Introduction

In this project, I wanted to see if we can determine whether tissue type (Lung, Liver, or Heart) can be predicted from human gene expression data

Datas used

GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt

- Metadata with information about each biological sample in the GTEx project

GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct

- gene expression measurements from thousands of human tissue samples collected

The GTEx project provides gene expression data across various human tissues.

Each sample includes thousands of genes quantified using TPM.

For this project, we chose three tissues lung, liver and heart – left ventricle

Data Preparation

Selected relevant tissue samples from the metadata

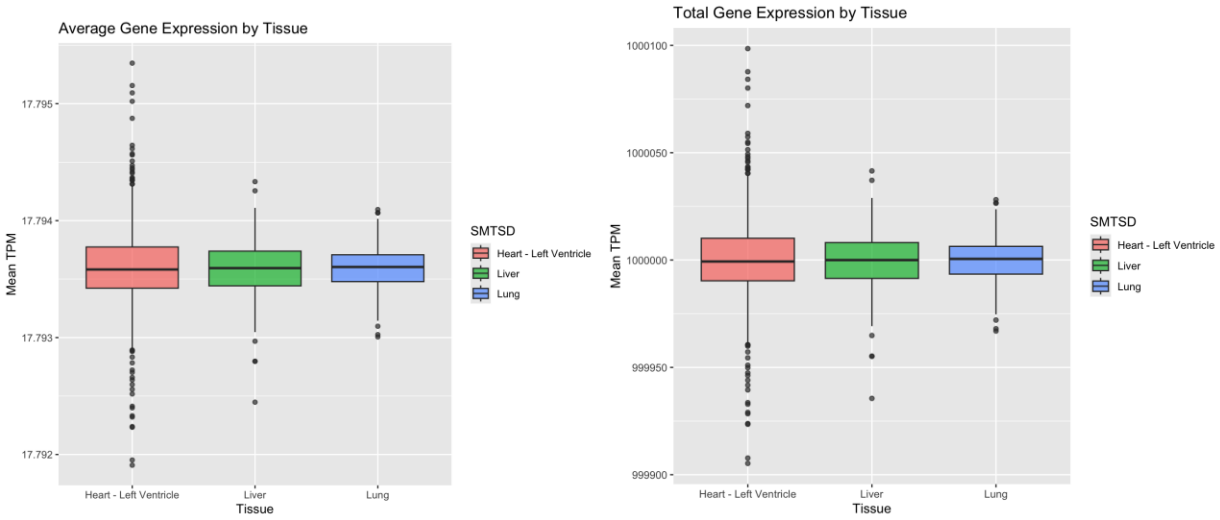
Matched samples between the metadata and the gene expression

Filtered to common IDs only

matrix: genes became columns, samples became rows

Exploring Data

Average and Total TPM by Tissue



Checked out both average and total gene expression by tissue

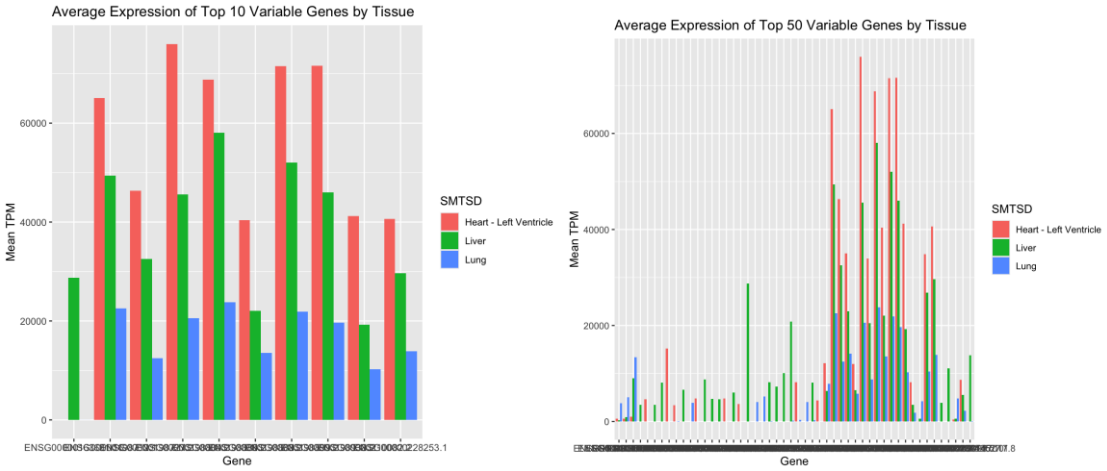
Compared using boxplots

The boxplots show that

- Both average and total gene expression by tissue is relatively the same with similar distribution
- Heart has the most outliers followed by liver and lung
- biological variation between tissues likely lies not in the overall amount of expression, but in which specific genes are expressed differently

Chose top 10 and 50 most variable genes by expression across tissues and measured and compared the expression

Top 10 & 50



Top 10

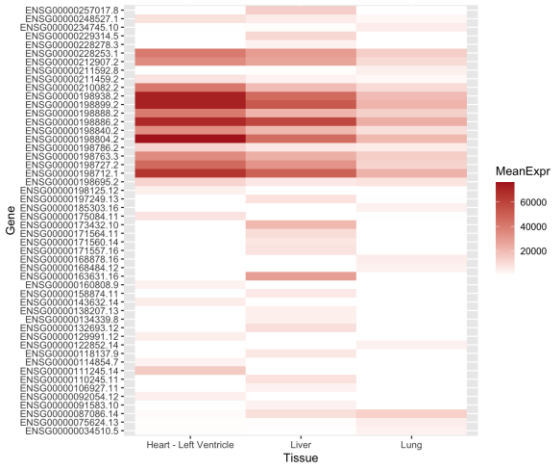
With the bar plot, we can clearly see most of these genes are strongly expressed in heart, and less in liver and lesser in lung

Top 50

Some genes, all 3 genes are highly expressed

Some tissue specific genes, especially liver

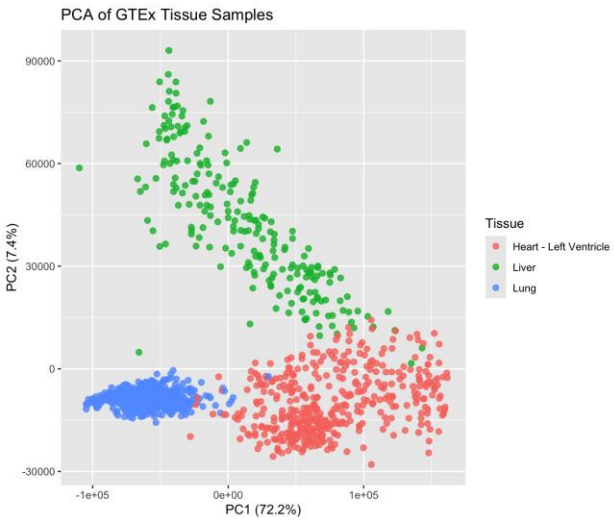
Heatmap



Similarly to top 50 bar graph, heart mostly has the highly expressed genes, and lesser in liver and lung

PCA

Use of PCA helps us visualize relationships between different variables



PC1 explains 72.2% of the variation and PC2 explains another 7.4%

This PCA graph shows

- The samples are grouped by tissue
- Lung is the most tight cluster

We can tell that the gene expression naturally differentiates tissues

Supports and strong signal of the idea of gene expression can be used to clarify tissues

Linear Regression

```
> summary(lm_avg)

Call:
lm(formula = AverageExpr ~ SMTSD, data = reg_df)

Residuals:
       Min         1Q     Median         3Q        Max
-1.682e-03 -1.367e-04  4.110e-06  1.347e-04  1.756e-03

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept)  1.779e+01  1.397e-05  1.273e+06   <2e-16 ***
SMTSDLiver  -6.017e-06  2.385e-05 -2.520e-01    0.801
SMTSDLung    2.281e-06  1.847e-05  1.230e-01    0.902
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0002905 on 1233 degrees of freedom
Multiple R-squared:  0.0001075, Adjusted R-squared:  -0.001514
F-statistic: 0.0663 on 2 and 1233 DF,  p-value: 0.9359
```

No meaningful difference in average gene expression across Lung, Liver, and Heart samples

Suggests looking at individual genes rather than overall mean

Multiple R-squared: 0.0001075

→ The model explains only 0.01% of the variation in average gene expression.

Adjusted R-squared: Negative

→ This indicates the model performs worse than just using the mean.

p-value for F-test: 0.9359

→ The overall model is not statistically significant.

Multinomial Logistic Regression

Training: 80% of samples

Testing: 20% of samples

From top 50 most variable genes

Classification model

```
>  class_metrics
                                        Tissue Precision Recall    F1
Heart - Left Ventricle Heart - Left Ventricle     1.000  0.989 0.994
Liver                                   Liver     1.000  1.000 1.000
Lung                                     Lung     0.991  1.000 0.996
```

Precision: how many predicted tissue labels were actually correct

Measured 1,000 for both heart and liver, meaning the model never mislabeled them. Lung with .991 with high precision

Recall: how many of the true samples were correctly identified

Perfect score for liver and lungs. Heart with .989 with high recall

F1 Score: mean of precision and recall

Liver scoring perfect score and others with very high score, reflecting excellent model performance

K-means Clustering

To explore whether gene expression patterns could naturally separate the samples by tissue type without using the tissue labels

k=3

Cluster sizes:

Cluster 1: 353 samples, Cluster 2: 467 samples, Cluster 3: 167 samples

```
> table(Cluster = km$cluster, Tissue = train_scaled_df$SMTSD)
       Tissue
Cluster Heart - Left Ventricle Liver Lung
      1                   340    12    1
      2                     5     1  461
      3                     0   167    0
```

Table above shows the confusion matrix between K-means clusters and actual tissue types

As we can see it is very accurate

To assess clustering quality, we reported the Between-cluster sum of squares divided by total sum of squares:

BetweenSS / TotalSS = 58.7%

58.7% of the total variation in the data is explained by the clustering

Summary

Our analysis revealed clear, tissue-specific gene expression patterns across 3 tissues (lung, liver, and heart – left ventricle)

Average and total expression were similar for each genes

PCA demonstrated that samples from each tissue cluster naturally in reduced dimensions, even before any supervised modeling.

A multinomial logistic regression model, trained on only the top 50 genes, achieved extremely high accuracy—over 99%—in classifying tissue type.

K-means clustering, an unsupervised method, was able to recover tissue groupings with good alignment to true labels, even without any training labels.