Task 01 - Data Science Project

Kartika Waluyo & Vrinda Rajendar Rajanahally

1000555 & 1129446

Project Overview

This project aims to utilise AI to predict gene expression across different human tissues.

The first task requires exploration of the gene expression data and tissue donor metadata to make an informed selection of tissues to take forward into the prediction project.

Task Description

To explore and visualise the count and proportion of donor and gene overlap between all pairs of tissue samples.

To carry out this task and understand each of the data frames better, four matrices containing the following information can be created:

1. Overlapping donor count: matrix that represents the count of overlapping donors between all pairs of tissues.
2. Overlapping donor proportion: matrix that represents the proportion of overlapping donors between all pairs of tissues.
3. Shared gene count: matrix that represents the count of overlapping expressed genes between all pairs of tissues.
4. Shared gene proportion: matrix that represents the proportion of expressed overlapping genes between all pairs of tissues.

Furthermore, each of the above matrices are visualized using heatmaps.

Additionally, bar plots are produced to understand the proportion of overlapping donors and shared genes amongst tissues, based off of different parameters.

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

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

The count of overlapping donors between all pairs of tissue samples

To visualise the count of overlapping donors between all possible pairs of tissues, a matrix is created to tabulate the count of overlapping donors for all pairs of tissues. It is a 54 x 54 square matrix, in which the row and column names correspond to the tissue names.

The diagonal elements of the resulting matrix give the total count of donors for each tissue. On the other hand, the non-diagonal elements give the total overlap of donors between two particular tissues. It is important to note that the count of overlapping donors between two tissues say Liver and Whole Blood is equal to the count of overlapping donors between Whole Blood and Liver. Hence, it is a symmetric matrix.

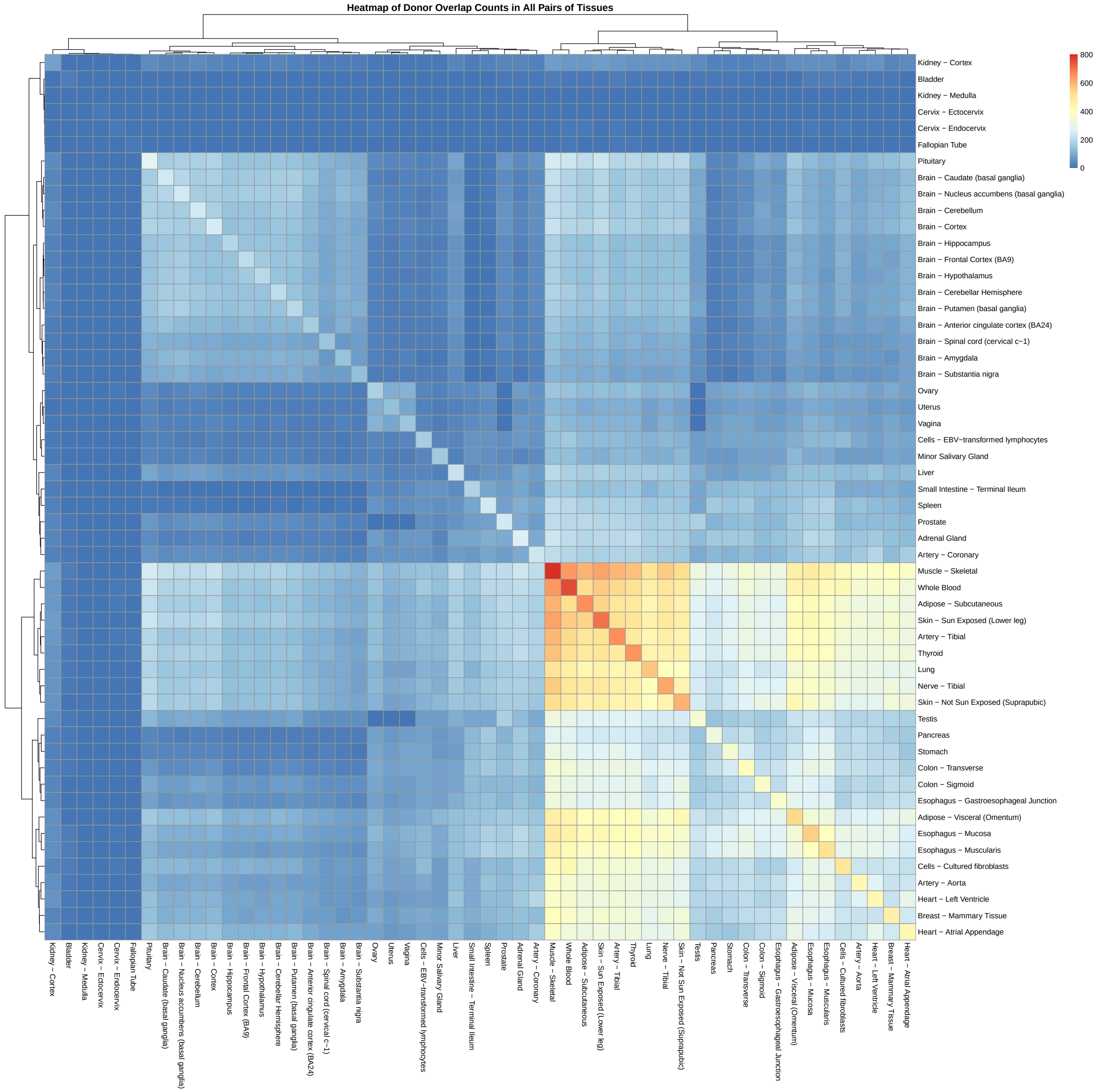

Figure 1: Heatmap representing no. of overlapping donors for each tissue pair

In this symmetrical matrix, rows and columns are clustered by similarity in terms of donor overlap. Since each element in Figure 1 is represented by a colour in the heatmap, it is easy to identify the tissue pairs with the least and most overlapping donors.

From Figure 1, Whole Blood and Muscle - Skeletal is one example of two tissues that have a high count of overlapping donors. On the other hand, Kidney - Medulla and Pancreas pair has an extremely low count of overlapping donors. The diagonal stands out as it gives the total count of overlapping donors per tissue. Muscle - Skeletal is the tissue with the highest number of donors whereas Kidney - Medulla seems to have the lowest number of donors.

As for the clustering, the resulting heatmap has clustered similar tissues close to each other based on overlapping count of donors. A portion of the bottom right quadrant (starting from Muscle - Skeletal row and column) displays a cluster that has the highest number of overlapping donors.The rest of the heatmap has clustered tissues that display an average to low count of overlapping donors.

The proportion of overlapping donors between all pairs of tissue samples

In this task, a matrix is created to tabulate the proportion of donors for all pairs of tissues. This is a 54 x 54 square matrix, in which the row and column names correspond to the tissue names.

In this proportion matrix, it can be observed that all diagonal elements will be equal to 1 since those cells represent the total proportion of donors per tissue. However, this matrix is non-symmetrical because every non-diagonal element corresponds to a different row tissue and a different column tissue while calculating the proportion per element. Each of them take a value between 0 and 1 inclusive and can be expressed in terms of percentages. For example, the proportion of overlapping donors between Liver and Whole Blood is 0.241 i.e overlapping donors between Liver and Whole Blood donors make up 24.1% of Whole Blood donors.

Every cell that is coloured red in Figure 2 denotes full proportion which is the proportion of the tissue to itself, or extremely high proportion. The other cells are likely to be values lesser than 1.

It is important to note that every column tissue acts as a numerator and every row tissue is the denominator. The column tissue's proportion is based on the number of overlapping donors shared with the row tissue and is divided by the total donors a row tissue contains. For example: according to Figure 2, the proportion of overlapping donors between Liver and Whole Blood is very high. Subsequently, the proportion of overlapping doors between Whole Blood and Liver is comparatively lower.

As for the clustering, there is an obvious cluster starting from the leftmost column up to the Esophagus - Muscularis column, which means that the donors of each column tissue make up high percentages of donors in almost all row tissues. The row tissues have also been clustered in such a way that the top most row tissues have the least proportion of overlapping donors, but as we go down the row, the proportion increases.

Another interesting observation is of the rows and columns of Testis and Vagina, whose corresponding cells have the least shared proportion of donors, a value very close to zero. This acts as a good check to see if the table gives us an accurate depiction of the overlapping donors across all tissue pairs.

Figure 3 visualises the proportion of overlapping donors of all Tissues with Whole Blood. It helps us identify that all donors of Cervix - Endocervix, Fallopian Tube, Kidney - Medulla are also donors of Whole Blood.

Figure 4 depicts the proportion of overlapping donors of Whole Blood in all other Tissues. Muscle - Skeletal has the highest proportion of overlapping donors with Whole Blood whereas, Kidney - Medulla has the least. It gives us a picture of how many Whole Blood donors have also donated other tissues.

Note: this graphical representation considers Whole Blood as the denominator.

The count of shared genes between all pairs of tissue samples

Next we wanted to investigate the similarity of tissues in terms of gene expression. To start off, a count matrix that records the count of shared genes across different pairs of tissues is created. It is a 54 x 54 square matrix, in which the row and column names correspond to the tissue names.

We observe that the diagonal elements of this matrix give us the total count of genes for each tissue. The non-diagonal elements give us the count of shared genes between two tissues. For example, the count of shared genes between Liver and Whole Blood is 11124, which is equal to the count of shared genes between Whole Blood and Liver. Hence, this matrix is also symmetrical.

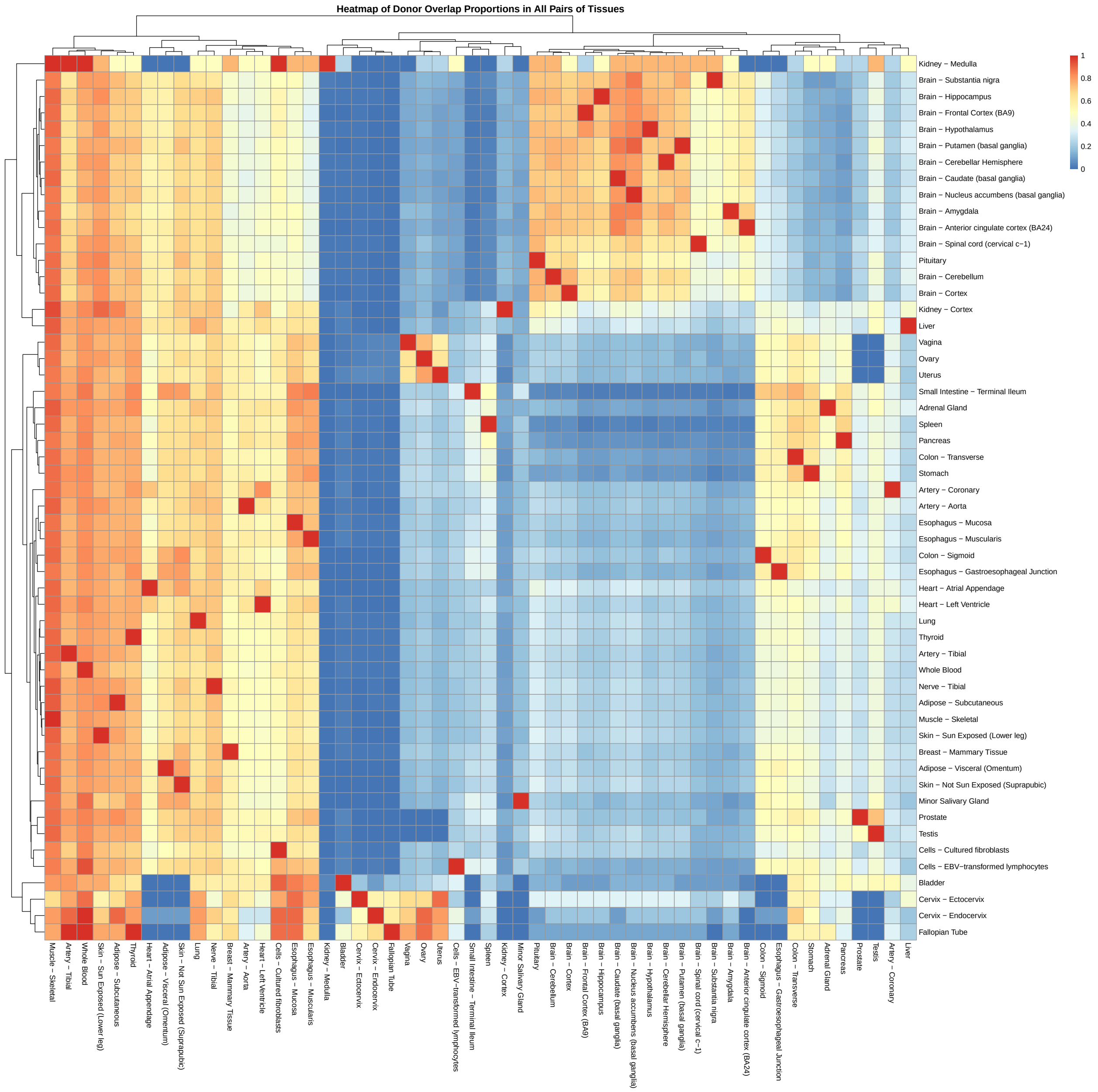

Figure 2: Heatmap representing proportion of overlapping donors for each tissue pair

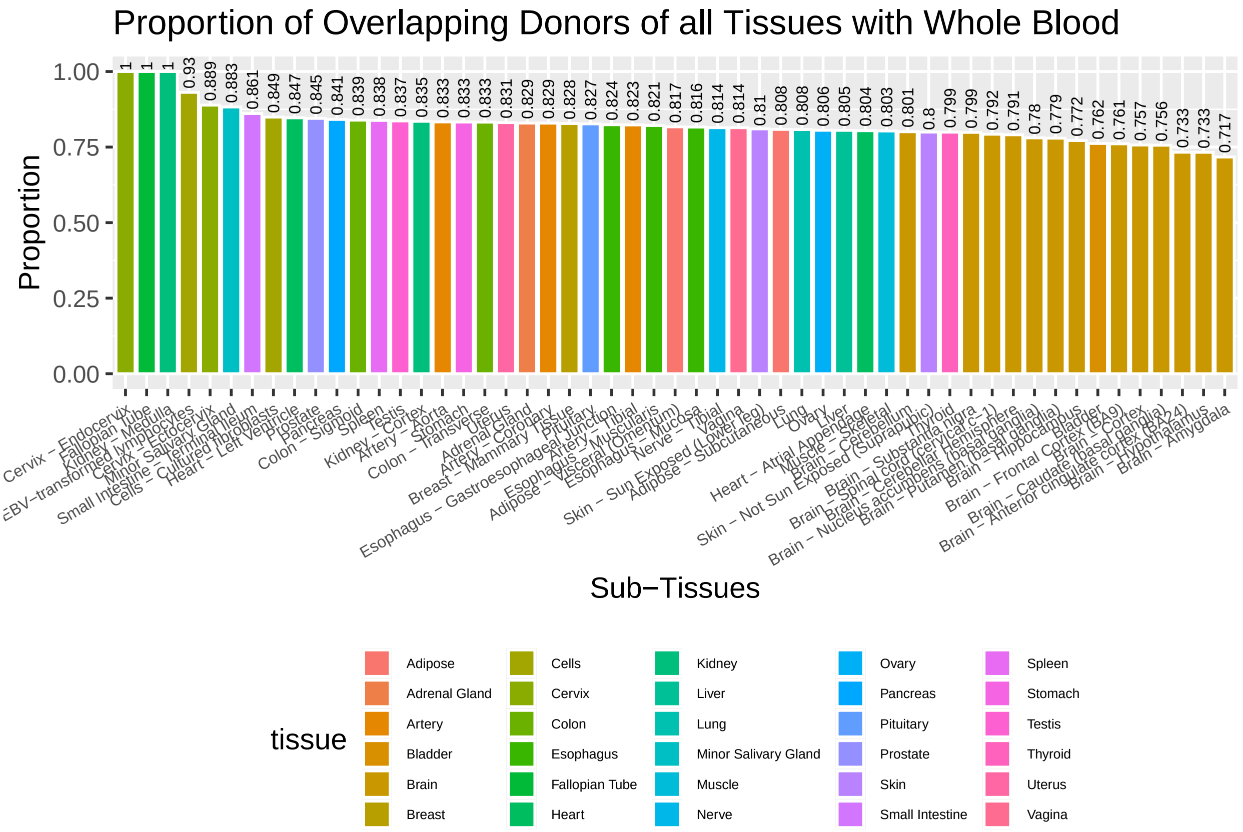

Figure 3: Bar plot representing the proportion of overlapping donors of all tissues with whole blood

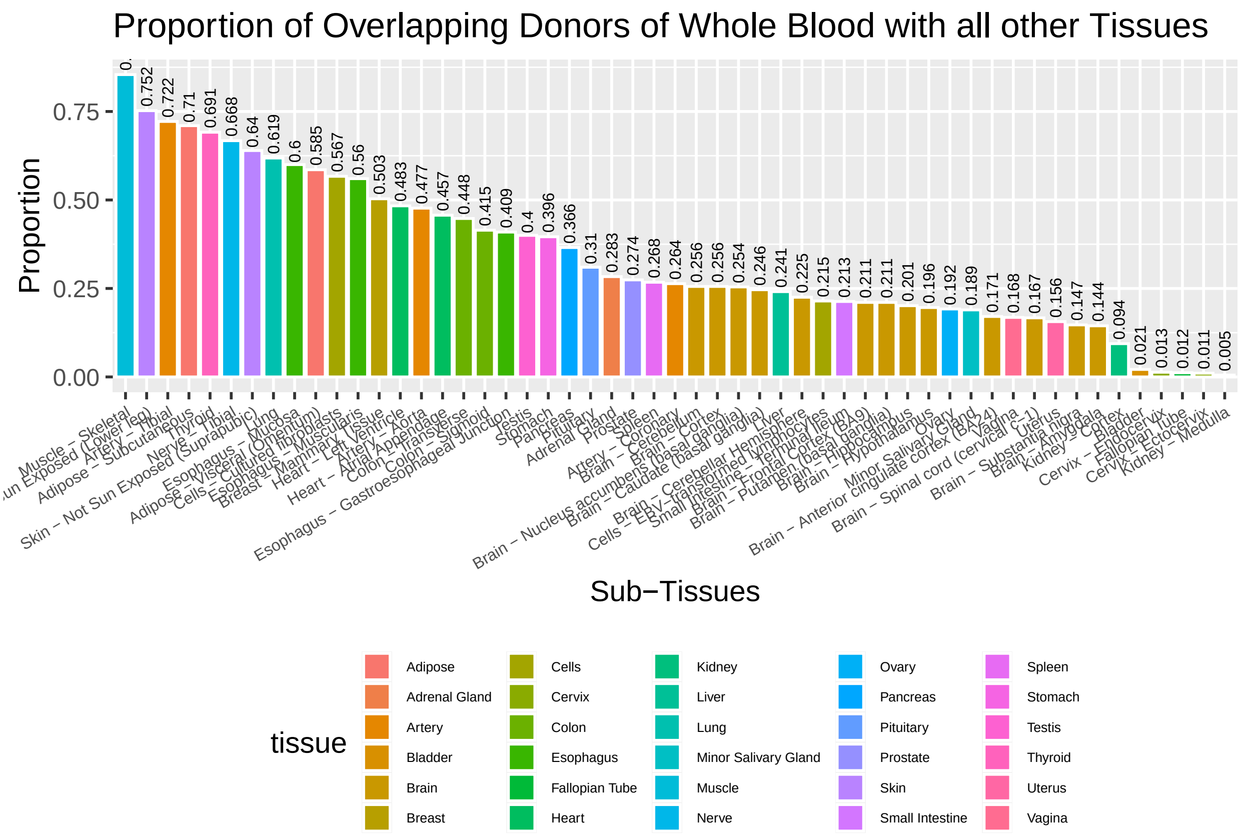

Figure 4: Bar plot representing the proportion of overlapping donors of whole blood with other tissues

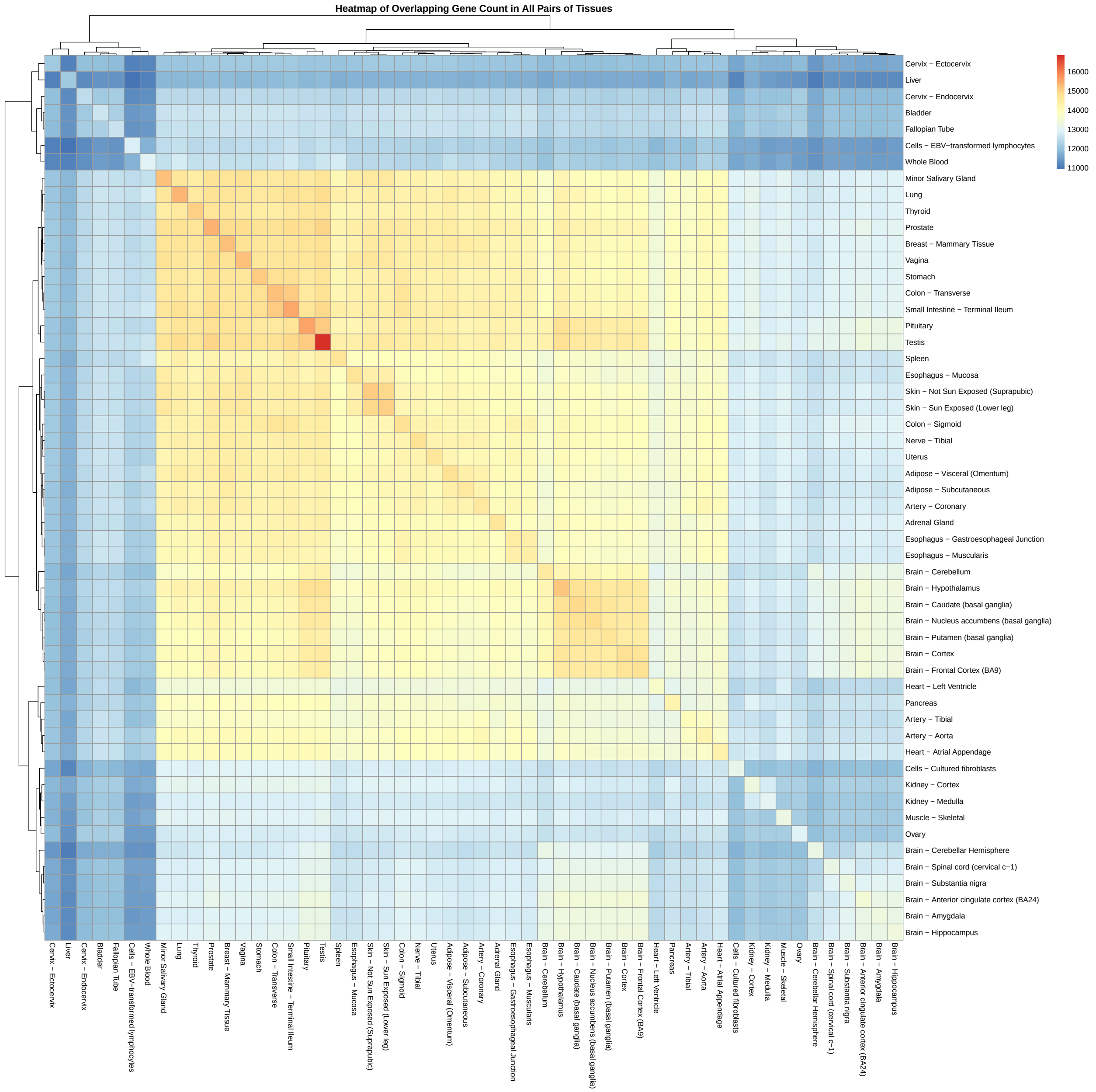

Figure 5 is a visual representation of the count matrix described above, in which the rows and columns represent all 54 tissues. In this symmetrical matrix rows and columns are clustered by similarity in terms of shared genes. Since each element in the resulting count matrix is represented by a colour in the heatmap, it is easy to identify the tissue pairs with the least and most count of shared genes.

From this heatmap, Brain - Cortex and Brain - Frontal Cortex (BA9) is one example of two tissues that have a high count of shared genes. Additionally, Liver and Brain - Cerebellar is an example of a tissue pair that has an extremely low count of shared genes. The diagonal in the heatmap stands out since it denotes the total count of genes of a tissue. Testis is the tissue with the highest number of expressed genes and Liver seems to have the lowest count of genes.

As for the clustering, we observe that similar row tissues and column tissues are clustered close to each other, according to count of shared genes. The middle portion of Figure 5, coloured in shades of red is a cluster that has a high to average count of shared genes. The rest of the heatmap has clustered tissues with low count of shared genes.

The proportion of shared genes between all pairs of tissue samples

To obtain the appropriate heatmap, a matrix that records the proportion of shared genes between different pairs of tissue is created. It is a 54 x 54 square matrix, in which the row and column names correspond to the tissue names.

The diagonal elements of this proportion matrix will all be equal to 1, as they represent the total proportion of genes per tissue. This resulting proportion matrix will be non-symmetrical.

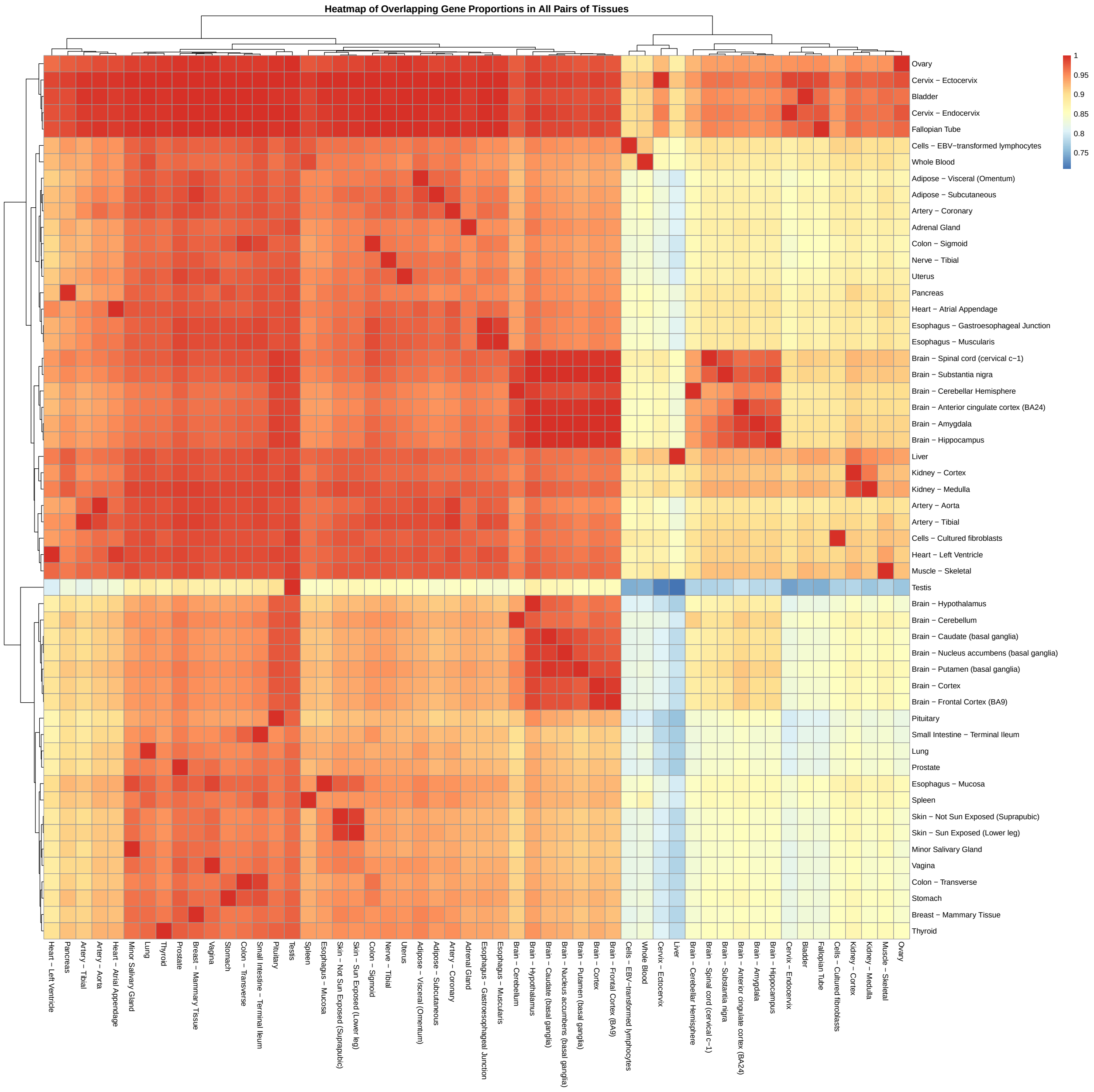

Figure 6 gives a graphical representation of the proportion of shared genes in different pairs of tissues, in which all rows and columns represent all 54 tissues. Every cell that is coloured red denotes full proportion which is the proportion of genes of the tissue to itself or a very high proportion. The other cells are likely to be values lesser than 1.

In this heatmap, we consider the column tissues as numerators and row tissues as denominators. While calculating each proportion, the number of genes shared between the column tissue and row tissue is divided by the total genes expressed in the row tissue. According to the heatmap above, the proportion of shared genes between Testis and every other tissue seems to be on the lower side. This is a salient outlier that can be noticed straight away. On the other hand, the proportion of shared genes of every other tissue with Testis is fairly high.

In terms of clustering, all similar gene proportions have been clustered together in the rows and columns. Majority of the heatmap is shades of red - which signifies that those column tissues make up a large proportion of their corresponding row tissues. Similarly, the portion of the bottom right quadrant, shaded in blue, denotes that the column tissues make up a small proportion of their corresponding row tissues.

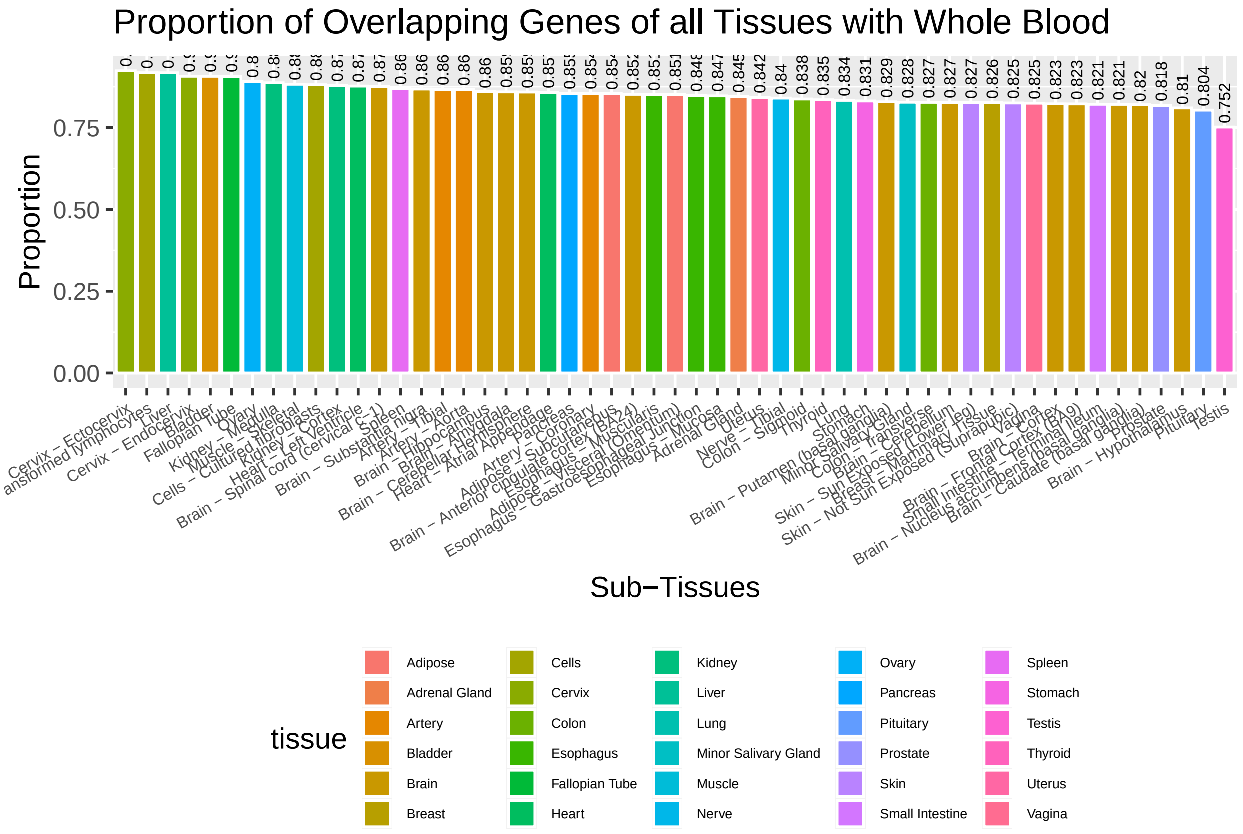

Figure 5: Bar plot representing the proportion of overlapping genes of all tissues with whole blood

Figure 7 is a visualisation of all overlapping genes of other Tissues with Whole Blood. Here, we observe that the proportion of shared genes of Testis and Whole Blood is the least. Whole Blood and Cervix - Ectocervix has the highest proportion of shared genes in this case.

Figure 8 shows the proportion of shared genes of Whole Blood with all other tissues. The Spleen has the highest proportion of shared genes, which are also found in Whole Blood. The proportion of shared genes in Liver and Whole Blood is the least amongst the lot.

Note: Figure 8 considers Whole Blood as the denominator.

Furthermore, to investigate the tissues that are the most similar to Whole Blood, we create the following lists:

1. List of the top 10 tissues that share most of its donors with Whole Blood

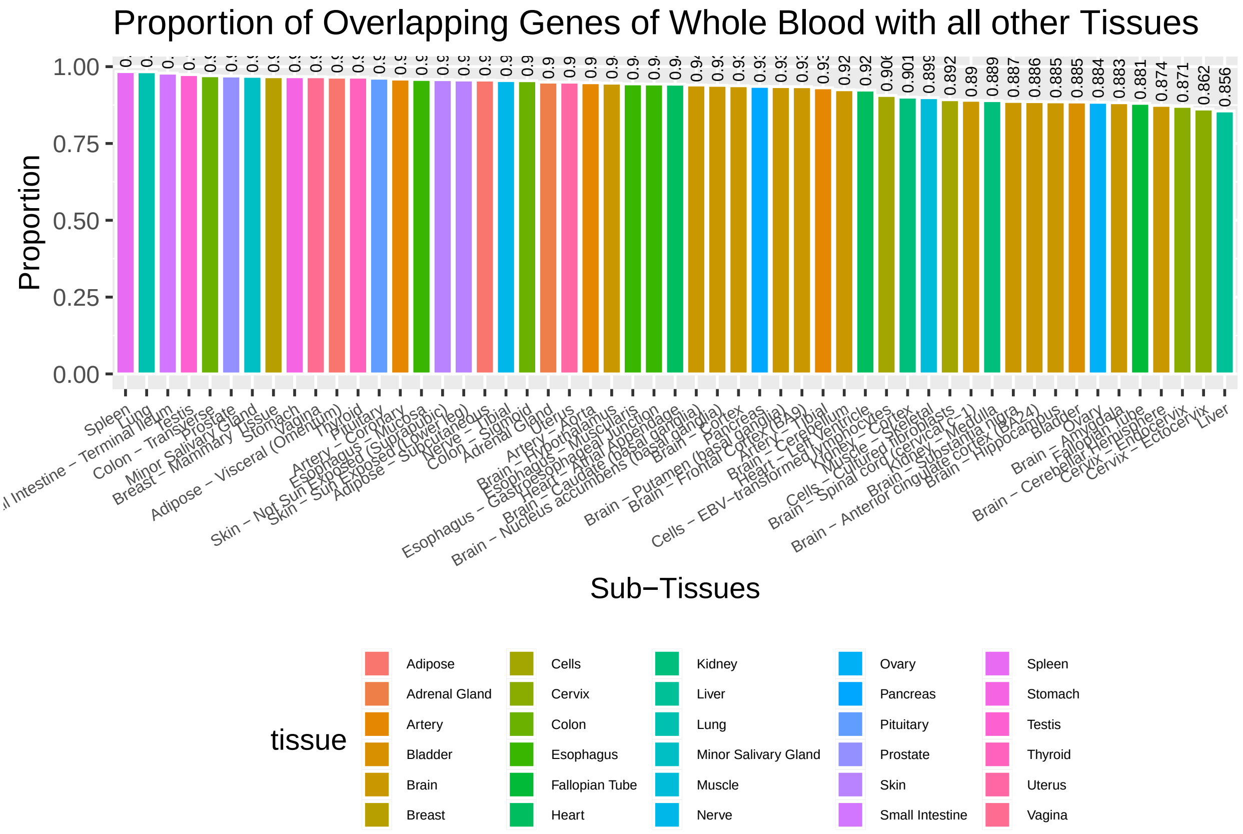

Figure 6: Bar plot representing the proportion of overlapping genes of whole blood with all other tissues

```
whole_blood_row_donor_top10[,c(1,3)]
```

```
##                                     proportion    tissue
## Muscle - Skeletal                    0.8543046    Muscle
## Skin - Sun Exposed (Lower leg)       0.7523179      Skin
## Artery - Tibial                      0.7218543    Artery
## Adipose - Subcutaneous               0.7099338   Adipose
## Thyroid                              0.6913907   Thyroid
## Nerve - Tibial                       0.6675497     Nerve
## Skin - Not Sun Exposed (Suprapubic)  0.6397351      Skin
## Lung                                 0.6185430      Lung
## Esophagus - Mucosa                   0.6000000 Esophagus
## Adipose - Visceral (Omentum)         0.5854305   Adipose
```

2. List of the top 10 tissues that share most of its genes with Whole Blood

```
whole_blood_row_gene_top10[,c(1,3)]
```

```
##                                  proportion               tissue
## Spleen                            0.9839218               Spleen
## Lung                              0.9833064                 Lung
## Small Intestine - Terminal Ileum  0.9787676      Small Intestine
## Testis                            0.9741519               Testis
## Colon - Transverse                0.9703054                Colon
## Prostate                          0.9694592             Prostate
## Minor Salivary Gland              0.9684591 Minor Salivary Gland
## Breast - Mammary Tissue           0.9671513               Breast
## Stomach                           0.9671513              Stomach
## Vagina                            0.9669975               Vagina
```

In short, we summarise the ‘Proportion of Overlapping Donors of all Tissues with Whole Blood’ and ‘Proportion of Overlapping Genes of all Tissues with Whole Blood’ barplots and concise them to a list.

The tissue(s) that appear in both lists are as follows:

```
## [1] "Lung"
```

The tissue Lung is found to be the most similar to blood based proportion of on overlapping donors and proportion of shared genes, amongst all 54 tissues. This gives us a good starting point for choosing which tissues we can use for our initial analysis.

Conclusion

The heatmaps created are a great starting point to explore the data dealt with in this study. It gives us a clear-cut picture based on the count and proportion of overlapping donors and shared genes between all combinations of tissues. The most useful observations are those that have a very high count and high proportion of donors and genes, with other tissues. Those tissues can be taken for further analysis.

From the various bar plots created, we can observe the many similarities that each of the tissues possess in comparison to Whole Blood. For example, some tissues like the Spleen which shares a high proportion of genes with Whole Blood. Since they share a high number of genes.

This task projects the similarities that can be observed among many sets of tissues, based on the count of donors and genes. It helps us identify which tissues are likely to behave in the same manner, based on these two factors.

Overall, this task not only allows for exploration of the dataframes involved, but also sets a premise to explore other measures of similarity between tissues and genes, like correlation.