Chris Day, Samuel Bryan, Enleo Dahal

Dr. Kai Kang

STT 303

5/7/25

GTEx Final Project

In our project, we decided to take from the GTEx dataset. GTEx is a public project which studies the adult genotype tissue expression. For our dataset, we decided to use the median gene level TPM (transcripts per million) organized by each tissue. we will be analyzing the RNA sequence and use various methods to see if there is any correlation and causation with variables to one another. It will involve models such as regression, classification, and clustering.

Based on our analysis of the available data provided by the GTEx Portal provided by the Broad Institute, we decided on using the median gene level of each tissue. This was chosen in order that we may have a broad scope in how the the tissues had effect and correlated with one another. First was needed to clean the dataset. When the headers, were at the top of the column, it was read in order to display the data. Along with this was to scale the data so that we would have accurate predictions of the tissues to one another. When performing quantitative analysis of the data, the rows that had name and description needed to be removed in order to perform our models.

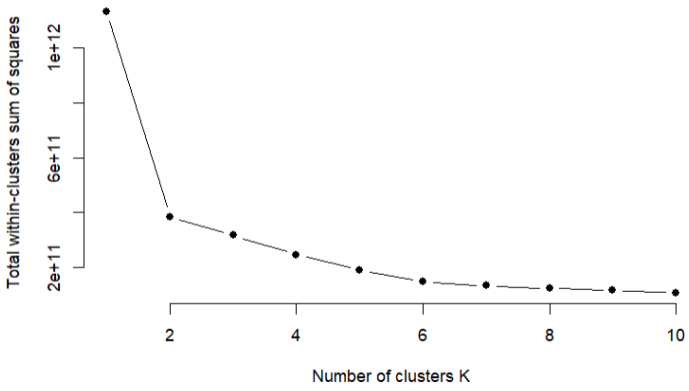

Figure 1

In clustering, to find the the K-means of the data we needed to determine how many clusters was needed. This was done by performing the elbow method above by seeing where error was at its lowest point without overfitting. This is what is referred to as the “elbow point”, representing the best amount of clusters to start with for our data.

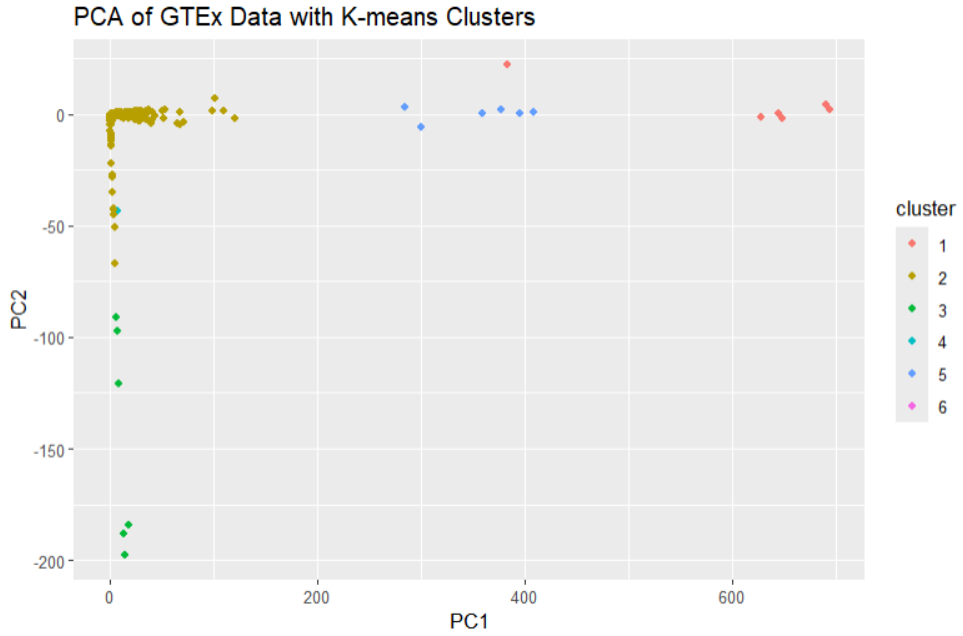

Figure 2

Above was the result of the K means clusters created by comparing PC1 and PC2 variables of the GTEX data to one another. With these clusters, we are able to tell where part of the data aggregates along with what outlying clusters are identified, that are closely related to other data points. With this analysis, we were able to determine that clusters 1 and 6 presented themselves as outliers from where the rest of the data aggregated when using the PC1 and PC2 metrics.

Moving along to regression, the same GTEX data was used to answer the question being asked: Is there a linear relationship between the expression levels of the two different types of tissues? To answer this question, we conducted two different regression models, each comparing two different sets of tissues. Each of the points on the respective graphs represents a pair of expression levels for a specific gene or sample. The x-coordinate is the expression level in one tissue and the y-coordinate being the expression level in another tissue. It is also important to reinforce the fact that this set is measured in Transcripts Per Minute (TPM), so the “40000” that you see on the model is the same as “0.4” when scrolling through the data set.

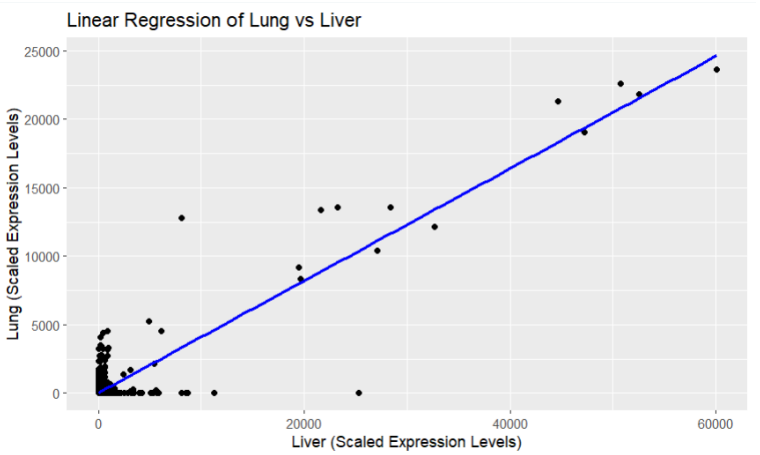

Figure 3

Here, we can identify the relationship between the Lung and Liver tissues and observe their scaled expression levels. After calculating the correlation coefficient in R, r = 0.91, which is a strong positive linear relationship, meaning that the tissues heavily influence one another. This result makes sense as there is a lot of build up right around the line, at the bottom of the regression model, and as the line continues, there are only one or two evident outliers.

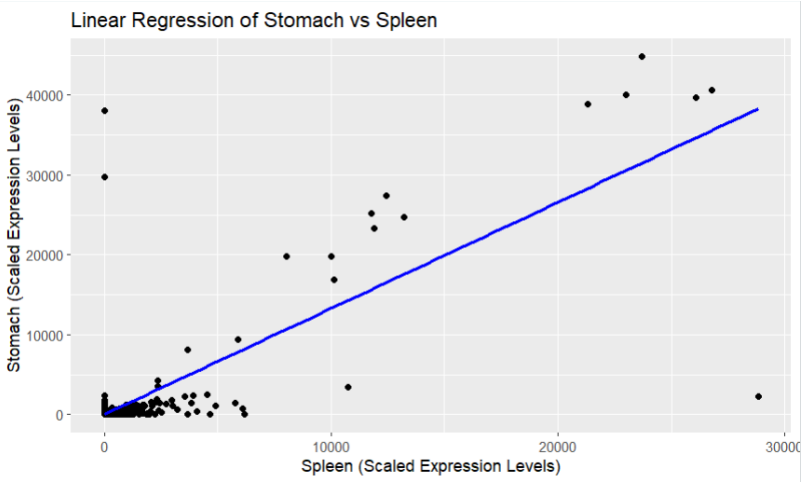

Figure 4

For this model, we look at the stomach and spleen tissues. From the immediate eye test, it can be said that the correlation would not be as strong as there is a greater variability in the data points. Also, compared to the last model, there is more clumped-up data, but it isn't as symmetrical. This also plays into a weaker correlation coefficient. After running it through R, r = 0.8. Though this is still a strong positive correlation where the tissues will

influence one another by a good bit, it is not as strong as the previous experiment with the lung and liver.

Another metric that evaluates the validity of the model is the Root Mean Square Error (RMSE). This is a measure used to gauge the prediction of errors of a regression model. In the context of Figure 3, on average, the predicted scale expression levels for lung tissue differ from the actual observed by 108 units. In terms of Figure 4, it was about 297 units. For the case of both models, the high RMSE's implies that the model does not perfectly capture the relationship between the two sets of tissues. It is especially evident in Figure 4 with an RMSE almost three times that of Figure 3. While the model captures a relationship between the tissues, there is more to the relationship beyond the regression.

Clustering

Using GTEx10 data set we merged each sample's filters gene-expression profile with the subject metadata combine in sample attribution and subject phenotype. We had to re-express the data as orthogonal "eigengenes" ranked by explained variance. This was done because there were tens of thousands of genes which is too many dimensions to perform a principal component analysis (PCA). The first 2 PCs alone captured –47% of the total transcriptional variation (PC1-29%, PC2=18%). When the graphs were generated and color coded by tissue there were clusters of similar tissue types clumped together. When the graphs were color coded by Age and Sex there was no clear cluster to be seen. We next projected each sample into the firs 10 PCs which retained >60% of the total variance while

discarding noise. We ran it with K-means of K=3. The 3 distinct cluster that appeared were as such.

Cluster	Dominant tissues in the contingency table	Biological theme (shorthand)
1 (red)	Adipose, colon, esophagus, heart, skeletal muscle	High-energy / contractile & metabolic
2 (green)	Brain, whole blood, heart (second mode)	Neuro-haemato-cardiac
3 (blue)	Skin, blood vessel, thyroid, peripheral nerve, testis	Peripheral epithelial / endocrine

In conclusion, age and sex had little to no effect on gene expression except for sex based specific tissue that are distance in men and women. The clearest cluster was tissue type based which clumped together to make distinct clusters. K-mean and dimensional reduction was able to group the expression that could be interpreted to have similar themes, but that correlation was not abandonly clear.