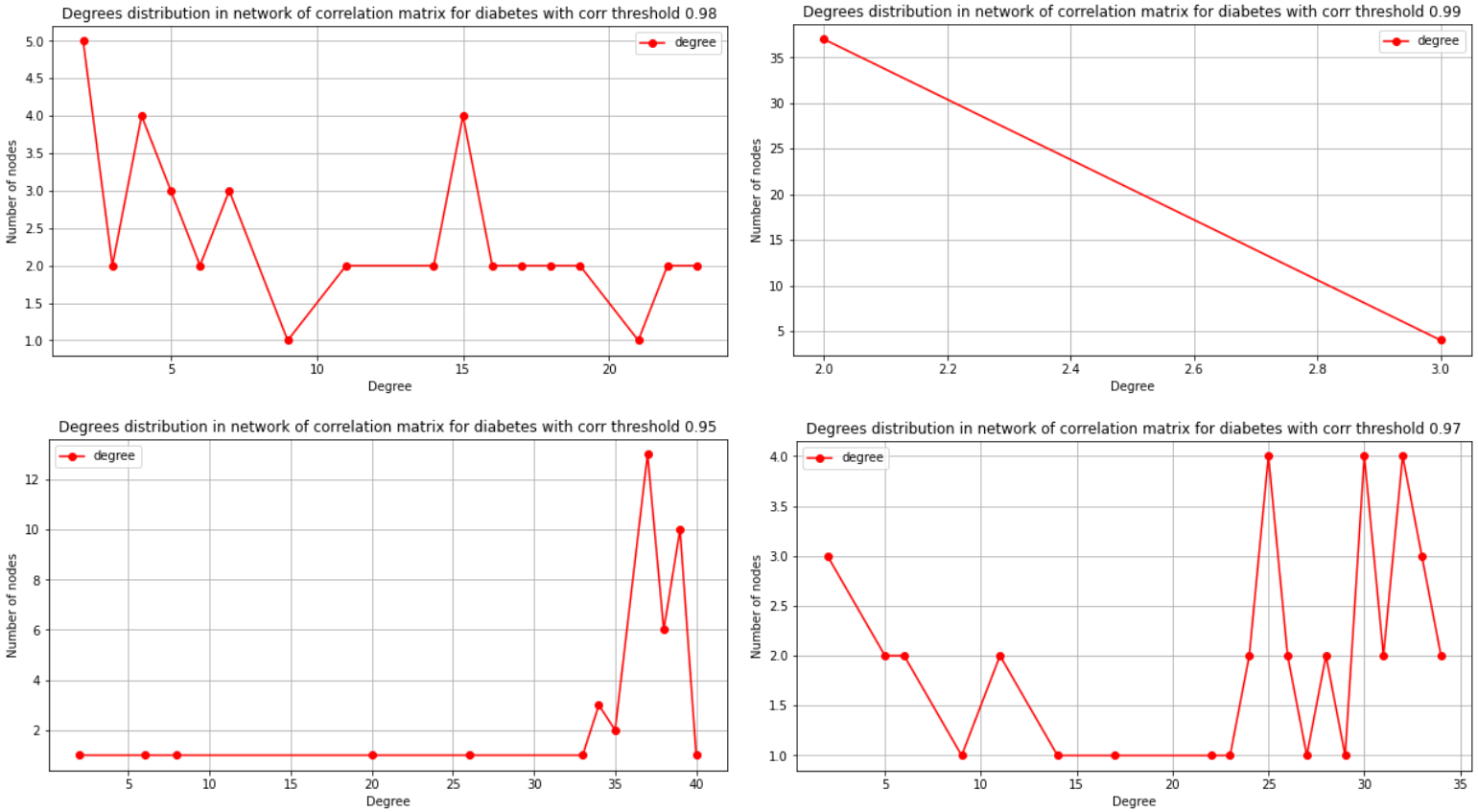


Figure 40: Distribution of degree of correlation for threshold values of 0.95, 0.97, 0.98 and 0.99 for the brain aging dataset.

DISCUSSION

Table Summary and Phenotype Clustering Discussion

For the Human Data for Insulin Resistance and Diabetes in Muscle data frame, it was not expected for male and female patients to cluster as gender was found to be an insignificant predictor in both datasets. This was confirmed and visually represented in Figures 26 and 27; both clusters overlap. As expected, age correlates with the clusters in the second dataset, as age was previously found to be a significant phenotype in the second dataset.

Decision Tree Discussion

Concerning the decision trees for hemoglobin and BMI, the prediction accuracy obtained is not a high value which can have multiple explanations. The sample size of 50 is small and leaves less than 10 data points to test the model. There were no genes that showed high positive or negative correlation with the absolute maximum value going up to only 0.63. Figure 6 (Supplementary file) shows the correlation distribution of the genes with respect to all

phenotypes with numerical values. None of them portrayed correlation values beyond 0.7. Overall this method to find the highly correlated genes and using them to train the gene expression data of the top correlated genes to predict the phenotypes of a test data can be considered a good method to understand the contribution of gene expression towards certain phenotypes which are related to biological diseases. Concerning the decision trees for age, the prediction accuracy for the first dataset was 92%, and the prediction accuracy for the second dataset was 95%. It makes sense that the prediction accuracy for the second dataset is higher compared to the first dataset, as the second dataset has to do with brain aging.

Neural Networks Discussion

It was expected for the errors in dataset 2 (brain aging dataset) to be lower than in dataset1; that the gene expression in the second disease model is more correlated with age compared to the first model. Furthermore, as previously stated, the more you increase the threshold, the higher the error and the less the steps. In other words, the higher the threshold, the lower the steps and the higher the error. By definition, the threshold is the partial derivative of the error function. Inputting a high threshold would be equivalent to decreasing the model accuracy. As the top 5 positively and negatively correlated genes with age (10 most significant genes) were utilized for these models, huge differences in error when altering the threshold were not observed. If one instead used genes that were not very correlated or were quite insignificant as model inputs, one would expect large error differences corresponding to the threshold changes.

Pearson and Spearman Discussion

The correlation matrix for both the datasets showed positive values and close to 1 indicating that the genes are highly correlated among themselves.

It is also observed that as we increase the threshold of correlation the number of connected nodes for each node which is indicated by the degree is distributed highly on the lower side meaning all nodes are connected to very few nodes indicating that strong correlation exists only among a few genes. A possible reason could be that they share the same pathway.

DAVID Analysis Discussion (Part I)

In the second Dataframe, we were able to identify gene signatures that correlated with age that suggested that that disease model is reliant on age and progressively changes the gene expression as the patients get older. Thus, age itself could be predictive of prognosis or outcome of these specific patients. The gene expression in the second disease model was more correlated with age compared to the first model which makes sense because the second disease model is studying the brain while the first one is studying diabetes. The changes in gene expression are more obvious or more correlated with age with the brain disease model.

DAVID Results and Discussion (Insulin Resistance)

Insulin resistance is a detectable metabolic defect in humans with T2D that is present in diabetic parents' offspring and predicts incident diabetes. Insulin signaling changes, mitochondrial dysfunction, endoplasmic reticulum stress, oxidative stress, and inflammation have all been linked to the development of cellular insulin resistance. While insulin resistance is characterized by decreased glucose uptake into skeletal muscle, the underlying molecular defects that mediate insulin resistance and diabetes risk in humans are unknown. Over 30 polymorphic loci have been identified in individuals with T2D by genome-wide association studies, but identification of genetic markers of the insulin-resistant state has proven elusive thus far.

Interestingly, striated muscle activator of Rho signaling (STARS; official gene name **ABRA**, or actin-binding Rho activating protein), a known activator of SRF transcriptional activity, was the top-ranking differentially expressed gene. Increased STARS expression may contribute to increased SRF target gene expression in insulin resistance because STARS is known to activate SRF via actin-dependent induction of nuclear accumulation of MKL1.

AGE/RAGE signaling has been extensively researched in a variety of disease states, particularly diabetes. The AGE/RAGE signaling mechanism is still poorly understood due to the complex nature of the receptor and multiple intersecting pathways [3]. The goal of this review is to highlight key aspects of AGE/RAGE-mediated vascular calcification as a diabetes complication. In both hyperglycemic and calcification conditions, AGE/RAGE signaling heavily influences both cellular and systemic responses to increase bone matrix proteins via PKC, p38 MAPK, fetuin-A, TGF-, NFB, and ERK1/2 signaling pathways.

Heart failure and related morbidity and mortality are on the rise, owing in large part to increases in aging, obesity, and diabetes mellitus. Clinical outcomes for patients with diabetes mellitus are significantly worse than those for those without diabetes mellitus. **Diabetic cardiomyopathy** refers to the presence of myocardial dysfunction in the absence of overt clinical coronary artery disease, valvular disease, and other conventional cardiovascular risk factors such as hypertension and dyslipidemia in people with diabetes mellitus.

DAVID Results and Discussion (Alzheimer's Disease)

Abnormalities in cellular bioenergetics have been found in patients with Alzheimer's disease (AD) and other neurodegenerative diseases. The pyruvate dehydrogenase complex, alpha-ketoglutarate dehydrogenase complex, and oxidative phosphorylation are the most commonly reported enzyme abnormalities (OXPHOS). Although genetic evidence for primary OXPHOS defects as a cause of AD is lacking, functionally significant reductions in OXPHOS enzyme activities appear to occur in AD and may be linked to beta-amyloid accumulation or

other neurodegenerative processes. OXPHOS defects may play an important role in the pathophysiology of AD because decreased neuronal ATP may increase susceptibility to glutamate toxicity. [4]

APPENDIX

ID	adj.P.Val	P.Value	F	Gene.symbol	Gene.title
1552731_at	0.000194	3.55e-09	29.45	ABRA	actin binding Rho activating protein
1568751_at	0.045533	2.31e-06	17.02	RGS13	regulator of G-protein signaling 13
1552732_at	0.045533	2.50e-06	16.89	ABRA	actin binding Rho activating protein
240244_at	0.054937	4.02e-06	16.1		
223973_at	0.082145	7.51e-06	15.08	MIR7-3HG	MIR7-3 host gene
207914_x_at	0.109733	1.20e-05	14.33	EVX1	even-skipped homeobox 1
231035_s_at	0.211407	2.88e-05	12.98	OTUD1	OTU deubiquitinase 1
242329_at	0.211407	3.50e-05	12.69	LOC401317///CREB5	uncharacterized LOC401317///cAMP …
224108_at	0.211407	3.78e-05	12.57		
1560069_at	0.211407	3.87e-05	12.54	PLEKHM3	pleckstrin homology domain containi…
226140_s_at	0.233791	4.70e-05	12.24	OTUD1	OTU deubiquitinase 1
219734_at	0.256798	6.08e-05	11.86	SIDT1	SID1 transmembrane family member 1
212779_at	0.256798	6.11e-05	11.86	KIAA1109	KIAA1109
211062_s_at	0.258876	6.63e-05	11.74	GPR78///CPZ	G protein-coupled receptor 78///carb…
211674_x_at	0.292226	8.21e-05	11.42	CTAG1A///CTAG1B	cancer/testis antigen 1A///cancer/test…
206039_at	0.292226	8.55e-05	11.36	RAB33A	RAB33A, member RAS oncogene fa…
200974_at	0.332524	1.05e-04	11.06	ACTA2	actin, alpha 2, smooth muscle, aorta
202149_at	0.332524	1.09e-04	11.01	NEDD9	neural precursor cell expressed, dev…
1553298_at	0.348793	1.34e-04	10.72	C17orf77	chromosome 17 open reading frame 77
239027_at	0.348793	1.37e-04	10.69	DOCK8	dedicator of cytokinesis 8
223486_at	0.348793	1.39e-04	10.67	GTPBP8	GTP binding protein 8 (putative)
209081_s_at	0.348793	1.40e-04	10.65	COL18A1	collagen type XVIII alpha 1 chain

Table 1: Table showing top differentially expressed (significant) genes for GSE25462

ID	adj.P.Val	P.Value	F	Gene.symbol	Gene.title
205626_s_at	0.00000259	4.73e-11	33.4	CALB1	calbindin 1
205625_s_at	0.00000276	1.01e-10	31.7	CALB1	calbindin 1
1552722_at	0.00000278	1.53e-10	30.8	ARPP21	cAMP regulated phosphoprotein 21
241672_at	0.00000489	3.58e-10	28.9	SERTM1	serine rich and transmembrane dom…
220030_at	0.00000523	5.82e-10	27.9	STYK1	serine/threonine/tyrosine kinase 1
213325_at	0.00000523	7.10e-10	27.5	NECTIN3	nectin cell adhesion molecule 3
218026_at	0.00000523	7.14e-10	27.5	COA3	cytochrome c oxidase assembly fact…
222774_s_at	0.00000523	7.65e-10	27.3	NETO2	neuropilin and tolloid like 2
222231_s_at	0.00000527	8.67e-10	27.1	LRRC59	leucine rich repeat containing 59
210675_s_at	0.00000561	1.03e-09	26.7	PTPRR	protein tyrosine phosphatase, recept…
224587_at	0.00000824	1.66e-09	25.8	SUB1	SUB1 homolog, transcriptional regula…
213219_at	0.00000835	1.83e-09	25.6	ADCY2	adenylate cyclase 2
220359_s_at	0.00000878	2.46e-09	25	ARPP21	cAMP regulated phosphoprotein 21
242825_at	0.00000878	2.54e-09	25	PLPPR5	phospholipid phosphatase related 5
235591_at	0.00000878	2.69e-09	24.9	SSTR1	somatostatin receptor 1
216307_at	0.00000878	2.71e-09	24.8	DGKB	diacylglycerol kinase beta
1553864_at	0.00000878	2.87e-09	24.7	GPR26	G protein-coupled receptor 26
221696_s_at	0.00000878	2.89e-09	24.7	STYK1	serine/threonine/tyrosine kinase 1
220889_s_at	0.00001056	3.67e-09	24.3	CA10	carbonic anhydrase 10
232275_s_at	0.00001126	4.12e-09	24.1	HS6ST3	heparan sulfate 6-O-sulfotransferase 3
218888_s_at	0.00001185	4.73e-09	23.8	NETO2	neuropilin and tolloid like 2
223550_s_at	0.00001185	4.77e-09	23.8	CA10	carbonic anhydrase 10
214553_s_at	0.00001215	5.23e-09	23.6	ARPP19	cAMP regulated phosphoprotein 19
1560587_s_at	0.00001215	5.34e-09	23.6	PRDX5	peroxiredoxin 5

Table 2: Table showing top differentially expressed (significant) genes for GSE53890

GEO2R Results: Insulin resistance

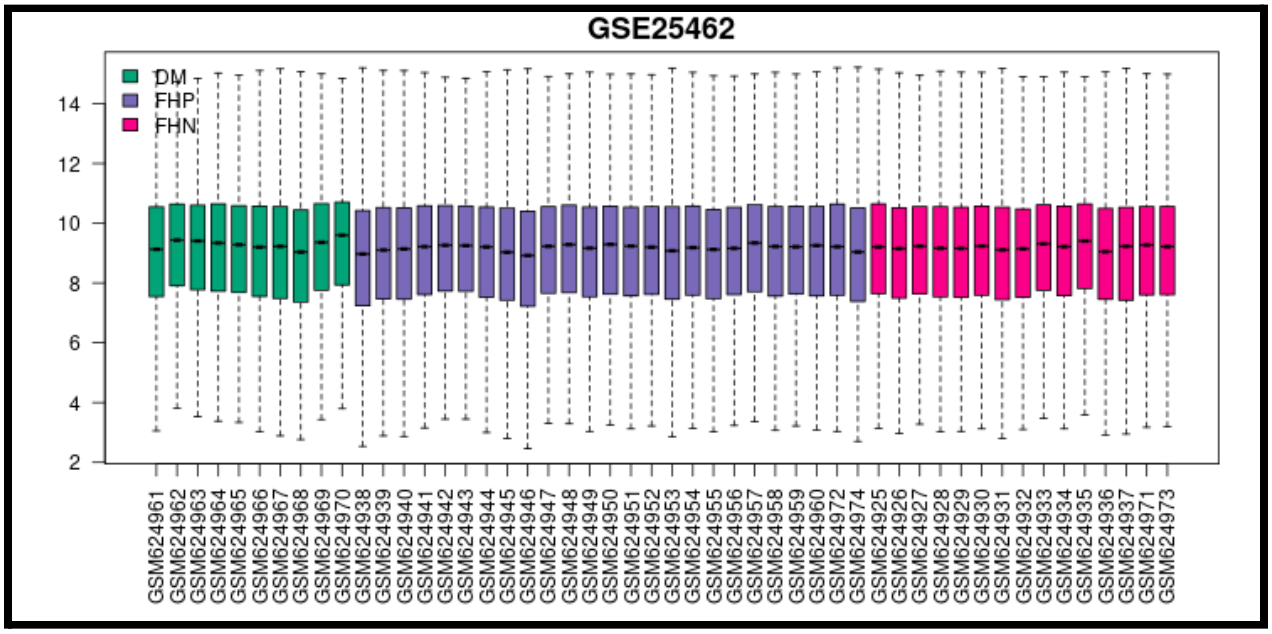
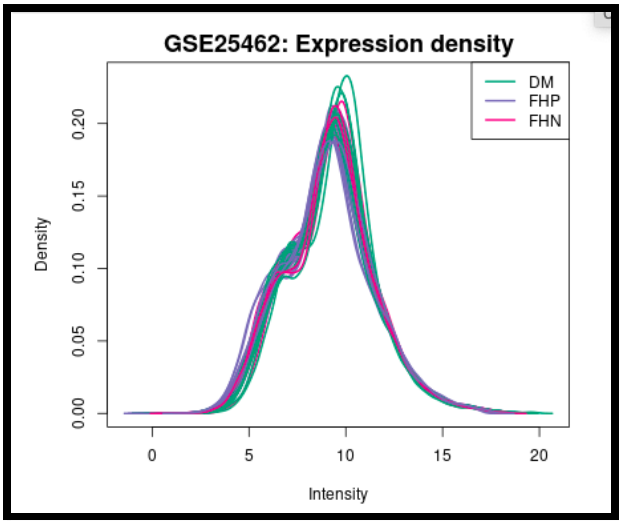


Figure 1 (a): Boxplot showing normalized samples (10 subjects with type 2 diabetes, 25 subjects with a family history of type 2 diabetes (one or both parents), and 15 subjects with no family history of type 2 diabetes)



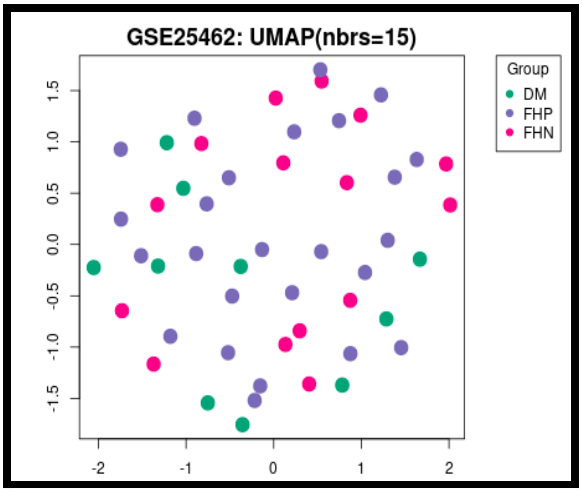


Figure 1 (b): Expression density plot

Figure 1(c): UMAP Plot

GEO2R Results: Brain

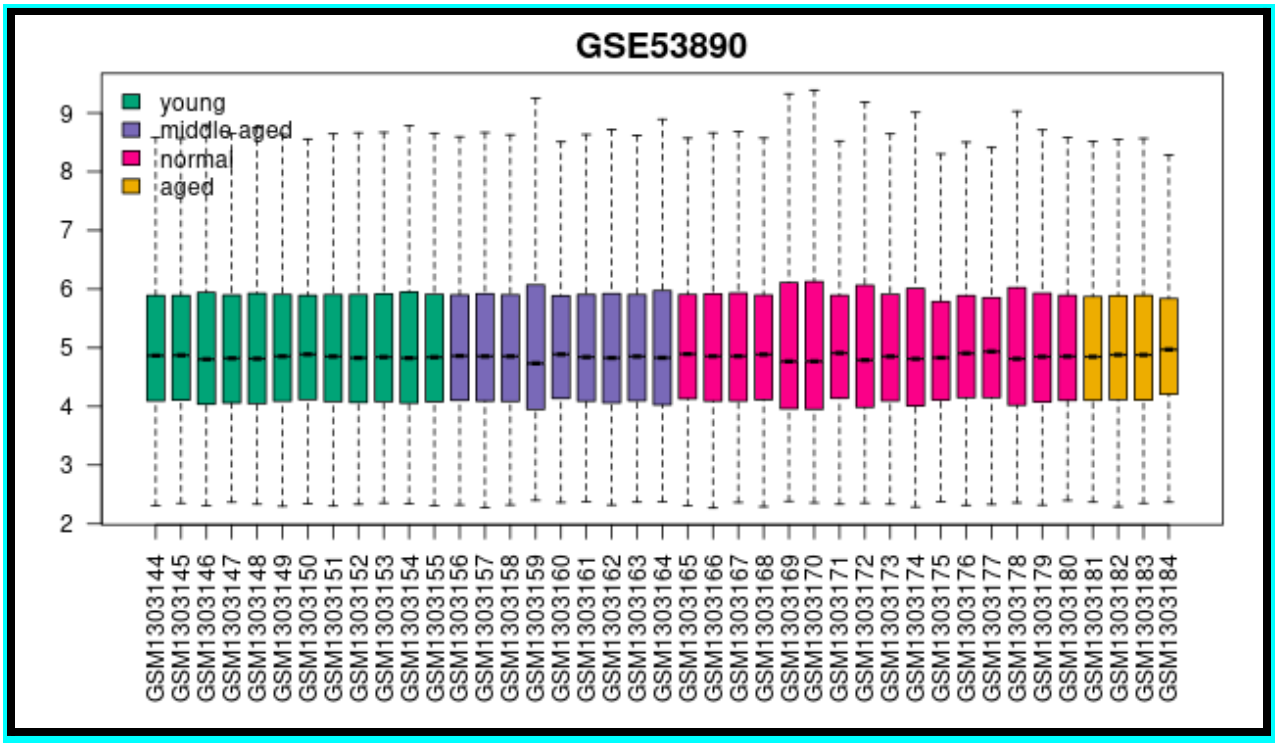
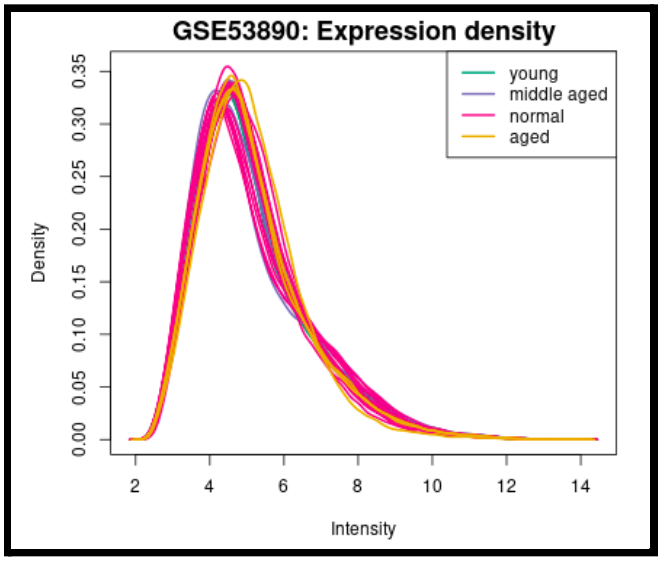


Figure 2 (a): Boxplot showing normalized samples (Comparison of 12 young (<40yr), 9 middle aged (40-70yr), 16 normal aged (70-94yr), and 4 extremely aged (95-106yr))



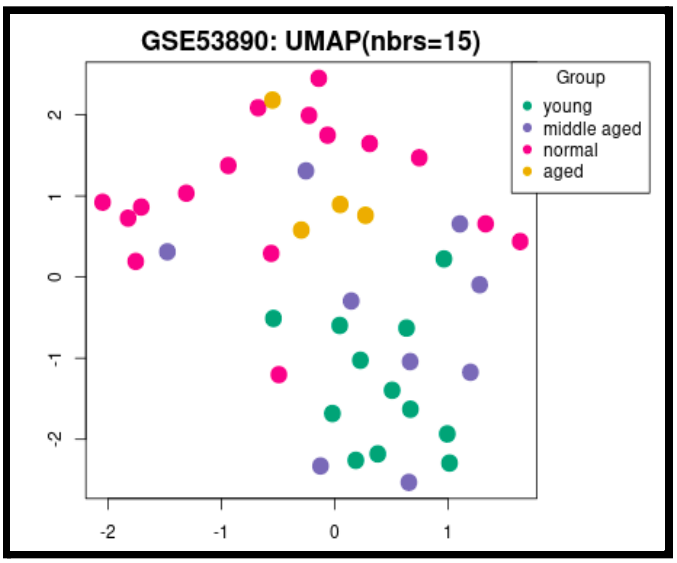
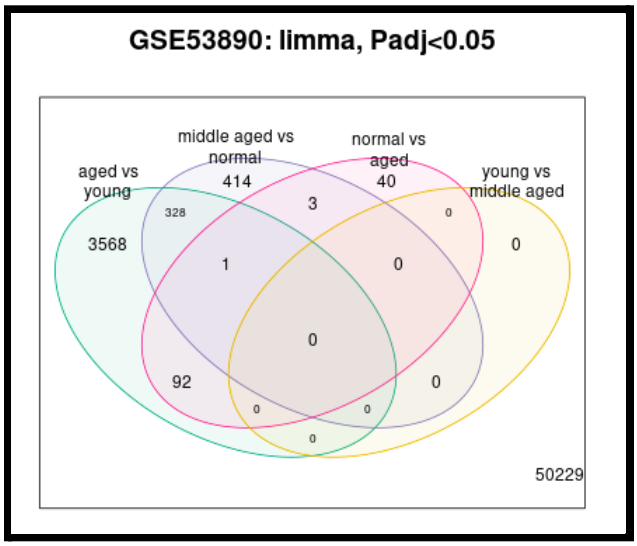


Figure 2(b): Expression density plot

Figure 2(c): UMAP Plot



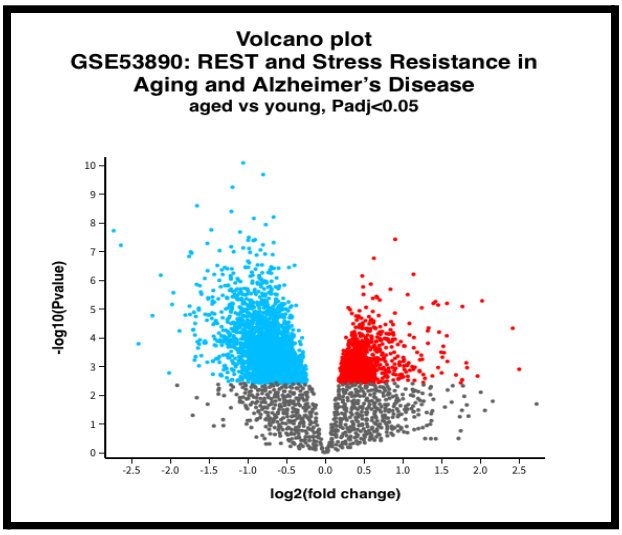


Figure 2(d): Venn diagram

Figure 2(e): Volcano Plot (Downloaded significant genes)

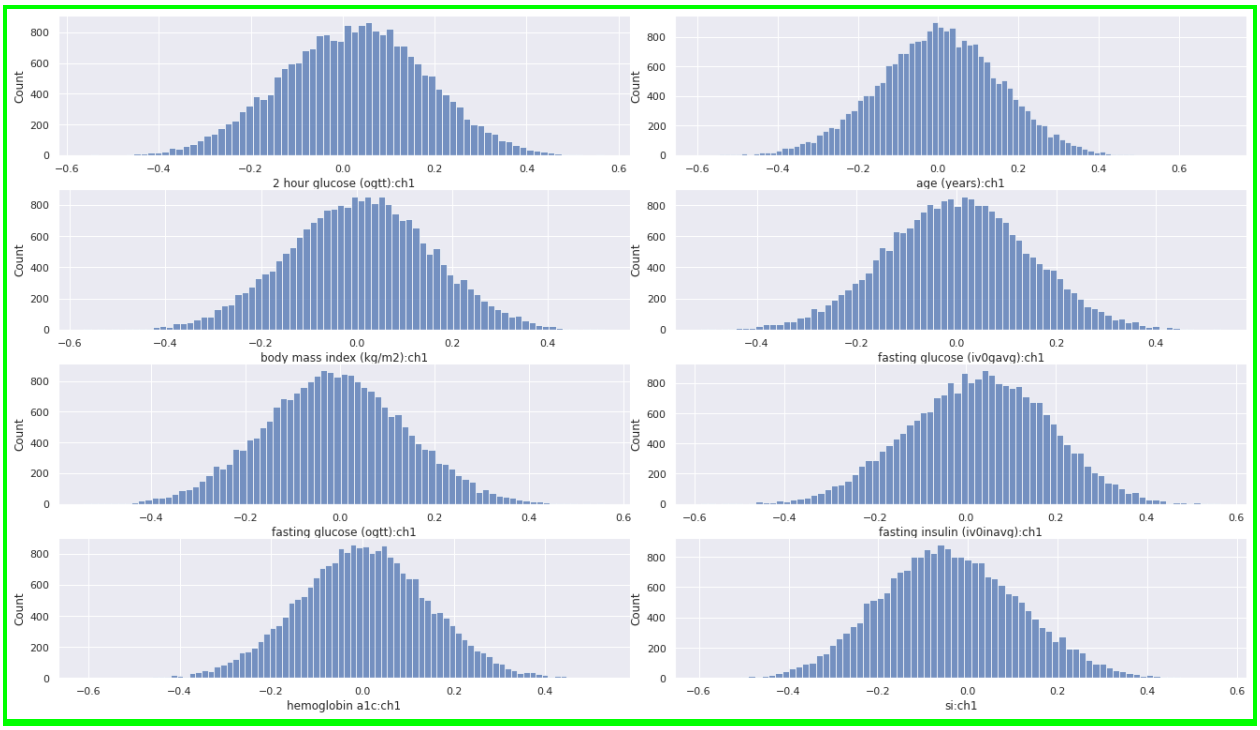


Figure 3: Histogram plots of counts of genes and the correlation values for correlation with eight different phenotypes – 2hr glucose, age, BMI, fasting glucose (iv0gavg), fasting glucose (ogtt), fasting insulin, hemoglobin, si:chi1

Pearson and Spearman Analysis Table Summary

	GSM624925	GSM624926	GSM624927	GSM624928	GSM624929	GSM624930	GSM624931	GSM624932	GSM624933	GSM624934	...
GSM624925	1.000000	0.879457	0.886103	0.898879	0.888601	0.870685	0.889174	0.872102	0.886620	0.879316	...
GSM624926	0.879457	1.000000	0.880947	0.889163	0.876872	0.869019	0.881717	0.867194	0.883715	0.874378	...
GSM624927	0.886103	0.880947	1.000000	0.894706	0.889969	0.880234	0.893109	0.875683	0.890796	0.888300	...
GSM624928	0.898879	0.889163	0.894706	1.000000	0.899736	0.883693	0.893220	0.886936	0.899955	0.895255	...
GSM624929	0.888601	0.876872	0.889969	0.899736	1.000000	0.876325	0.889117	0.880906	0.889211	0.883568	...

5 rows × 50 columns

```
diab_corr_s.head()
```

	GSM624925	GSM624926	GSM624927	GSM624928	GSM624929	GSM624930	GSM624931	GSM624932	GSM624933	GSM624934	...
GSM624925	1.000000	0.872757	0.879696	0.895001	0.883638	0.862210	0.884661	0.863938	0.879555	0.873313	...
GSM624926	0.872757	1.000000	0.867436	0.878375	0.863265	0.854306	0.869780	0.852049	0.868464	0.861545	...
GSM624927	0.879696	0.867436	1.000000	0.886278	0.878600	0.868694	0.883100	0.862494	0.878636	0.876848	...
GSM624928	0.895001	0.878375	0.886278	1.000000	0.890659	0.873364	0.884834	0.877021	0.891099	0.887230	...
GSM624929	0.883638	0.863265	0.878600	0.890659	1.000000	0.862625	0.878704	0.867136	0.876050	0.872145	...

5 rows × 50 columns

Figure 4 A: Pearson Correlation matrix (top) and Spearman Correlation matrix (bottom) for diabetes in muscle

	GSM1303144	GSM1303147	GSM1303148	GSM1303151	GSM1303155	GSM1303145	GSM1303146	GSM1303149	GSM1303150	GSM1303152	...
GSM1303144	1.000000	0.973389	0.979649	0.989122	0.979572	0.987534	0.985739	0.986037	0.986056	0.987668	...
GSM1303147	0.973389	1.000000	0.981695	0.978673	0.977545	0.970411	0.977313	0.977044	0.970797	0.974974	...
GSM1303148	0.979649	0.981695	1.000000	0.986755	0.983167	0.976872	0.983122	0.978612	0.976841	0.979308	...
GSM1303151	0.989122	0.978673	0.986755	1.000000	0.988704	0.988294	0.987904	0.986002	0.987090	0.987054	...
GSM1303155	0.979572	0.977545	0.983167	0.988704	1.000000	0.979402	0.985219	0.982783	0.978442	0.981068	...

5 rows × 41 columns

	GSM1303144	GSM1303147	GSM1303148	GSM1303151	GSM1303155	GSM1303145	GSM1303146	GSM1303149	GSM1303150	GSM1303152	...
GSM1303144	1.000000	0.970895	0.975567	0.985326	0.976211	0.982318	0.979908	0.981724	0.980054	0.983512	...
GSM1303147	0.970895	1.000000	0.978323	0.976637	0.975194	0.965461	0.974999	0.975413	0.963860	0.973361	...
GSM1303148	0.975567	0.978323	1.000000	0.983506	0.980129	0.970728	0.980181	0.975745	0.968905	0.975854	...
GSM1303151	0.985326	0.976637	0.983506	1.000000	0.985692	0.983094	0.983876	0.982712	0.980427	0.983605	...
GSM1303155	0.976211	0.975194	0.980129	0.985692	1.000000	0.975126	0.981837	0.980666	0.971581	0.979282	...

5 rows × 41 columns

Figure 4 B: Pearson Correlation matrix (top) and Spearman Correlation matrix (bottom) for brain aging

Figure 5: Network diagram of the correlation matrix of diabetes in muscle data

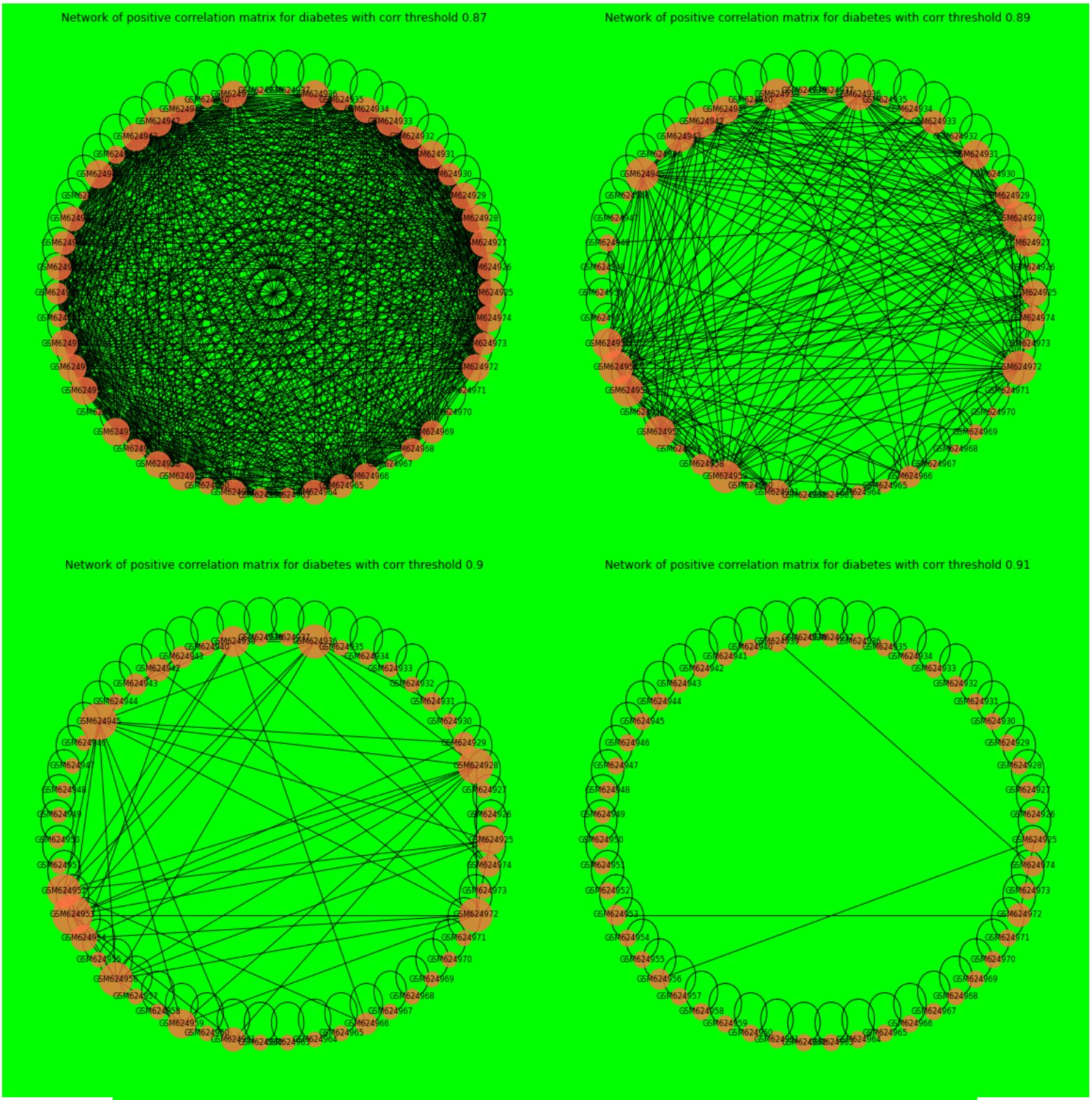


Figure 6: Network of correlation matrix for threshold values of 0.8, 0.87, 0.89, 0.9, and 0.91 for the diabetes dataset.

The edges in the network graph represent the coefficient themselves and the thickness of it is a measure of its weight. All the graphs showed similar thickness of the edges indicating that the coefficient values lie within a short range. The sizes of the nodes are a measure of the degree of correlation. It is observed that almost all coefficients have values more than 0.8 and 3 of them have values above 0.91. This indicates that these are highly correlated gene expressions.

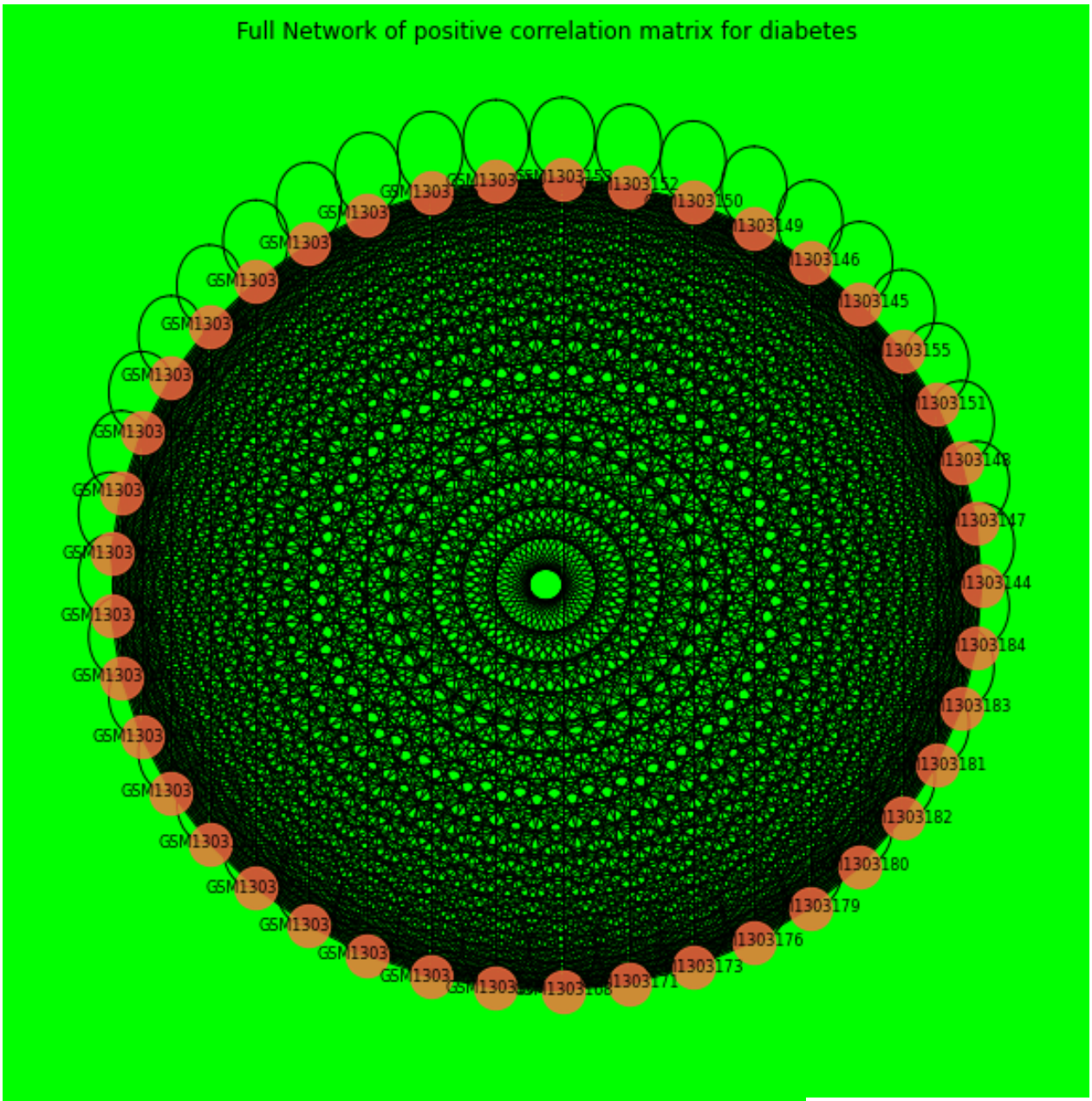


Figure 7: Network diagram of the correlation matrix of brain aging data

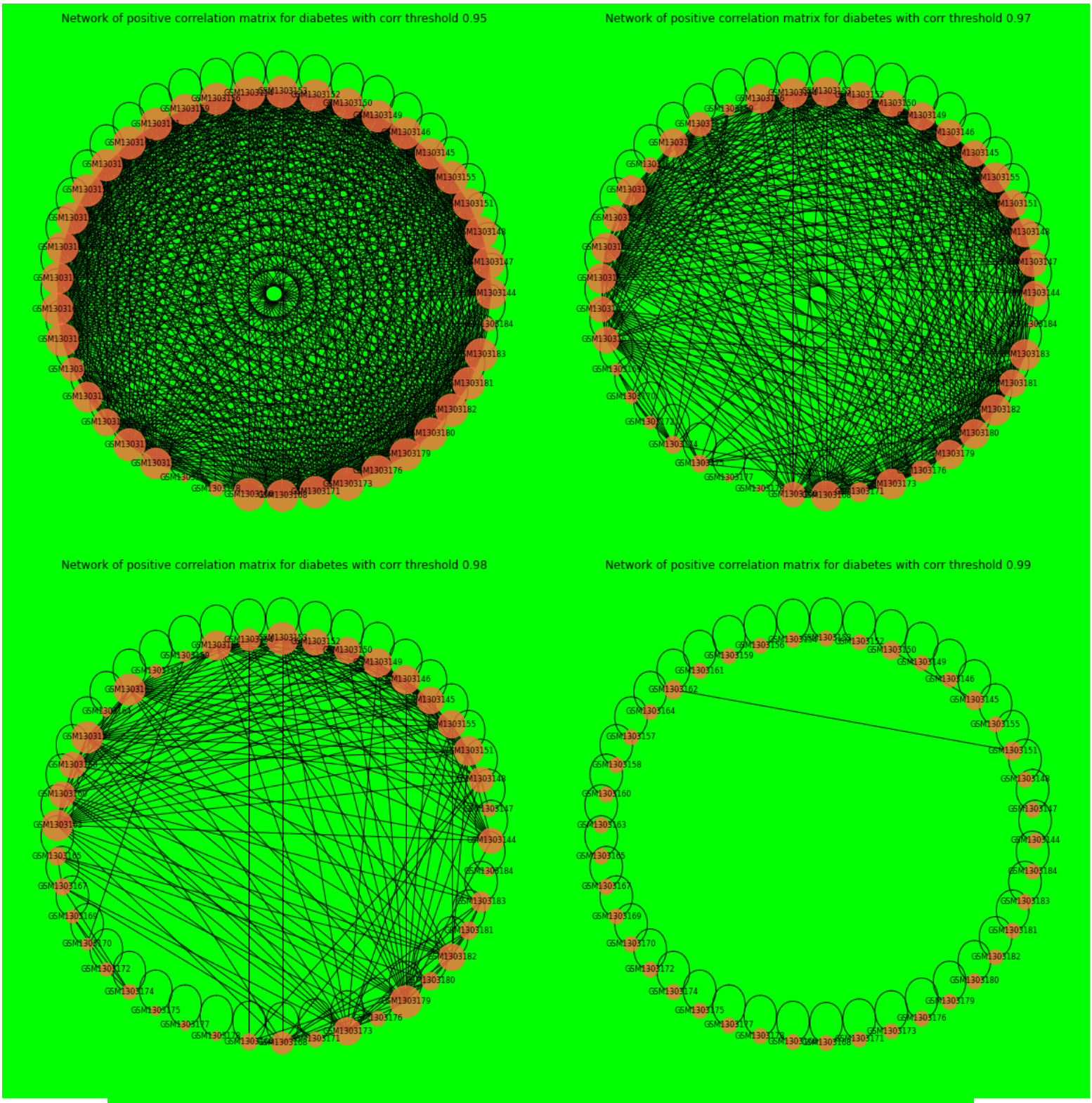
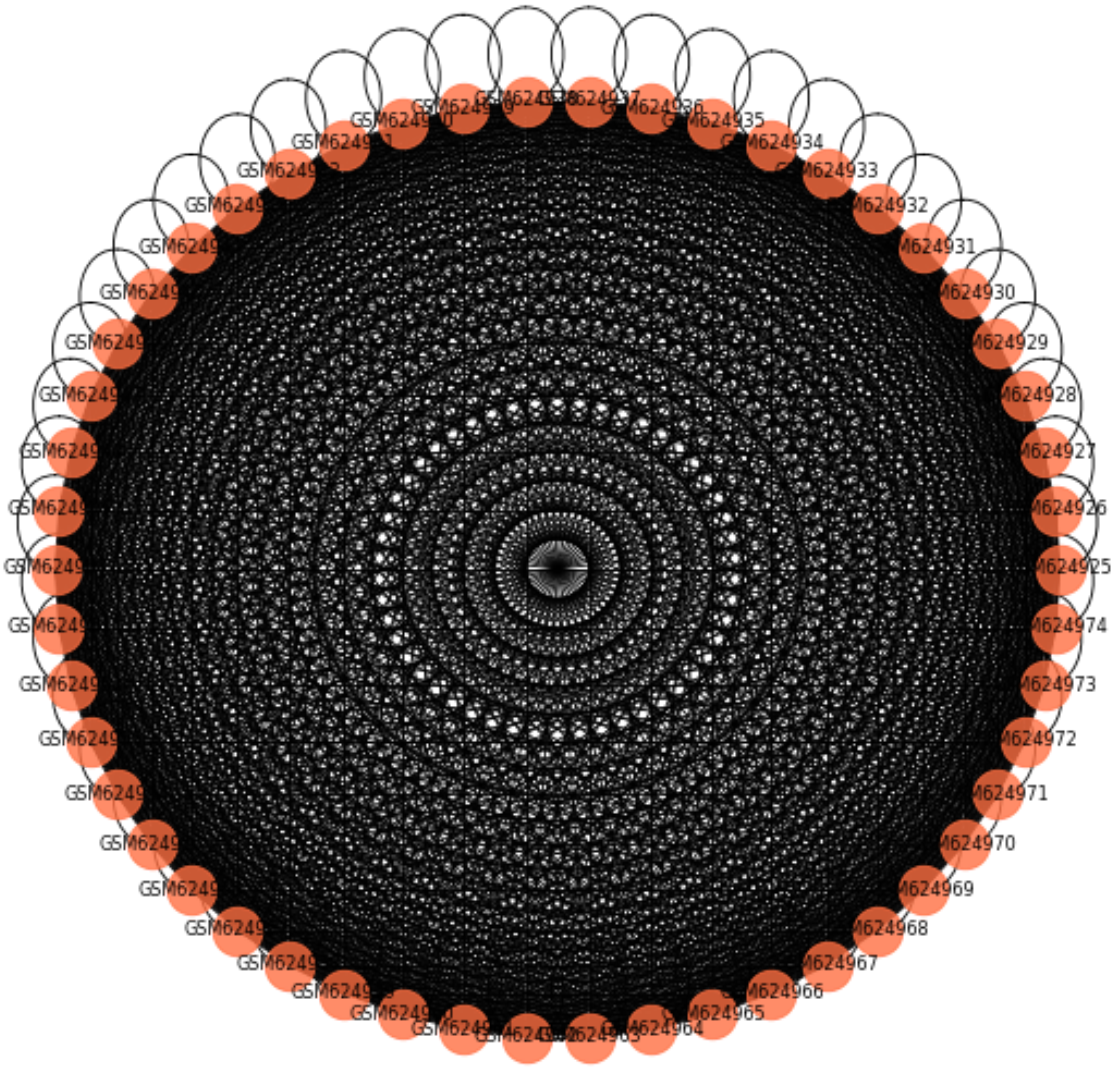


Figure 8: Network of correlation matrix for threshold values of 0.95, 0.97, 0.98 and 0.99 for the brain aging dataset.

Full Network of positive correlation matrix for diabetes



REFERENCES

1. Jin, W., Goldfine, A., Boes, T., Henry, R., Ciaraldi, T., Kim, E., Emecan, M., Fitzpatrick, C., Sen, A., Shah, A., Mun, E., Vokes, M., Schroeder, J., Tatro, E., Jimenez-Chillaron, J. and Patti, M., 2011. Increased SRF transcriptional activity in human and mouse skeletal muscle is a signature of insulin resistance. *Journal of Clinical Investigation*, 121(3), pp.918-929.

2. Lu, T., Aron, L., Zullo, J., Pan, Y., Kim, H., Chen, Y., Yang, T., Kim, H., Drake, D., Liu, X., Bennett, D., Colaiácovo, M. and Yankner, B., 2014. REST and stress resistance in ageing and Alzheimer's disease. *Nature*, 507(7493), pp.448-454.

3. Kay, A., Simpson, C. and Stewart, J., 2016. The Role of AGE/RAGE Signaling in Diabetes-Mediated Vascular Calcification. *Journal of Diabetes Research*, 2016, pp.1-8.

4. Shoffner, J., 1997. Oxidative phosphorylation defects and Alzheimer's disease. *neurogenetics*, 1(1), pp.13-19.