Task 1 - Data Science Project

Kartika Waluyo & Vrinda Rajendar Rajanahally

1000555 & 1129446

R Markdown

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 genes between all pairs of tissues.
4. Shared gene proportion: matrix that represents the proportion of overlapping genes between all pairs of tissues.

Each of the above matrices are visualized using heatmaps.

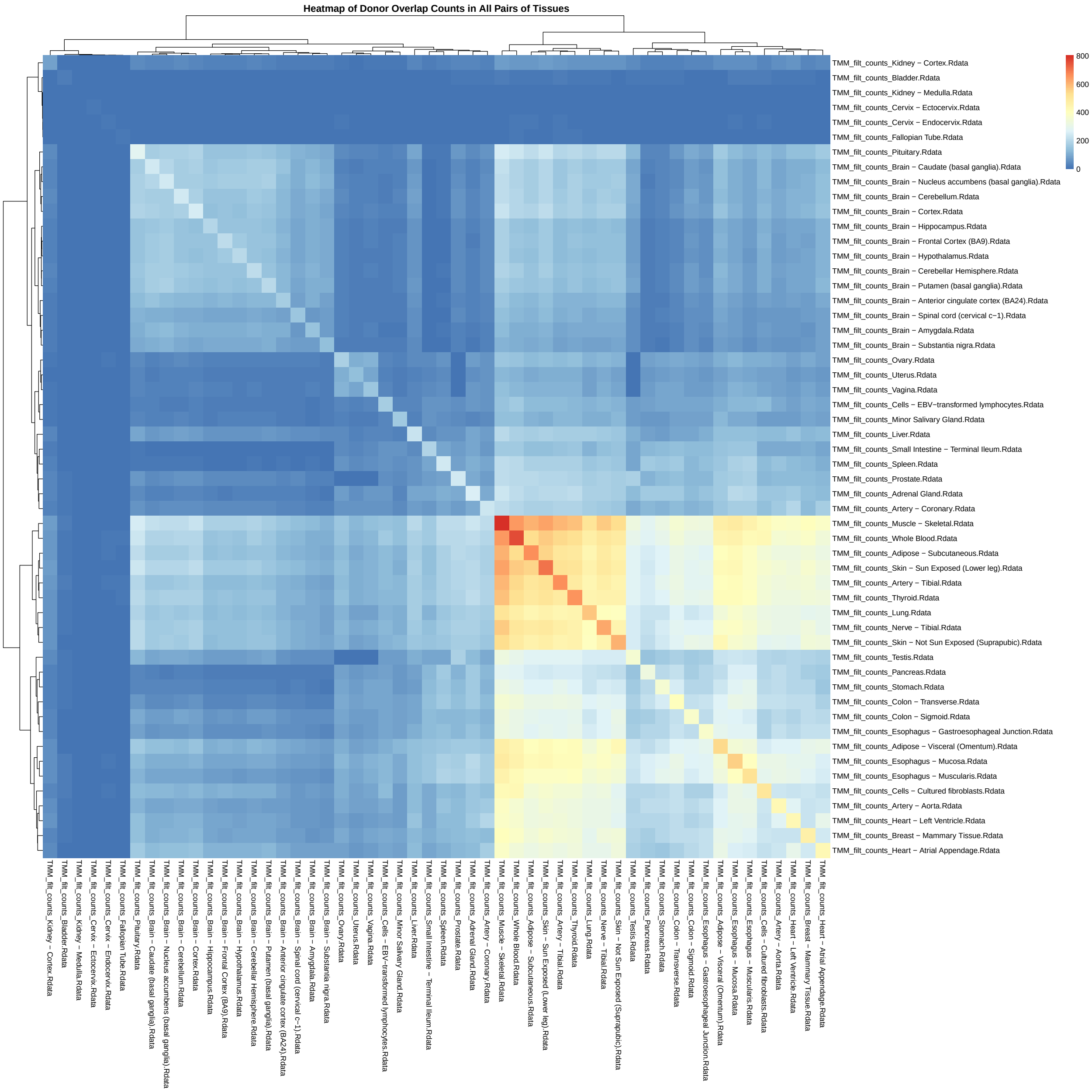
```
## Loading required package: edgeR
```

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

The count of overlapping donors between all pairs of tissue samples

The matrix 'tissue_donor_count' 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 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.



The above heatmap is a visual representation of the 'tissue_donor_count' matrix. It mirrors the matrix perfectly - the rows and columns represent all 54 tissues. With the help of the colour gradient (legend), it is easy to determine the maximum and minimum count of overlapping donors over all combinations of the tissues. Since each element in the 'tissue_donor_count' matrix is represented by a colour in the heatmap, it is easy to identify the tissue pairs with the least and most overlapping donors.

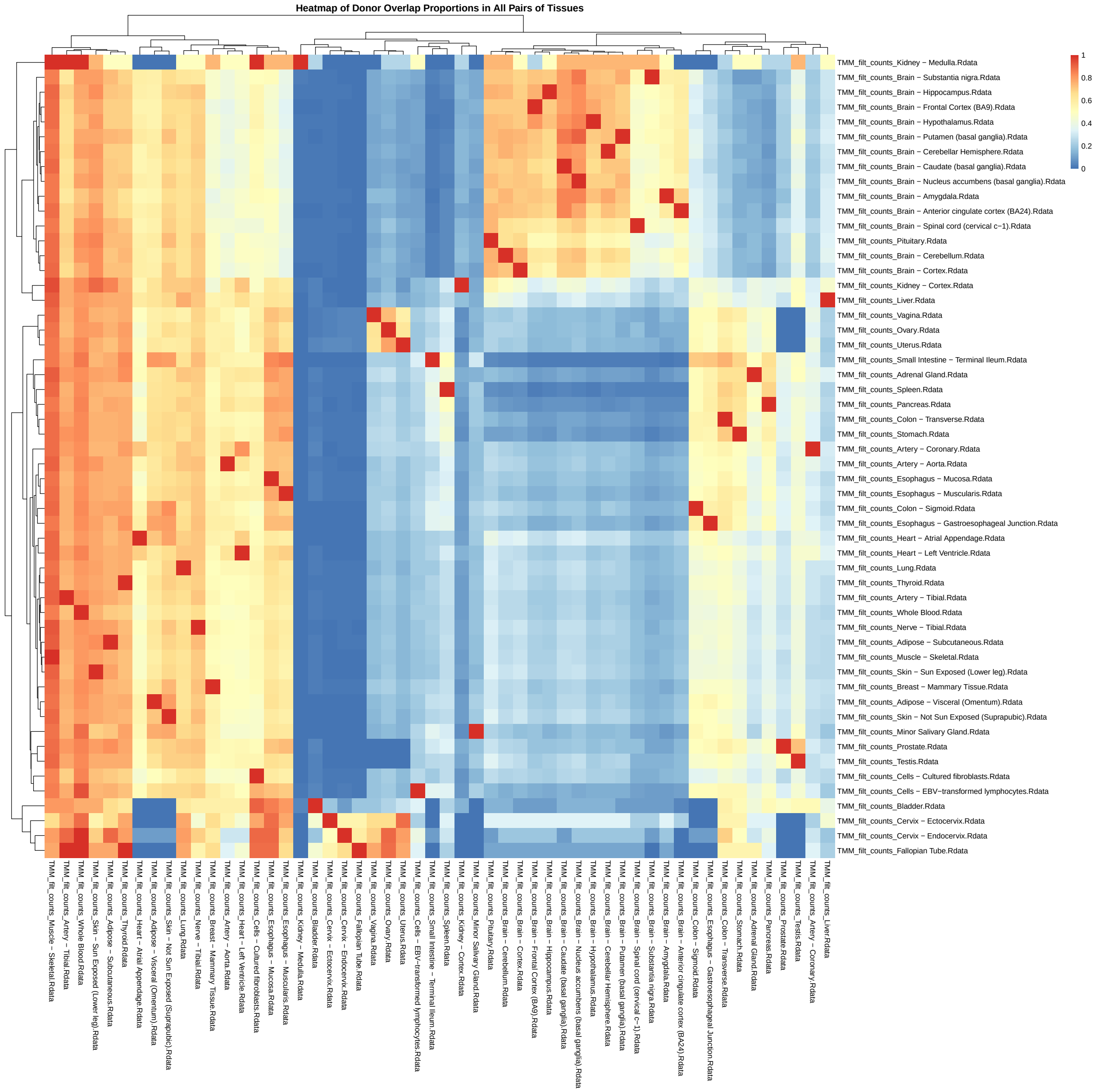
From this heatmap, Whole Blood and Muscle - Skeletal is one example of two tissues that have a high count of overlapping donors. On the other hand, Kidney - Medulla and Pancreas have an extremely low count of overlapping donors. In this visualisation, the diagonal stands out as it gives the total count 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, this heatmap has clustered similar tissues close to each other in this visualisation. A portion of the bottom right quadrant (starting from Muscle - Skeletal row and column) is a cluster that has the most count of overlapping donors. The rest of the heatmap has clustered tissues with average to low count of overlapping donors.

The proportion of overlapping donors between all pairs of tissue samples

In this task, the matrix ‘tissue_donor_proportion’ 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.

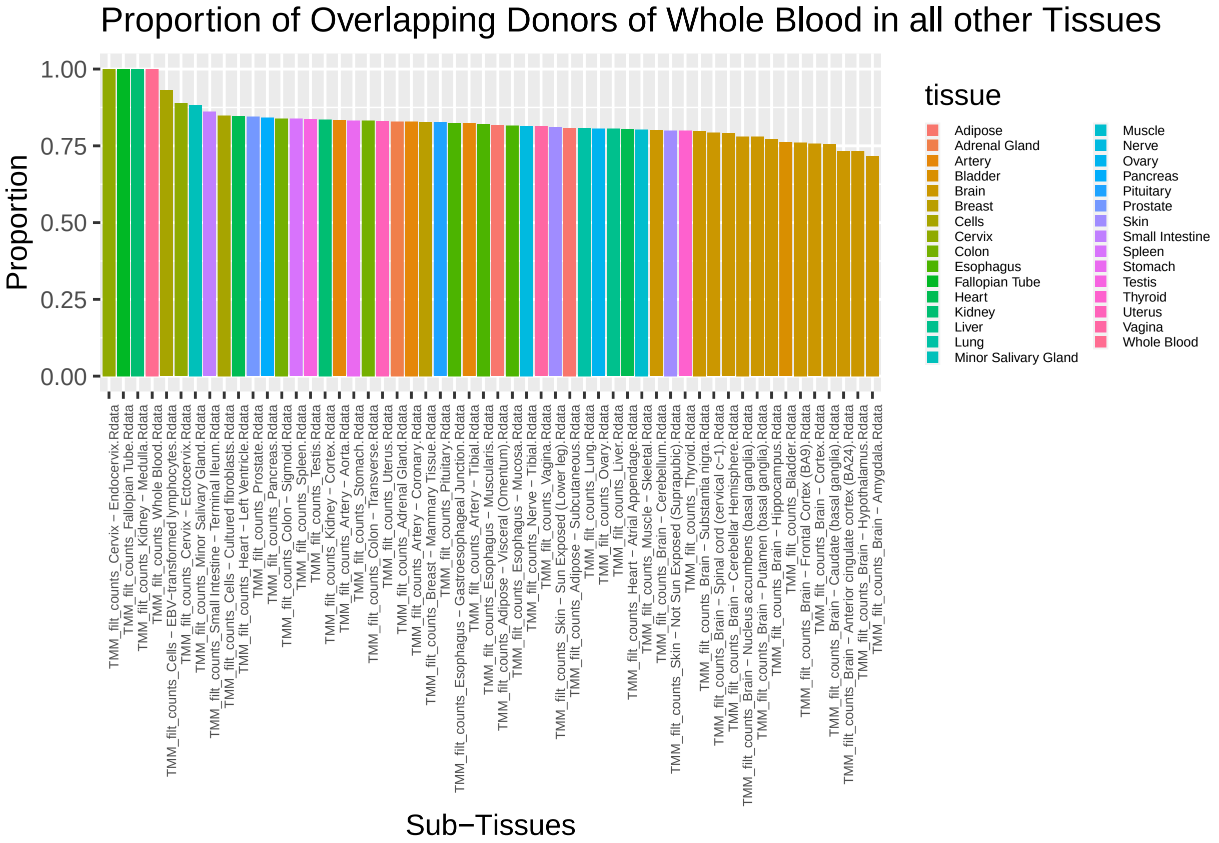
In this matrix, it can be observed that all diagonal elements will be equal to 1 since those cells represent the total proportion of donors per tissue. However, this matrix is non-symmetrical because every non-diagonal element corresponds to a different row tissue and a different column tissue while calculating the proportion per element. Each of them take a value between 0 and 1 (0 and 1 included) and can also 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.



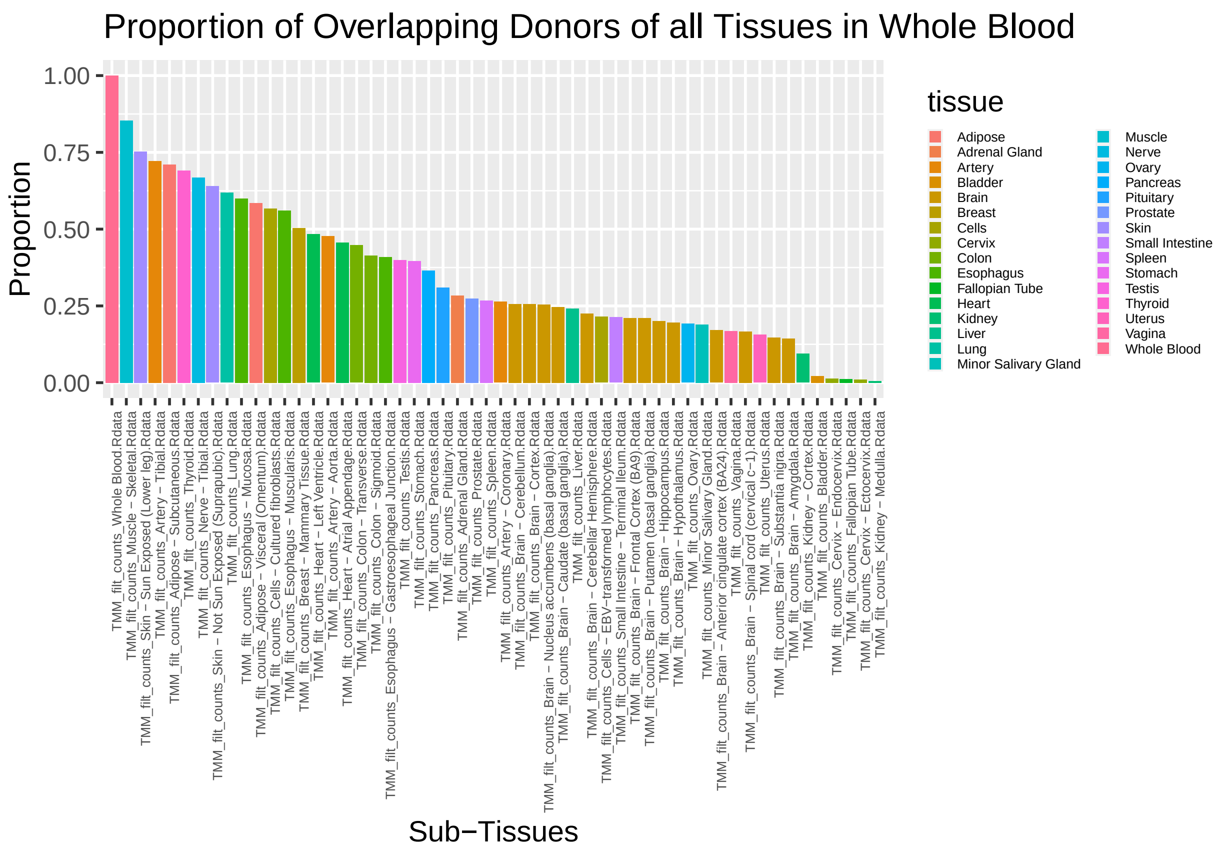
This heatmap gives a graphical representation of the 'tissue_donor_proportion', in which all rows and columns represent all 54 tissues. In this case, the colour gradient is useful to determine the proportion of each tissue to another tissue. Every cell that is coloured red denotes full proportion - which is the proportion of the tissue to itself. The other cells are likely to be values lesser than 1.

According to the heatmap, the proportion of overlapping donors between Whole Blood and Liver is very high. Subsequently, the proportion of overlapping donors between Cervix - Endocervix and Prostate is very low.

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 other tissues.



This bar plot visualises the proportion of overlapping donors of Whole Blood in all other tissues. This plot helps us identify that all donors of Cervix - Endocervix, Fallopian Tube, Kidney - Medulla are also donors of Whole Blood.

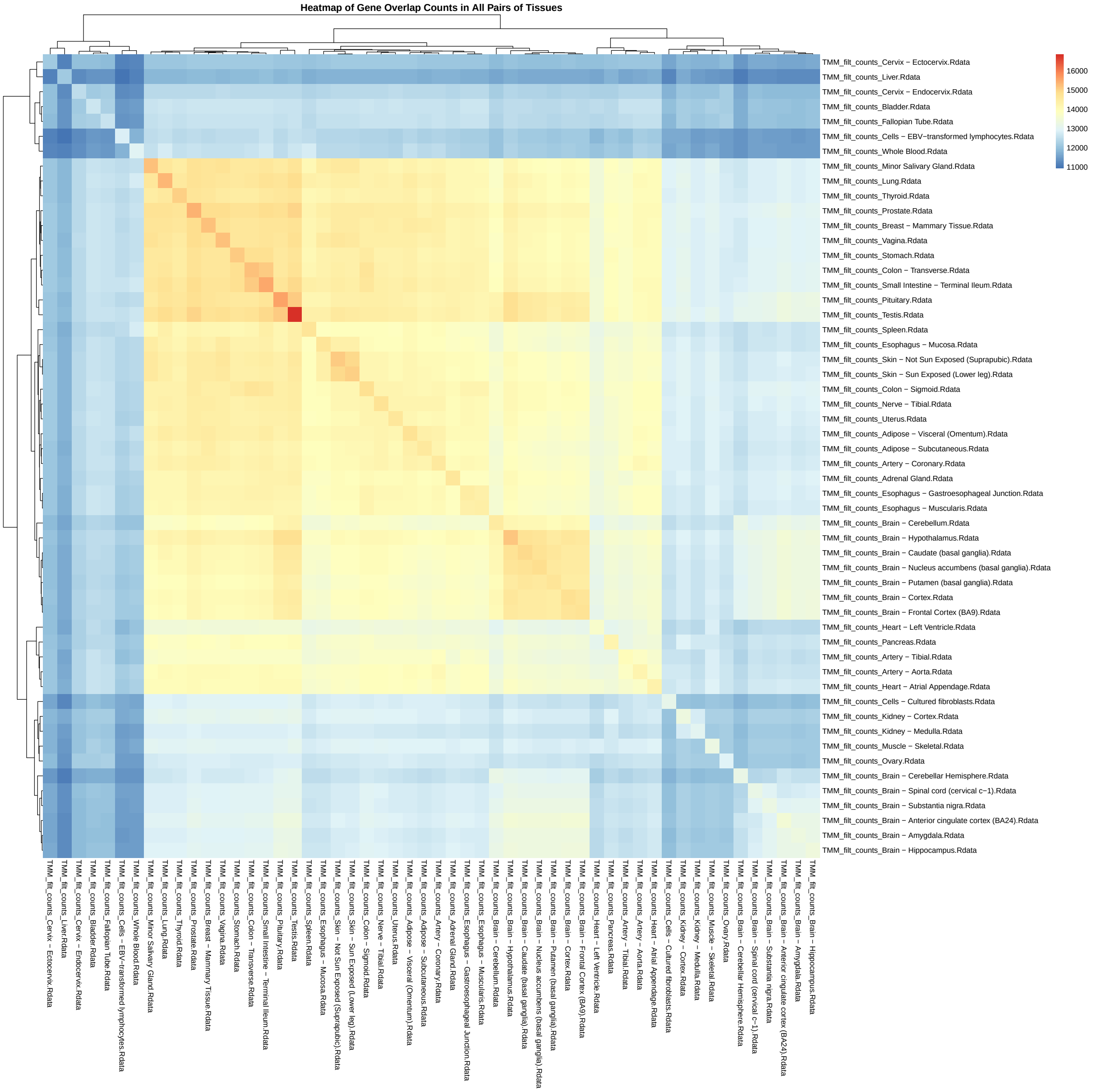


This bar plot depicts the proportion of overlapping donors of all tissues in Whole Blood. Muscle - Skeletal has the highest proportion of overlapping donors with Whole Blood whereas, Kidney - Medulla has the least. This could probably be because of the overall count of Kidney - Medulla donors is very low.

The count of shared genes between all pairs of tissue samples

The 'shared__gene__count' matrix records the count of shared genes across different pairs of tissues. It is a 54 x 54 square matrix, in which the row and column names correspond to the tissue names.

It is observed 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.



This heatmap is a visual representation of the ‘shared_gene_count’ matrix, in which the rows and columns represent all 54 tissues. With the help of the colour gradient (legend), it is easy to determine the maximum and minimum count of shared genes over all combinations of the tissues. Since each element in the ‘shared_gene_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.

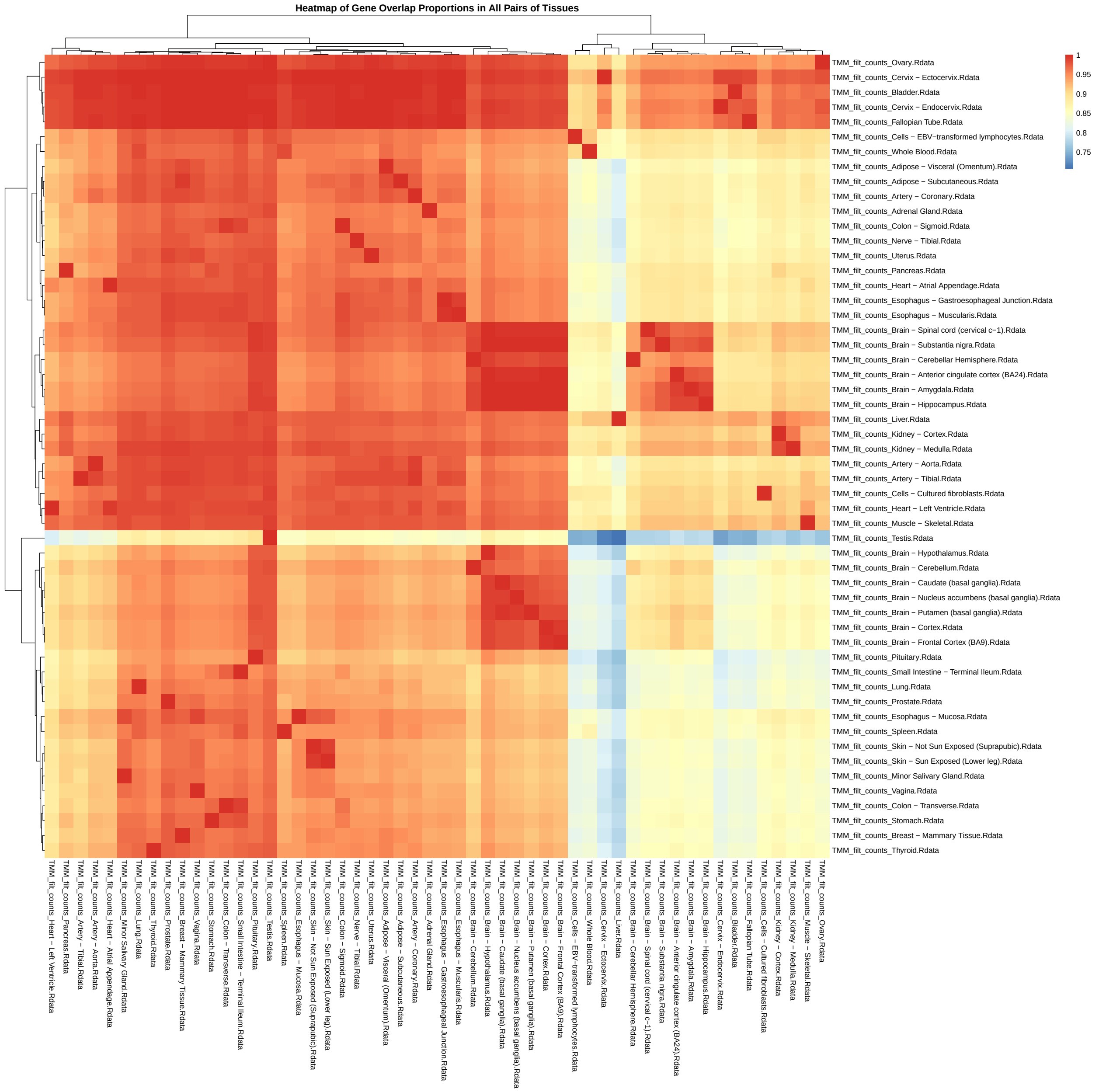
From this heatmap, Brain - Cortex and Brain - Frontal Cortex (BA9) is one example of two tissues that have a high count of shared genes. Additionally, Liver and Brain - Cerebellar is an example of a tissue pair that has an extremely low count of shared genes. The diagonal in the heatmap stands out since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

As for the clustering, this heatmap has clustered similar tissues close to each other in this visualisation, according to count of shared genes. The middle portion of this heatmap, coloured in shades of red and orange 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

The last matrix, 'shared_gene_proportion' records the proportion of shared genes between different pairs of tissues. It is a 54 x 54 square matrix, in which the row and column names correspond to the tissue names.

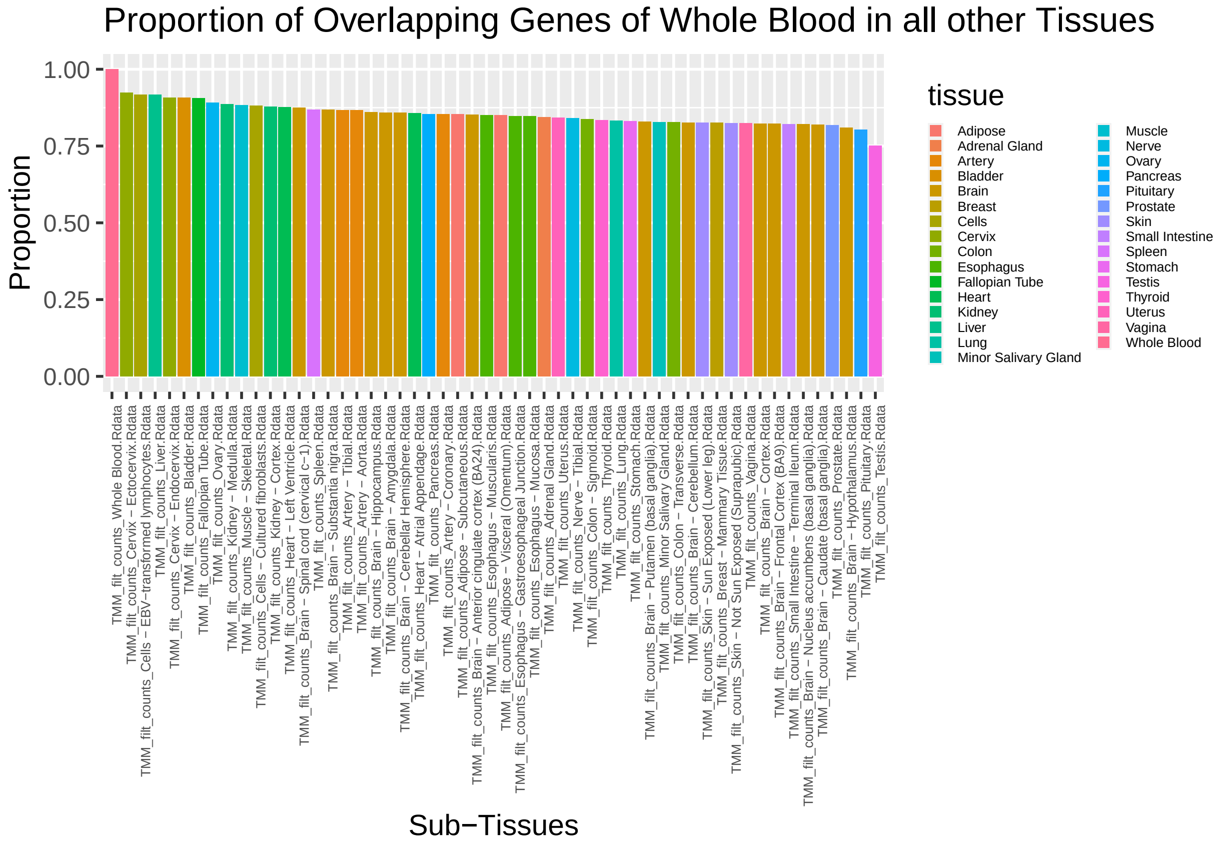
The diagonal elements of this matrix will all be equal to 1, as they represent the total proportion of genes per tissue. This matrix will also be non-symmetrical since each of the non-diagonal elements corresponds to a different row tissue and a different column tissue, while estimating the proportion. They can be expressed in a similar manner, just like the 'tissue_donor_proportion' matrix. For instance, the proportion of shared genes between Liver and Whole Blood is 0.856. In terms of percentage, overlapping genes between Liver and Whole Blood make up 85.6% of the total amount of genes found in Whole Blood.



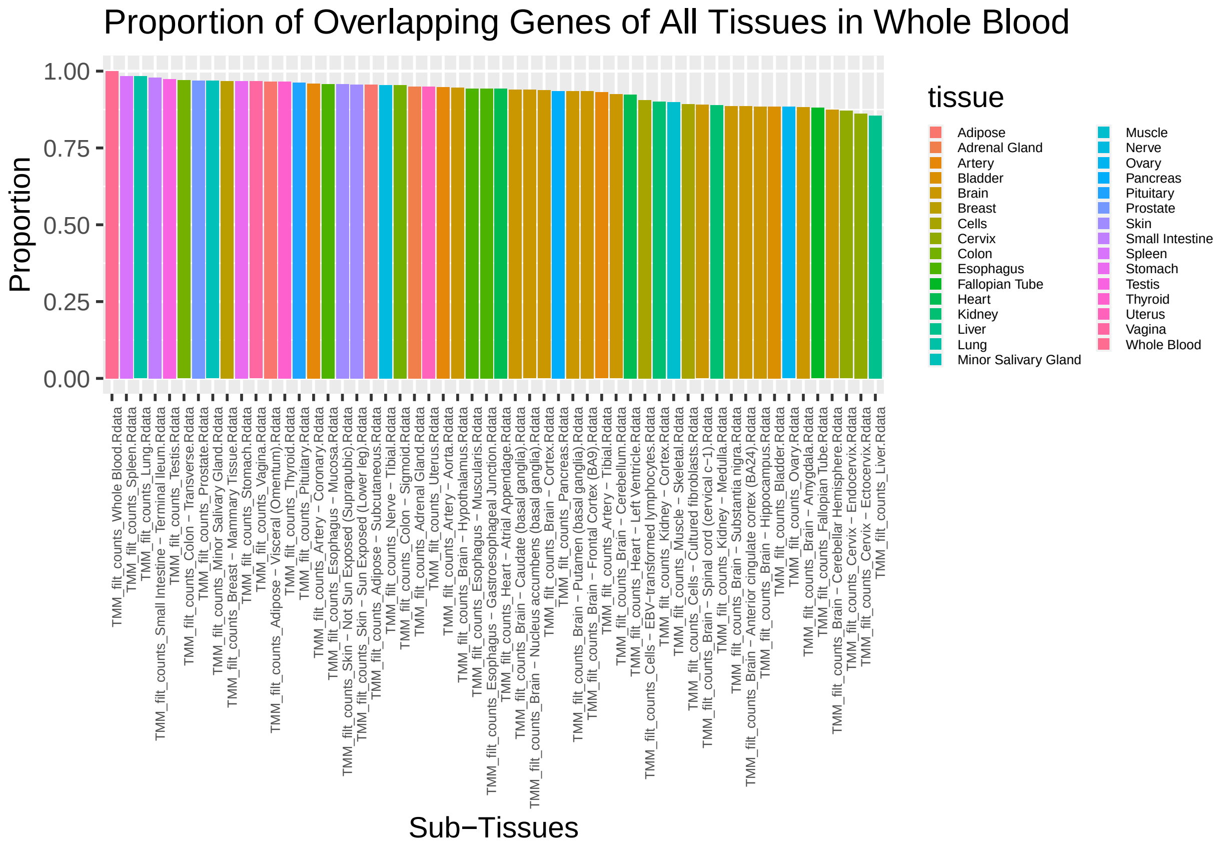
This heatmap gives a graphical representation of the 'shared_gene_proportion', in which all rows and columns represent all 54 tissues. In this case, the colour gradient is useful to determine the proportion of each tissue to another tissue. Every cell that is coloured red denotes full proportion - which is the proportion of genes of the tissue to itself. The other cells are likely to be values lesser than 1.

According to the heatmap, the proportion of shared genes between Stomach and Bladder is very high. Subsequently, the proportion of shared genes between Testis and Liver is very low.

Coming to clustering, all similar gene proportions have been clustered together. 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.



The above bar plot is a visualisation of all overlapping genes of Whole Blood in other tissues. Here, we observe that the proportion of shared genes in Whole Blood, that are also found in Testis is the least. Whole Blood and Cervix - Ectocervix has the highest proportion of shared genes in this case.



This bar plot shows the proportion of shared genes of all tissues in Whole Blood. 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.

Conclusion

From the various bar plots above, there are many tissues like the Spleen and Ectocervix that share a high proportion of genes with Whole Blood. It can be deduced that the Spleen and Ectocervix have similar characteristics of Whole Blood.

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