Task 2 - Data Science Project

Kartika Waluyo & Vrinda Rajendar Rajanahally

1000555 & 1129446

```
## Loading required package: edgeR
```

```
## Loading required package: limma
```

```
blood_shared_genes_unique = unique(blood_shared_genes)
```

```
corr_stats_list = corr_stats_list[!(corr_stats_list$tissueX=="Whole Blood"),]
```

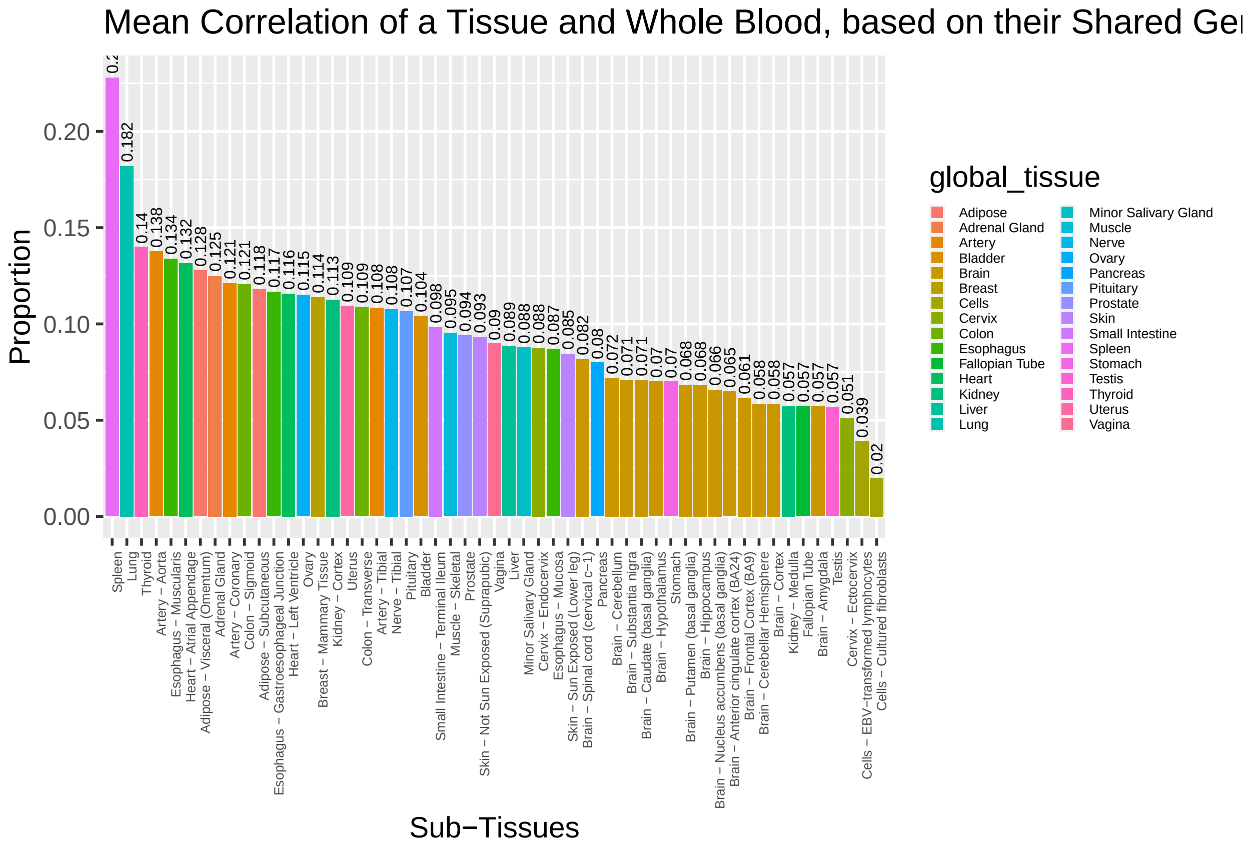
#Task Description

This task involves the exploration of correlation between gene expressions in tissues with the expression of the same genes in Whole Blood and their relationship. Additionally, the correlation of each gene found in Whole Blood with each tissue is also explored.

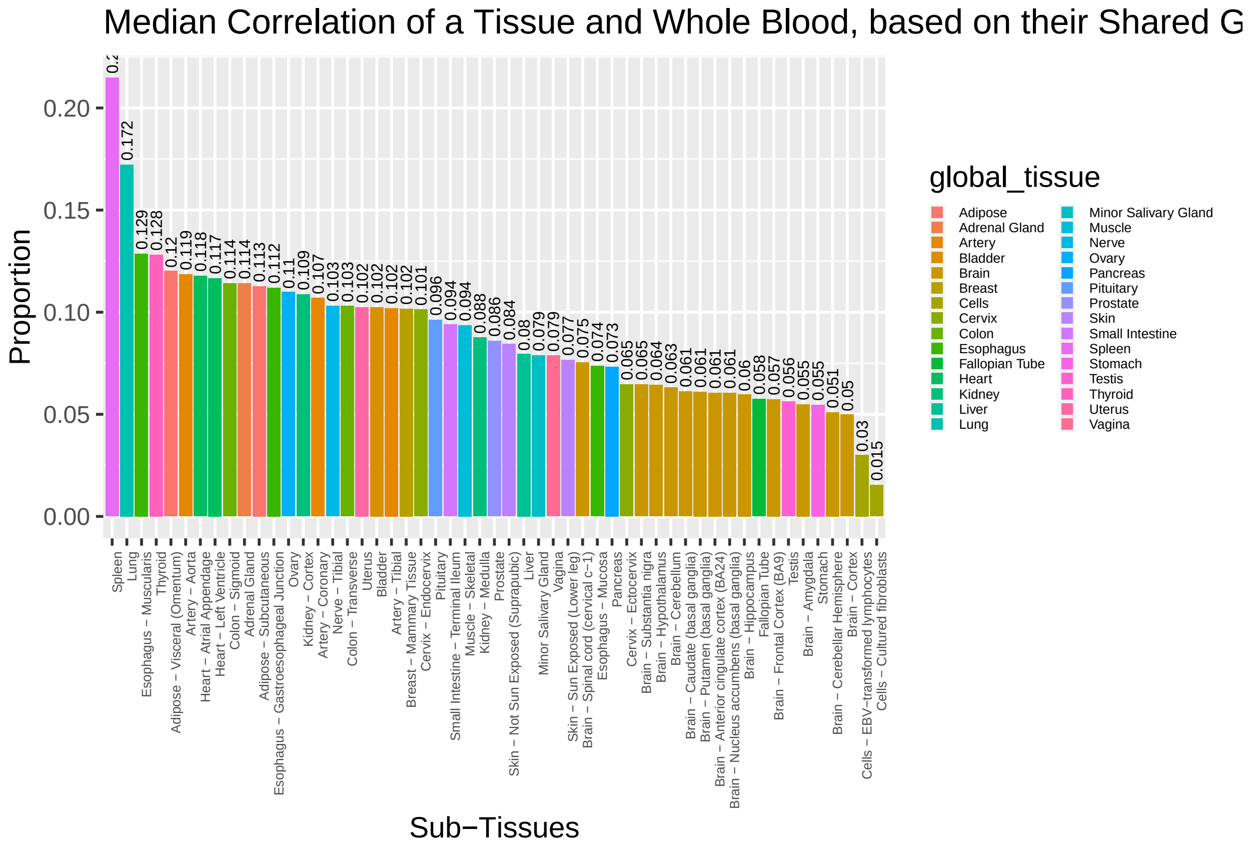
Various measures of correlation of a particular tissue and Whole Blood, based on their shared genes

To go about the first part of this task, we first create a vector that contains the correlation for each gene expressed in the tissues and Whole Blood. Mean and Median of the correlation for each pair is found to summarise the measure of correlation across all gene expressions. We also calculate plot the Absolute Mean and Absolute Median of correlations across the tissues and Whole Blood to understand the significance a bit better.

The mean and median correlation plots are as follows:



The above plot captures the mean correlation between a particular tissue and Whole Blood, based on the shared genes.

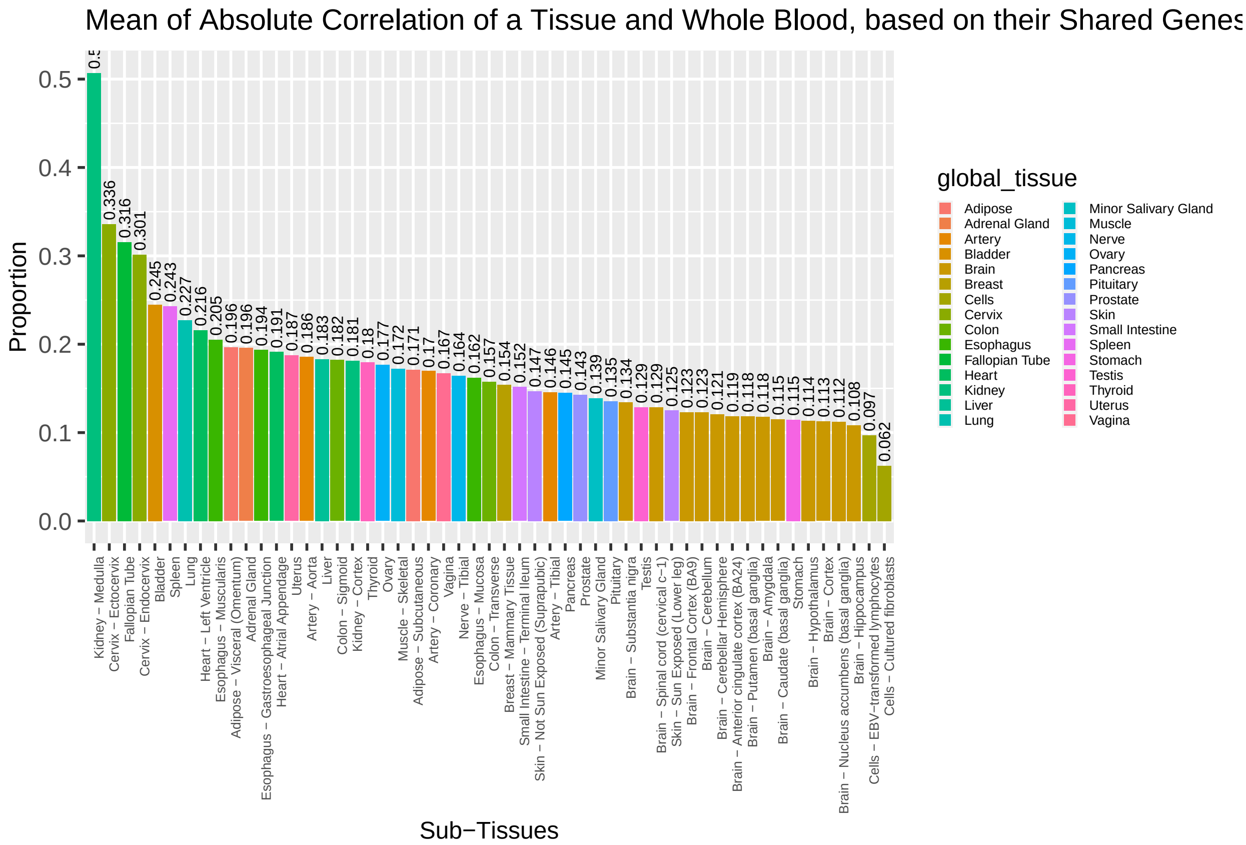


The plot above captures the median correlation between particular tissue and Whole Blood, based on their shared genes.

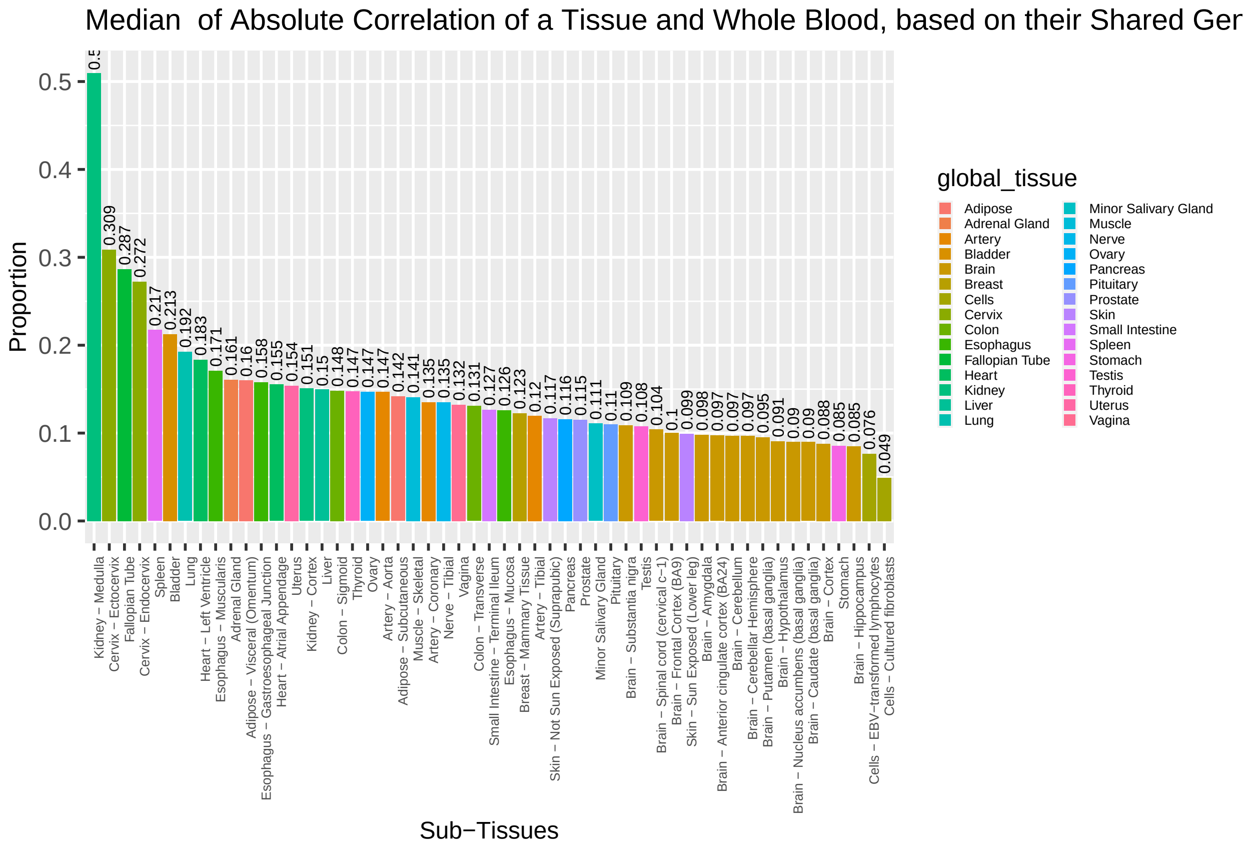
The mean correlation and median correlation can help us determine the how high or low the count of a particular shared gene in a tissue will be, when the count of the same gene is known in Whole Blood. It must be considered as an just an approximation of how high or low the count will be, and not as a predicting factor.

According to the two plots, we see that Spleen has the highest mean and median correlations and on the other hand, Cells - Cultured Fibroblasts have the least mean and median correlations with Whole Blood. It is also observed that all the individual tissues captured under 'Brain' have similar mean and median correlation values. Overall we can conclude that in this case, mean correlation and median correlation behave similarly. A closer look at each value of mean and median correlation values for each tissue suggests that they are roughly equal to each other.

Now considering the absolute correlation values, the following plots are produced:



The above plot represents the mean of absolute correlation between a tissue and Whole Blood, based on their shared genes.



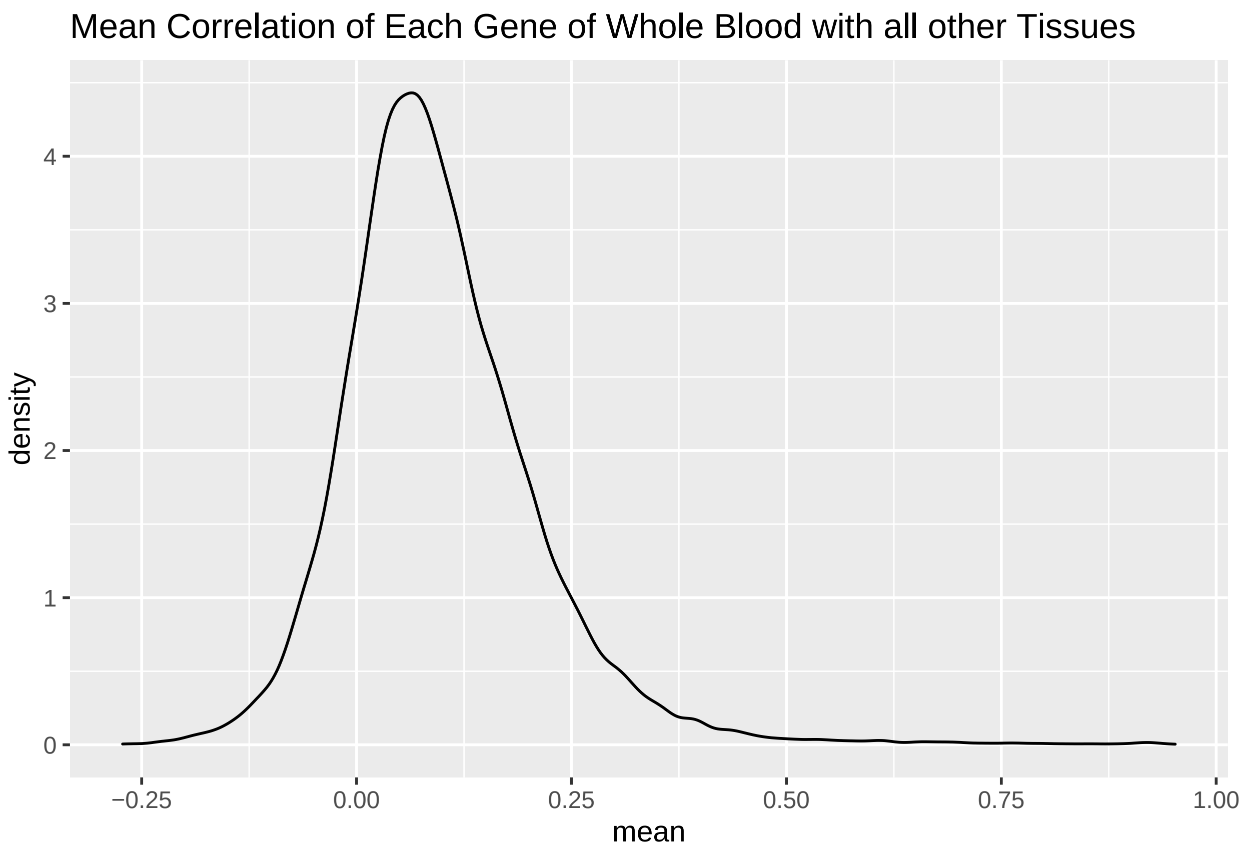
The above plot captures the median of absolute correlation between a tissue and Whole Blood, based on their shared genes.

In the context of this task, the absolute correlation is a measure that denotes the strength of the relationship between the tissue and Whole Blood, based on their shared genes. The direction of the relationship between the two is not accounted for while using absolute correlation. For each of the shared genes, we then take the absolute correlation and then continue to aggregate it using mean and median to get a summarised picture.

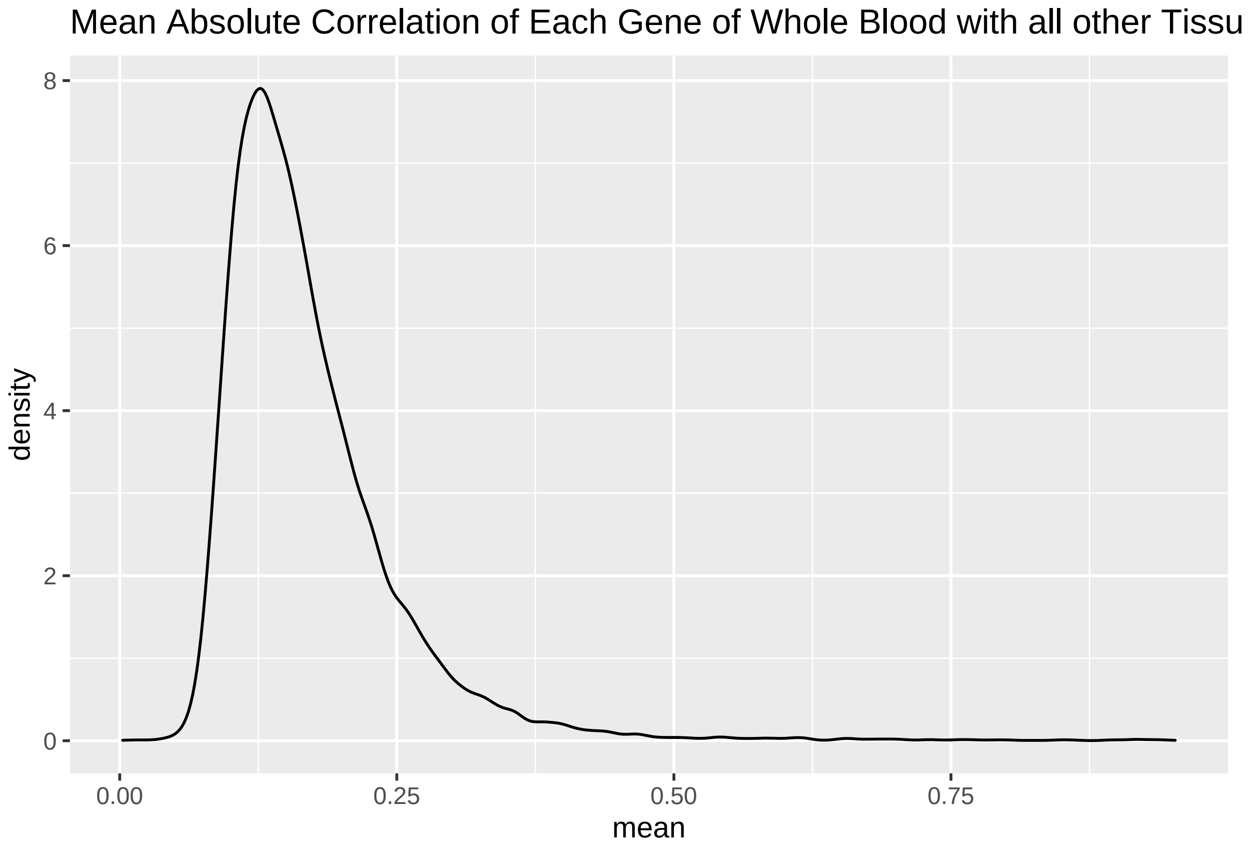
From the two plots, we can see that Kidney - Medulla has the strongest relationship with Whole Blood, based on their shared genes. Furthermore, Cells - Cultured Fibroblasts share the weakest relationship with Whole Blood. All specific tissues that fall under the 'Brain' category have roughly the same strength of relationship between Whole Blood. This can lead us to further conclude that the of mean and median of absolute correlation are roughly are the same values for each tissue.

Various measures of correlation between each gene found in Whole Blood with all other tissues

The second part of the task explores how each gene found in Whole Blood is correlated to the same gene found other Tissues. For this task, we will utilise mean correlation and median correlation to summarise the overall picture. Density plots are also to graphically represent the findings.



The above density plot represents the distribution of the mean correlation of all genes found in Whole Blood with all other tissues. The peak observed in this density plot suggests that the overall mean correlation is positive. While some genes of some tissues may be highly correlated with the genes of Whole Blood, not all of them are.



Since this plot considers absolute correlation of each gene in Whole Blood with the same gene found in all other tissues, it starts at zero. Majority of genes in other tissues have a correlation of about 0.10, but a very small number of them have a strong correlation.

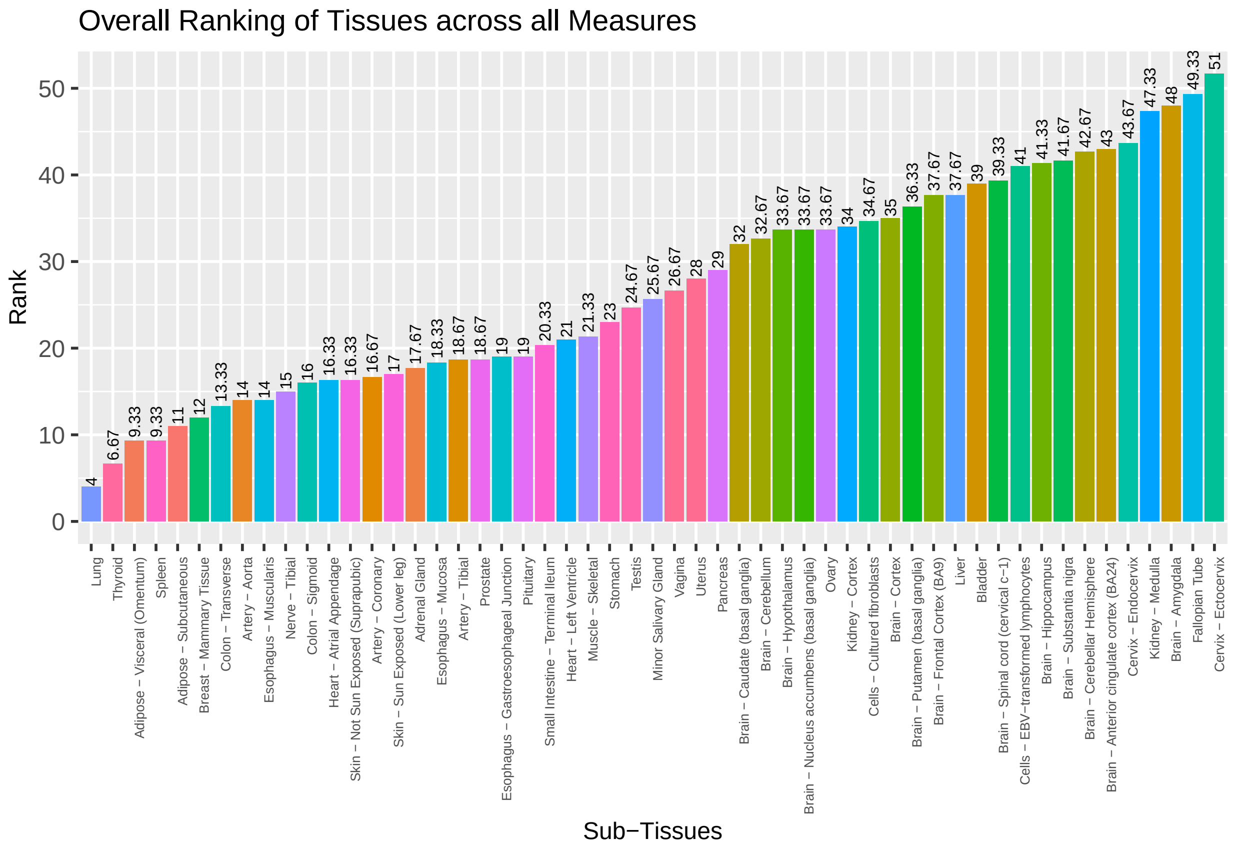
```
#install.packages("tidyft")
library("tidyft")
```

Ranking tissues based on various measures

The final part of the task involves ranking tissues based on three different measures: proportion of overlapping donors, proportion of shared genes and correlation. The rank per tissue per measure is averaged out to give us the final rank it holds.

The following list and bar plot represent the top tissues with the highest proportion of overlapping donors, proportion of shared genes and correlation. These are the best tissues to be chosen across all measures:

```
##                          subtissue proportion_donor rank_donor proportion_gene
## 37                            Lung        0.6185430          8       0.9833064
## 51                         Thyroid        0.6913907          5       0.9657666
## 2  Adipose - Visceral (Omentum)        0.5854305         10       0.9657666
## 48                          Spleen        0.2675497         26       0.9839218
## 1        Adipose - Subcutaneous        0.7099338          4       0.9562274
## 21      Breast - Mammary Tissue        0.5033113         13       0.9671513
##    rank_gene mean_corr rank_corr  avg_rank
## 37         2 0.1820026         2  4.000000
## 51        12 0.1402093         3  6.666667
## 2         11 0.1277764         7  9.333333
## 48         1 0.2279081         1  9.333333
## 1         18 0.1180928        11 11.000000
## 21         8 0.1138668        15 12.000000
```



Since ranks have been the most consistent unit across all measures so far, the ranking computation carried out can help with the selection of the best tissue across all measures.

#Conclusion

This task explores how shared genes across all tissues are correlated to Whole Blood. Each correlation value measured in this task helps us understand the kind of relationship every tissue shares with Whole Blood, based on the genes they share. A strong correlation is essential for analysis.

It also gives us a good starting point for choosing the right tissues to analyse and apply models to.