Data Science Project Pt. 1 (MAST90106)

Final Report, Semester 1

Kartika Waluyo, 1000555 Vrinda Rajendar Rajanahally, 1129446

Contents

1 Introduction

Gene expression is measured through RNA sequencing and indicates the "activation status" of all genes in a tissue. This is widely used in medical research to understand disease mechanisms and assess the effectiveness and safety of new treatments. Gene expression varies across tissues, but obtaining samples from heart, lung, brain and other organs is often not all that easy.

This project aims to develop a machine learning algorithm that can reliably predict gene expression in many different tissues when only the measurements of expression in blood are available. A proof of concept for this approach has been attempted by Basu et al, but is not compatible with unseen data (Basu et al., 2021). This project will start with the analysis of a publicly available dataset (GTEx) containing gene expression of 54 post-mortem tissues from almost 1,000 individuals to develop algorithms that produce a body-wide gene activation estimate from blood alone.

This project aims to develop an AI algorithm that reliably predicts gene expression in as many different tissues as possible, from blood. In the case of a successful outcome, the algorithm will potentially be published in a peer-reviewed journal and a web tool will be released to enable other researchers and companies to add value to their blood RNAseq data.

This report will cover the team's initial analysis of the data, which involves many tasks like creating different visualisations to understand how similar tissues are based on shared donors and shared genes, statistically evaluating those similarities. It also covers our approach to choosing the appropriate models for the data.

2 Background Knowledge

2.1 Introduction to Genetics

Proteins are large molecules that are complex and play critical and majority roles in the body. Proteins are encoded by genes, that are specific regions of the DNA through a process of two stages (Al Bhketan, 2020). The first stage is transcription which is the process where the DNA within a gene is transcribed into messenger Ribonucleic Acid (mRNA). The second stage is translation which is the process where the mRNA is translated into a protein.

The genetic instructions needed for cells to function are contained by the DNA. However, the cells of the same individual that contain the same genetic instructions can behave differently. That is due to the proteins, instead of the DNA of a gene, being responsible for cell functionalities. In spite of being very important, measuring the level of protein present in a cell is a hard procedure. However, it is easier to measure mRNA level which will give an estimation of protein levels. Over different cells, not all genes are expressed into mRNA, and when they do, it does not necessarily mean that they are expressed similarly. Genes are considered to be active when they are highly expressed, and inactive when they are lowly expressed (Al Bhketan, 2020).

2.2 Machine Learning

The machine learning methods we deemed appropriate to initially to explore in this project are Neural Networks and Random Forests.

2.2.1 Neural Network

Over the past few decades, neural networks modelling has been considered as one of the most powerful tools, and its ability to handle a huge amount of data made it very popular in the literature. Having deep

hidden layers in the models has recently become an interest that has started to surpass classical methods performance in many fields, especially in pattern recognition (Albawi et al., 2017).

Deep learning models have become popular in the bioinformatics field. Singh et al. (2016) used a unified CNN framework that automatically learns combinatorial interactions among histone modification marks to predict gene expression. Qi et al. (2012) used a deep multilayer perceptron (MLP) architecture with multitask learning to perform sequence-based protein structure prediction (Qi et al., 2012).

By using neural networks, the idea of setting up a lightly parameterised function shaped by human can be forgotten. Instead, it allows us to set up a highly parameterised function that is very flexible and will be conveniently shaped during the learning phase. To put it simply, a deep learning model automatically learns complex functions that map inputs to outputs and rules out the need to use hand-crafted features (Singh et al., 2016, i639–i648). Since the input data of this project has a very large feature dimension, neural network is considered as one of the suitable approaches.

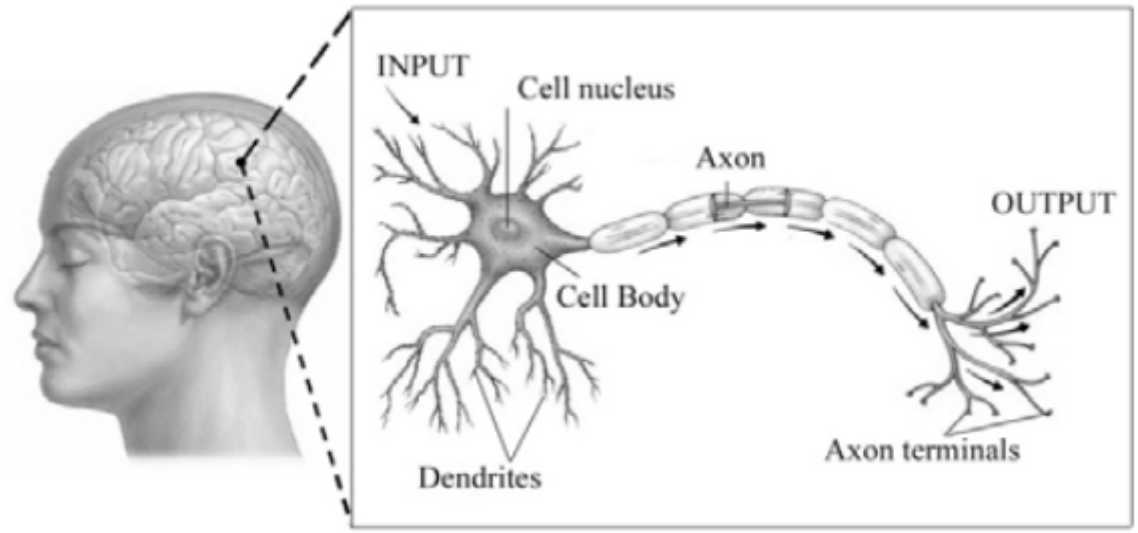

Figure 1: Neural networks in the human brain

The idea of neural network came from the most interesting organ in the human body, the brain. The human brain is made up of billions of basic units called neurons. **Figure 1** illustrates the basic neuron unit. The neuron is made up of dendrites, a cell body and an axon connecting to axon terminals. The brain will receive information or inputs which are then transferred into the cell body through dendrites. The cell body works as the processing unit, where all the learnt information is then transferred into outputs and passed down by the axon. The muscles then receive the outputs from the axon terminals for actions. McCulloch and Pitts first studied this concept in 1943 to form a mathematical model (Bakar et al., 2009).

Figure 2 shows a one hidden layer feed forward network with inputs $x_1, \ldots, x_i$, and output y_k. Each input has its own synaptic weight. The weights are then passed onto the hidden layer, which consists of several hidden neurons. A weighted summation of the inputs is performed by each neuron and then it passes a nonlinear activation function.

2.2.2 Random Forest

Over the years, random forests have become a prominently used technique in the field of biology and bioinformatics. Some fields where random forests are used are in classifying different types of samples using

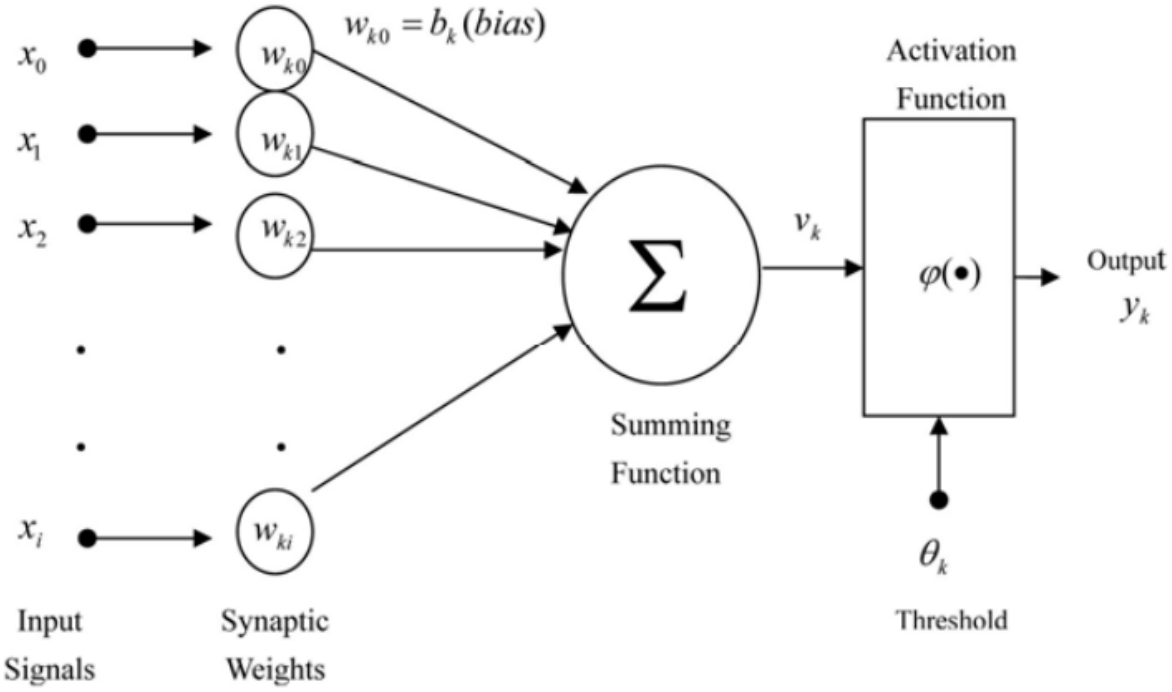

Figure 2: Hidden layers of a network

gene expression, identifying diseases associated with particular genes and others (Cutler et al., 2012).

A random forest is popularly known to be a classification algorithm that selects features randomly. It also utilises the concept of bagging samples and majority voting scheme, that make it better than decision trees (Cutler et al., 2012). Additionally, a random forest is a collection of many decision trees as seen in **Figure 3**, which is capable of classification and regression tasks. The learning process incorporates feature selection and interactions naturally (Qi, 2012).

Computational biology makes use of random forests as it works best with complex and multidimensional data (Qi, 2012). Moving to the advantages of using the random forest technique, this machine learning algorithm works best in large and high dimensional data which is very well suited in the context of our project. A large random forest is preferable as it accounts for a robust model, with better accuracy and predictive capabilities. It focuses on selecting the best variable for prediction, and can easily help with identifying variables that aren't very significant to the model. Bootstrapping and ensemble schemes prevent the model from overfitting and hence, pruning trees is not required. Another advantage is that this algorithm accounts for missing values in the data and continues to maintain accuracy (Hsueh et al., 2013).

A step further is to implement eXtreme gradient boosting or XGBoost (XGB), which is basically gradient boosted decision trees designed for speed and performance. The algorithm that is featured in XGB is that of gradient boosting, which is a popular supervised learning algorithm. This algorithm aims to predict a specific variable based on estimates of other models. With great computational and executional speed, it implements gradient boosting very fast with high efficiency and accuracy. This reflects it's modelling capabilities and it has become a highly preferential method for any regression and classification prediction problems (Chen et al., 2015).

3 Data Analysis

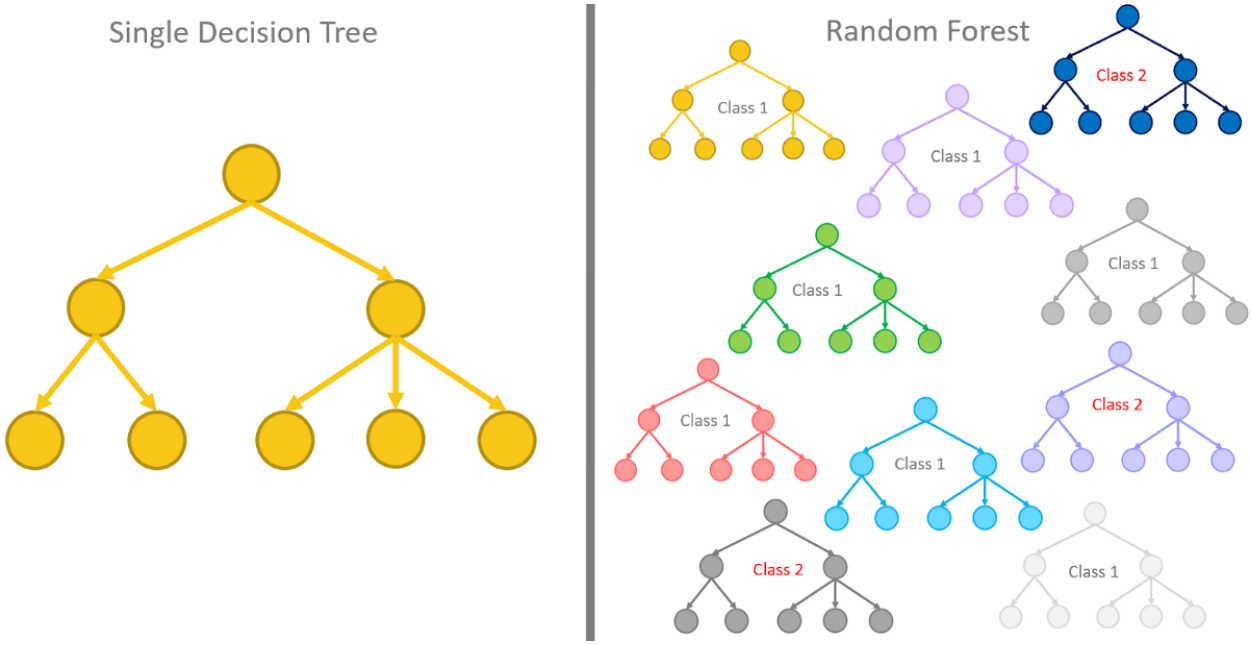

Figure 3: Single decision tree vs. Random Forest

3.1 Exploration of the Gene Expression Data and Tissue Donor Metadata

To 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.

3.1.1 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.

In this symmetrical matrix, rows and columns are clustered by similarity in terms of donor overlap. Since each element in **Figure 4** 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 4**, 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

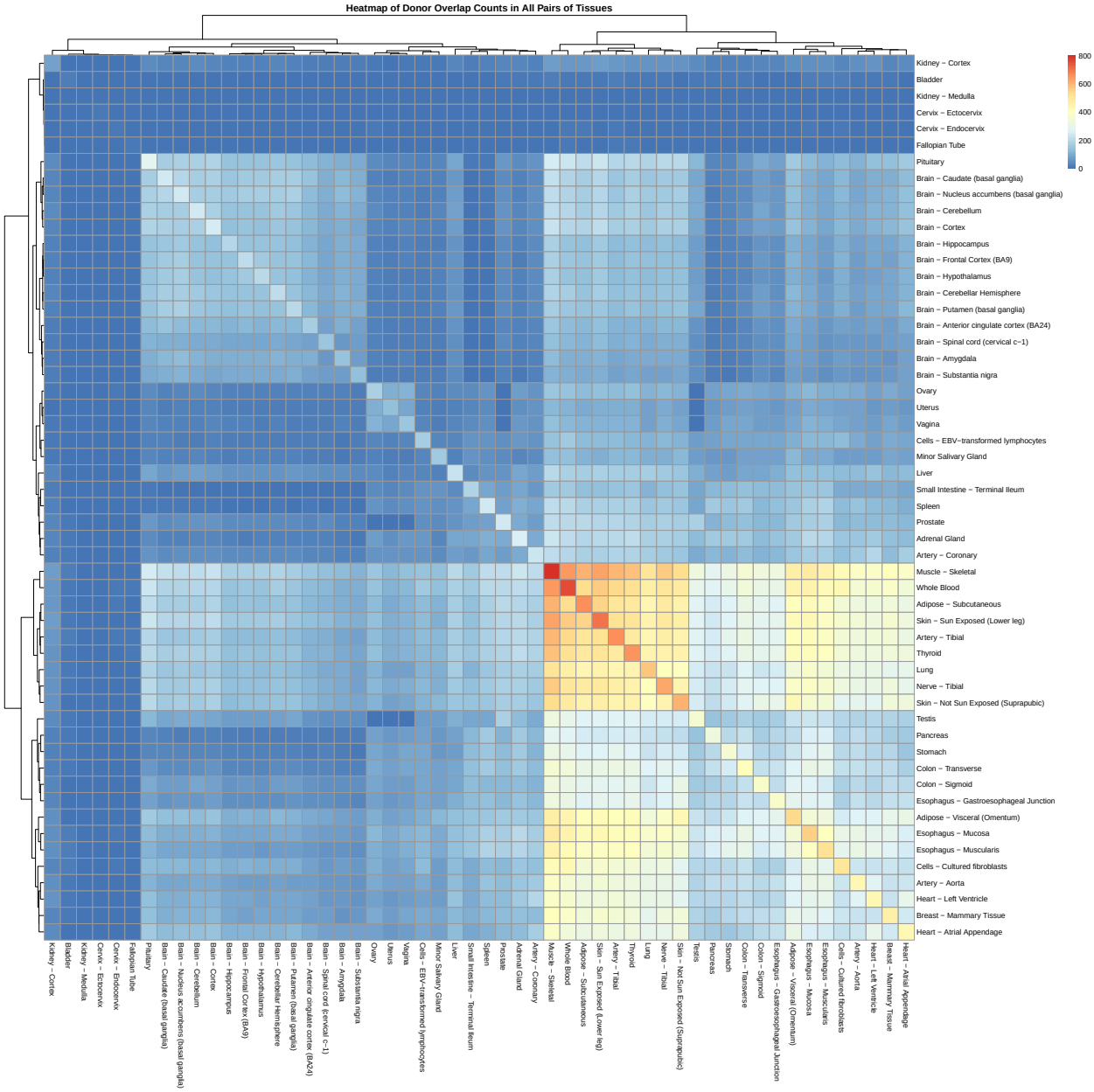

Figure 4: Heatmap representing no. of overlapping donors for each tissue pair

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.

3.1.2 The proportion of overlapping donors between all pairs of tissue samples

Here, 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 5** 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 5**, 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 7 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: **Figure 7** considers Whole Blood as the denominator.

3.1.3 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.

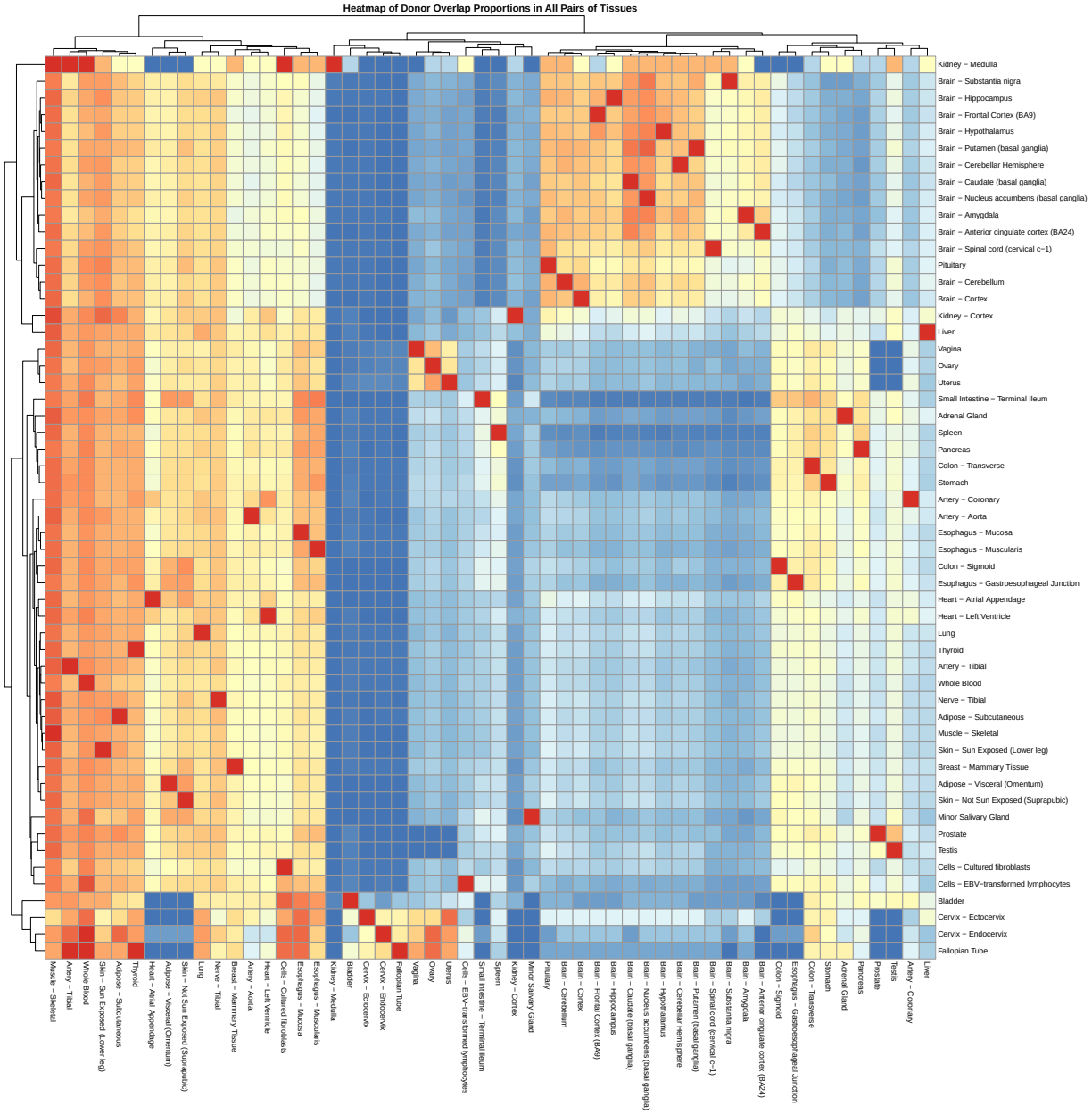

Figure 5: Heatmap representing proportion of overlapping donors for each tissue pair

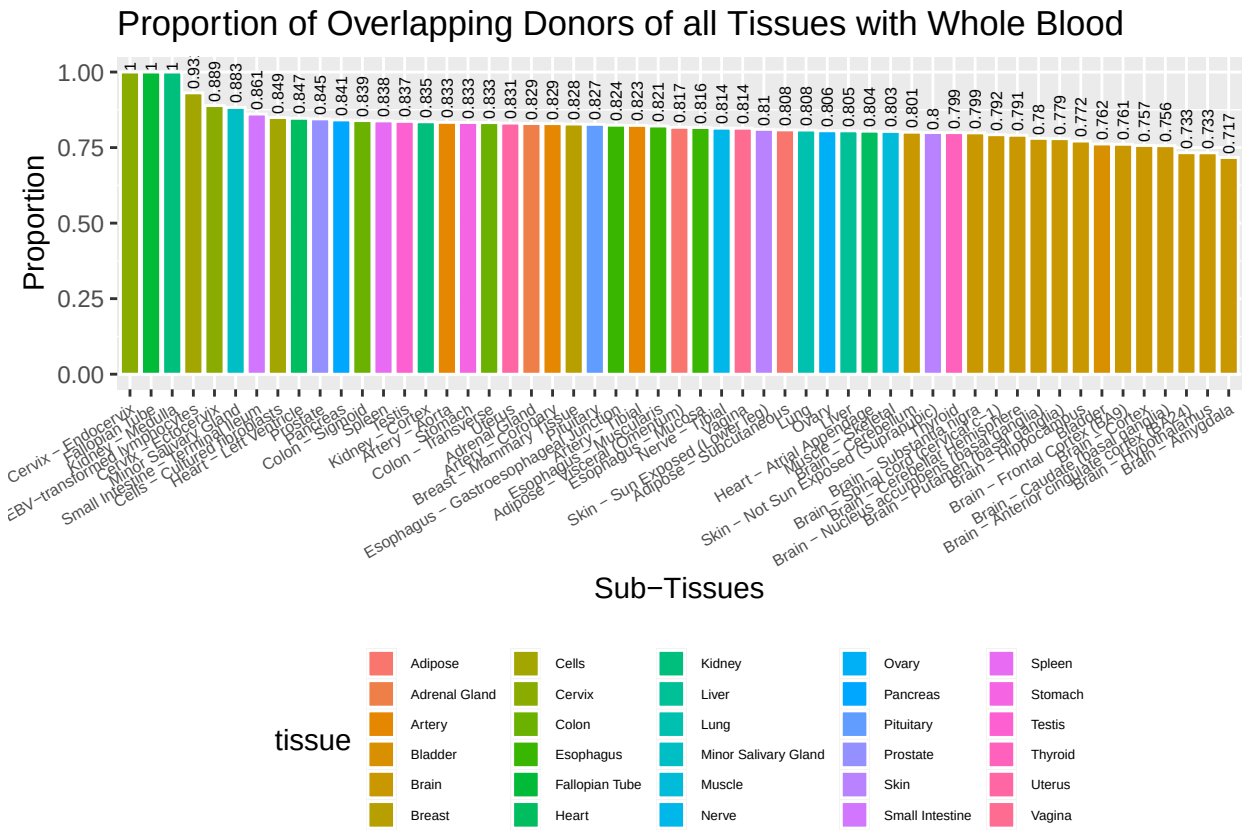

Figure 6: Bar plot representing the proportion of overlapping donors of all tissues with whole blood

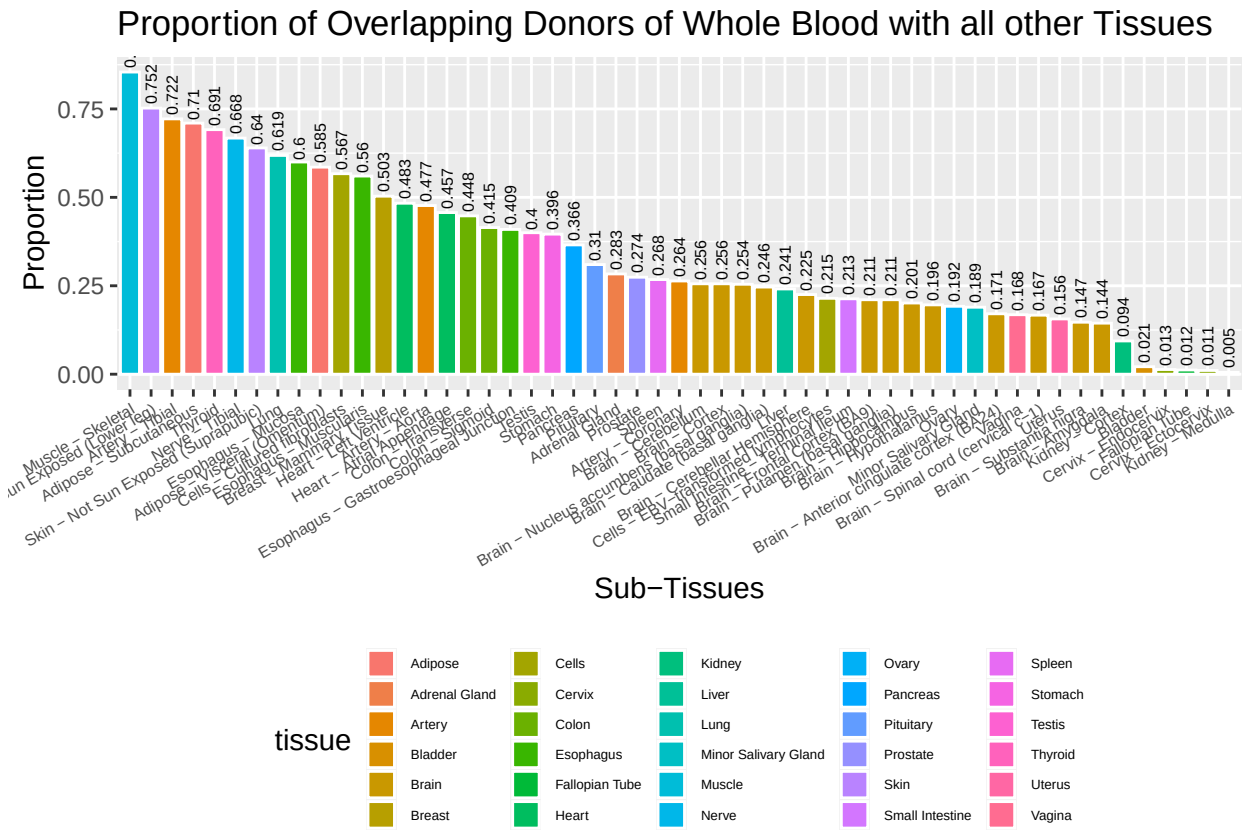

Figure 7: Bar plot representing the proportion of overlapping donors of whole blood with other tissues

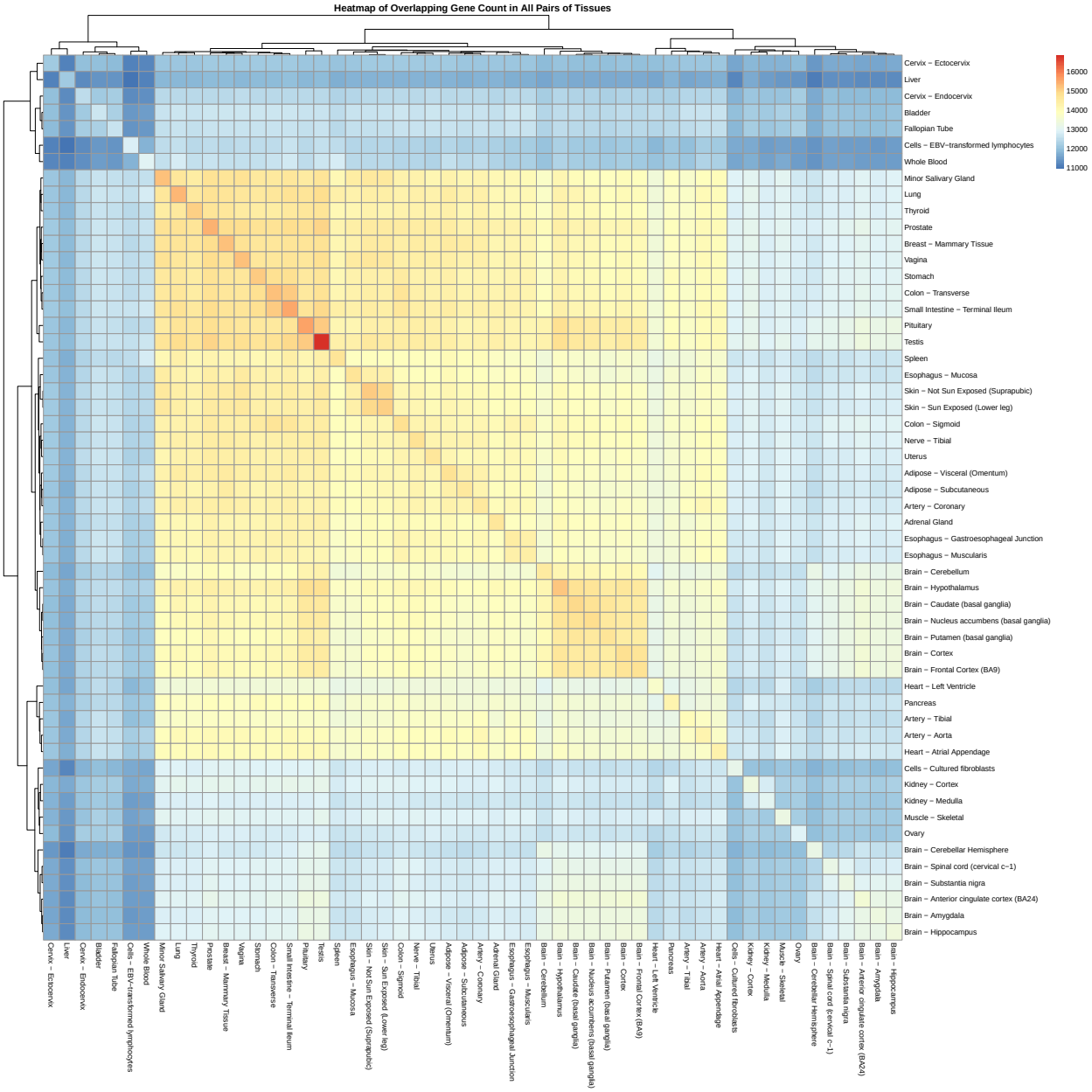

Figure 8: Heatmap representing the count of overlapping genes in all tissue pairs

Figure 8 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 8**, 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.

3.1.4 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 9 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 10 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 11 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 11** 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

```
##                                   proportion    tissue
```

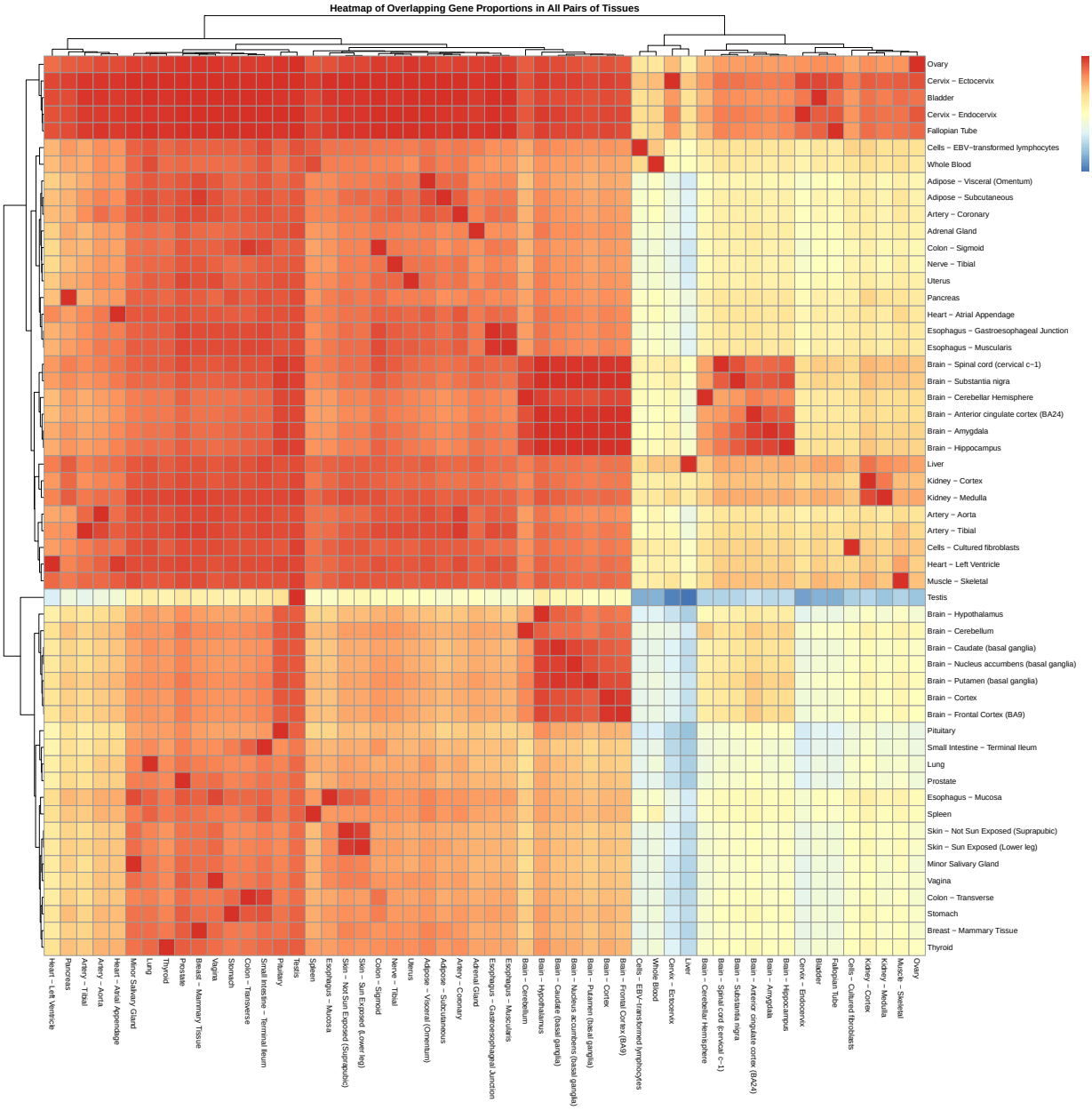

Figure 9: Heatmap representing the proportion of overlapping genes in all tissue pairs

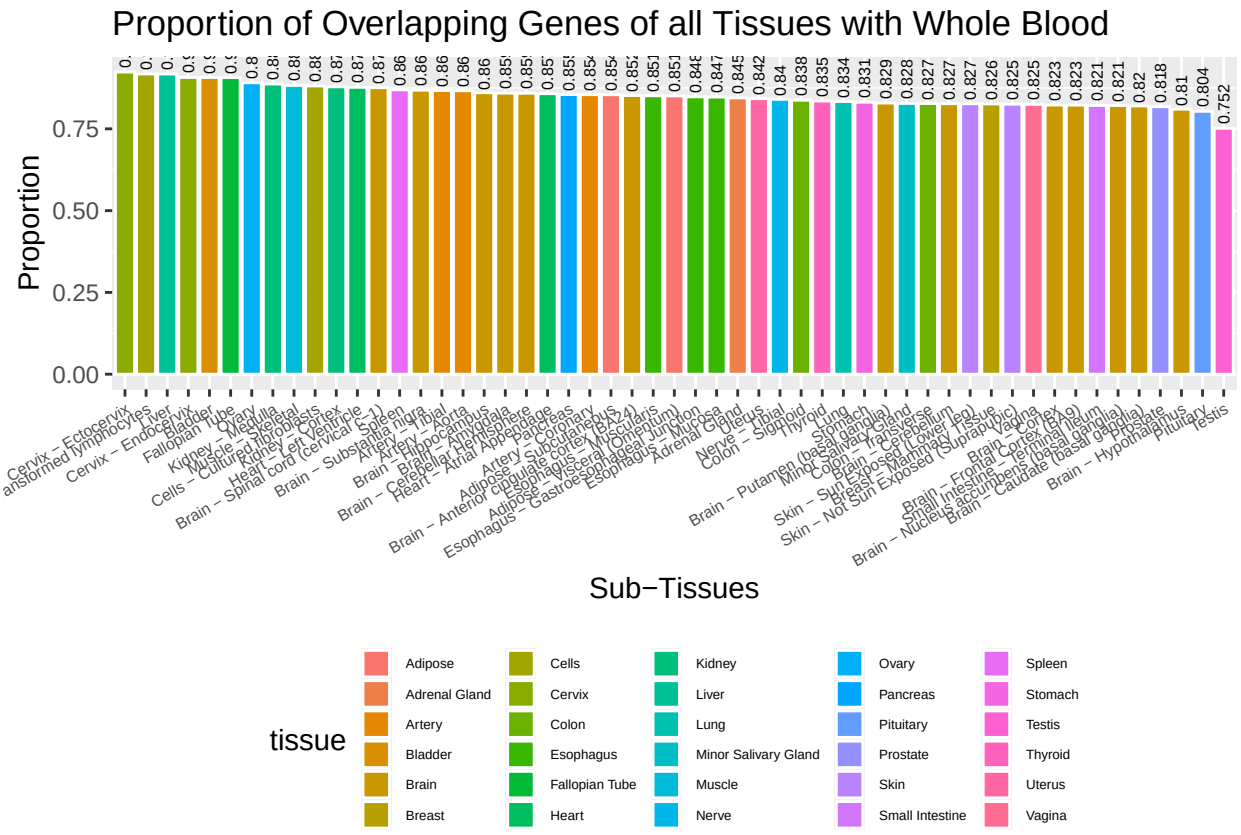

Figure 10: Bar plot representing the proportion of overlapping genes of all tissues with whole blood

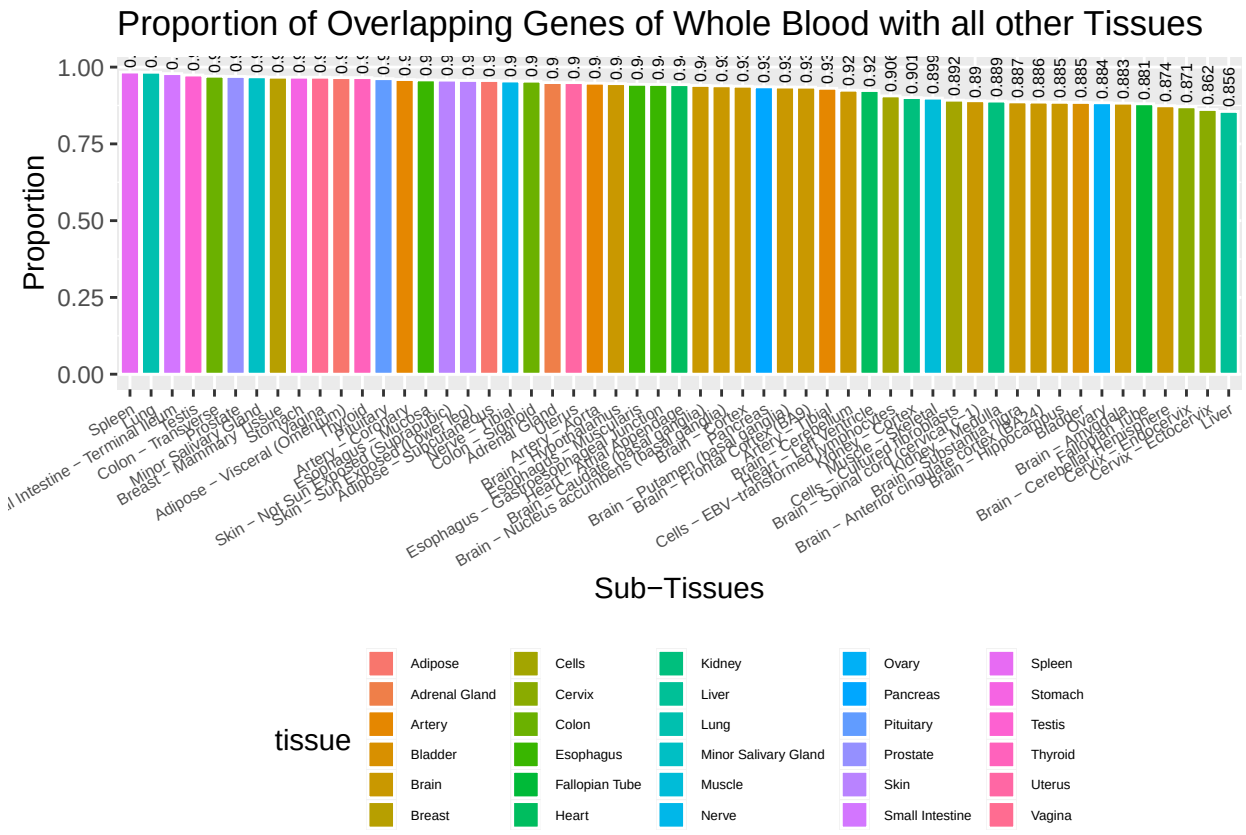

Figure 11: Bar plot representing the proportion of overlapping genes of whole blood with all other tissues

```
## 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

```
##                                  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.

3.1.5 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 section 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 section 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.

3.2 Tissues Correlations and Similarities

This section involves the exploration of pairwise correlation between each gene in whole blood's gene expressions with the same gene in other tissues' gene expressions. From there, the statistics of correlations of all genes in each tissue are computed. Additionally, the statistics of correlations of each gene across all tissues are also explored. Finally, the tissues are ranked on its similarity with whole blood on multiple measures.

3.2.1 Correlation of various tissues and with Whole Blood, based on their shared genes

We first create a vector that contains the pairwise correlation of gene x in blood with all other tissues. Mean and Median of the correlation for each tissue is found to summarise the measure of correlation across all gene expressions. We also calculate and plot the Absolute Mean and Absolute Median of correlations across the tissues and Whole Blood to understand the relationship between these a bit better.

The mean and median correlation plots are as follows:

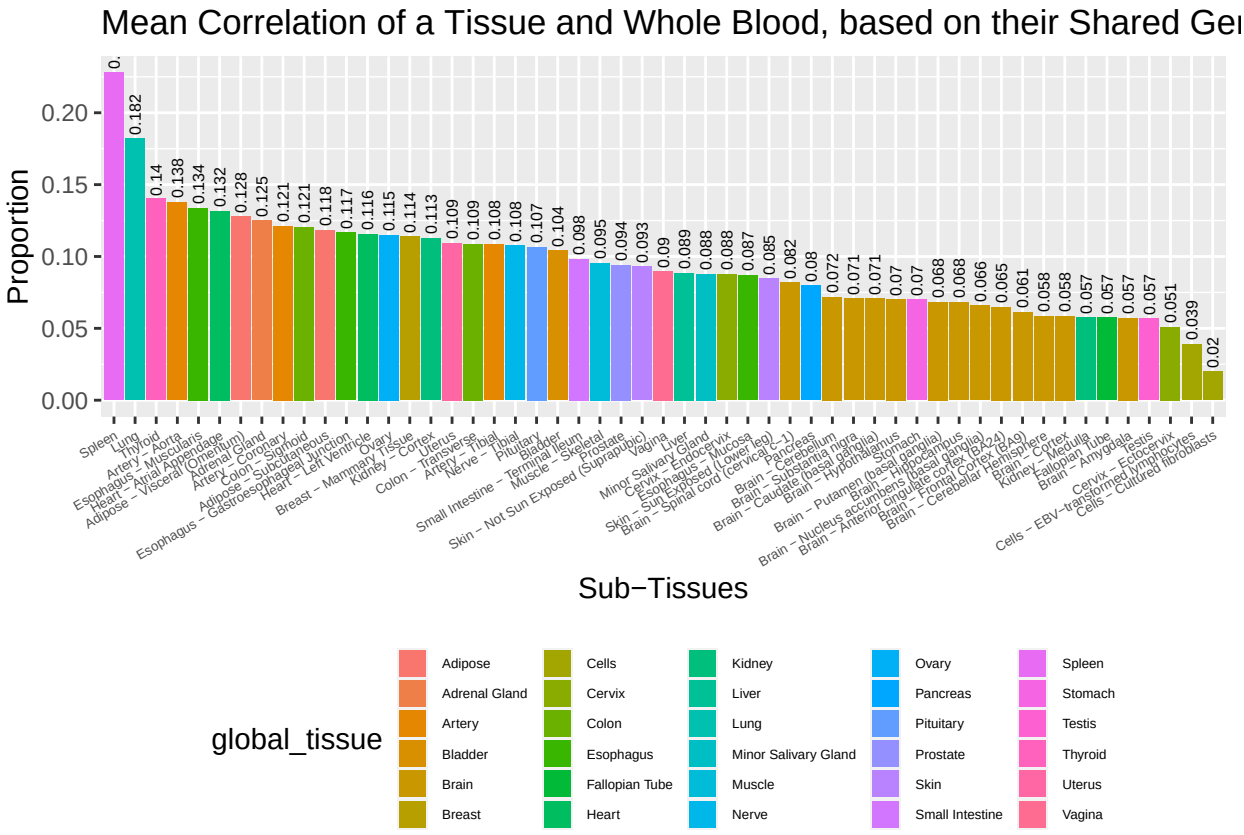

Figure 12: Mean Correlation of a Tissue and Whole Blood, based on their Shared Genes

Figure 12 captures the mean correlation between a particular tissue with Whole Blood, based on their shared genes.

Figure 13 captures the median correlation between particular tissue and Whole Blood, based on their shared genes.

The mean correlation and median correlation can help us to predict the how high or low the count of a particular shared gene in a tissue will be, when the expression of the same gene is known in Whole Blood.

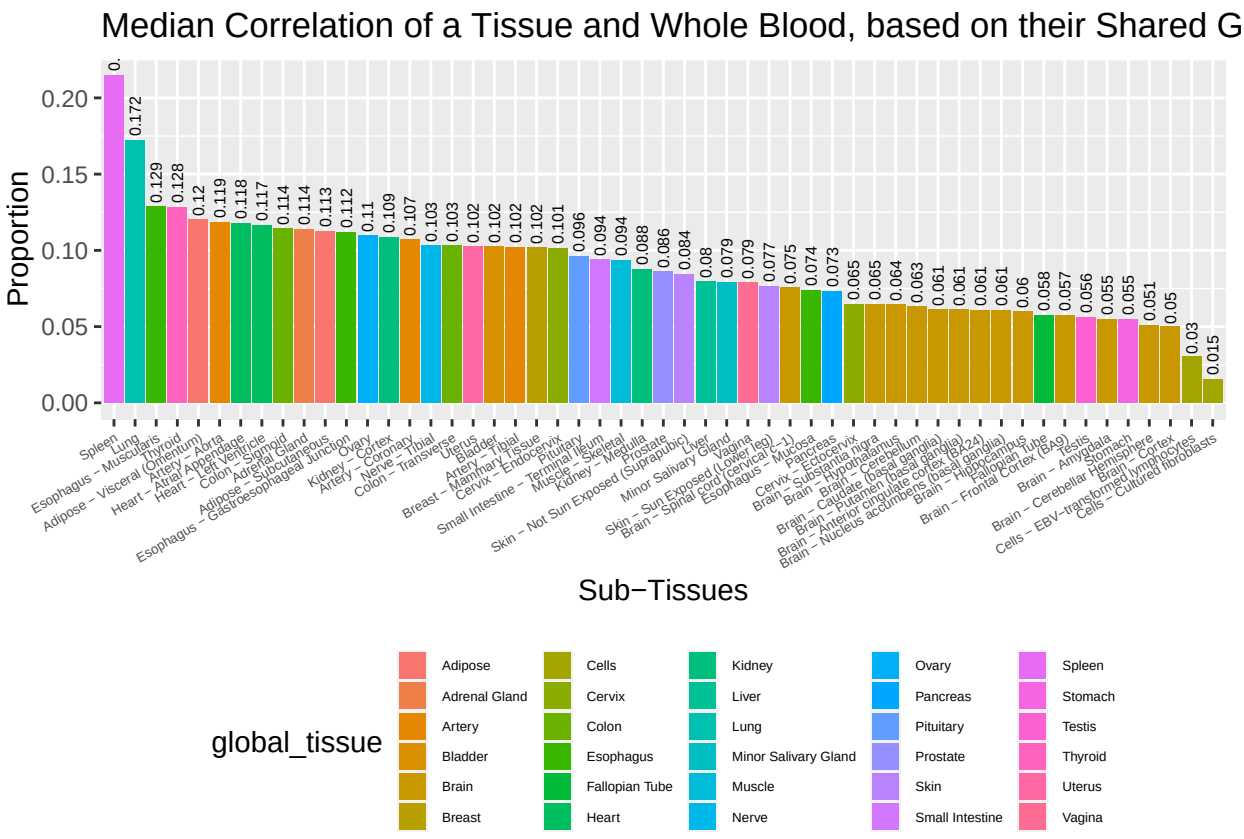

Figure 13: Median Correlation of a Tissue and Whole Blood, based on their Shared Genes

It must be considered as just an approximation of how high or low the count will be, and not as a determined factor.

According to **Figure 12** and **Figure 13**, we see that Spleen has the highest mean and median correlation with Whole blood. 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.

Next, to understand the strength of the relationship between the tissue and Whole Blood we calculated summary statistics for the absolute correlation.

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

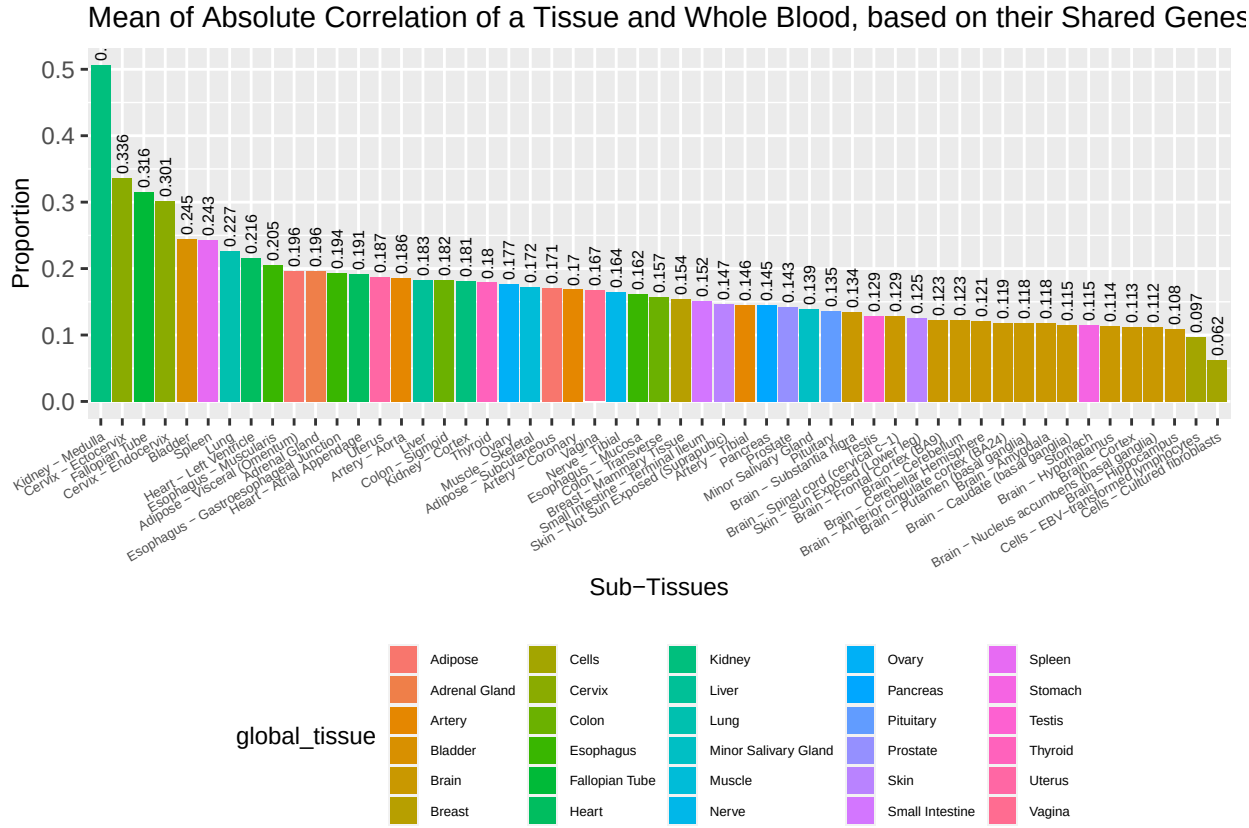

Figure 14: Mean of Absolute Correlation of a Tissue and Whole Blood, based on their Shared Genes

Figure 14 represents the mean of absolute correlation between a tissue and Whole Blood, based on their shared genes.

Figure 15 captures the median of absolute correlation between a tissue and Whole Blood, based on their shared genes.

In the context of this section, 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.

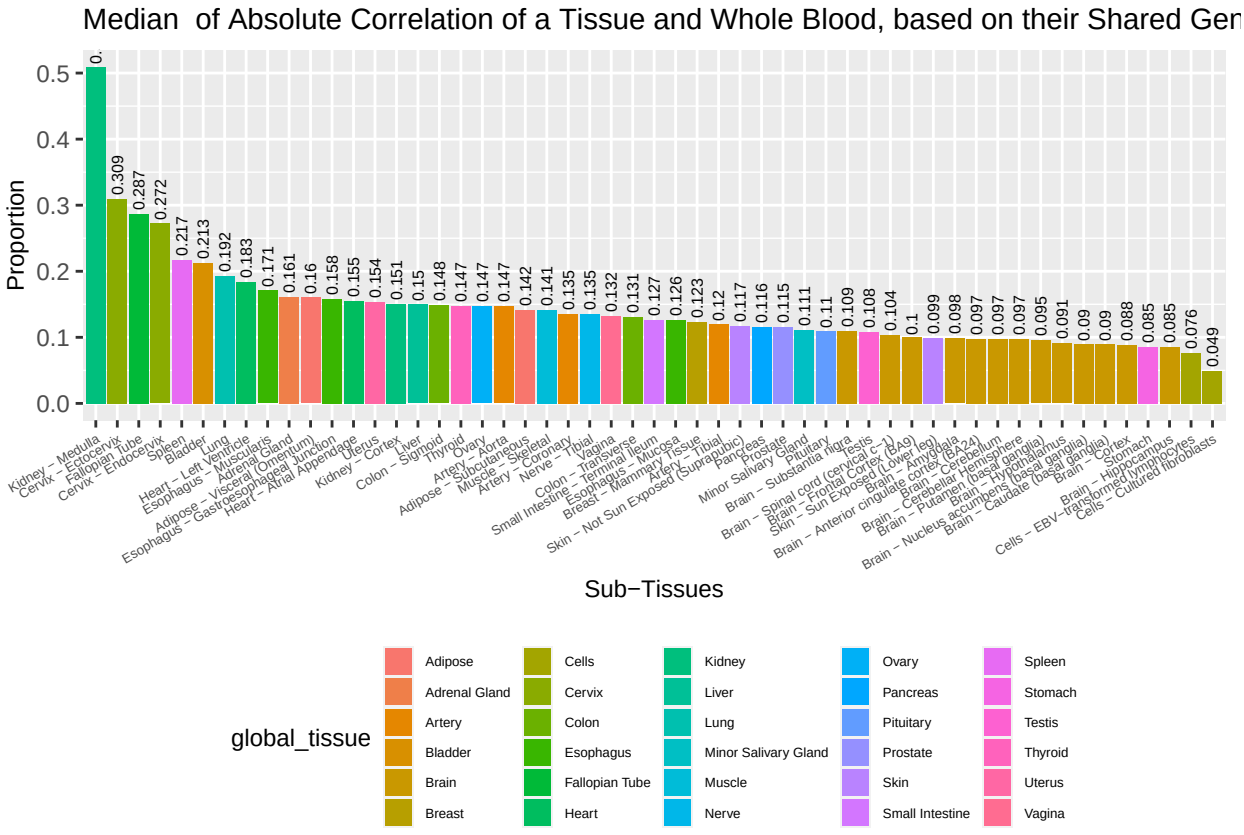

Figure 15: Median of Absolute Correlation of a Tissue and Whole Blood, based on their Shared Genes

From **Figure 14** and **Figure 15**, 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 tissues in the brain 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.

3.2.2 Correlation of each gene found in Whole Blood across all other tissues

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

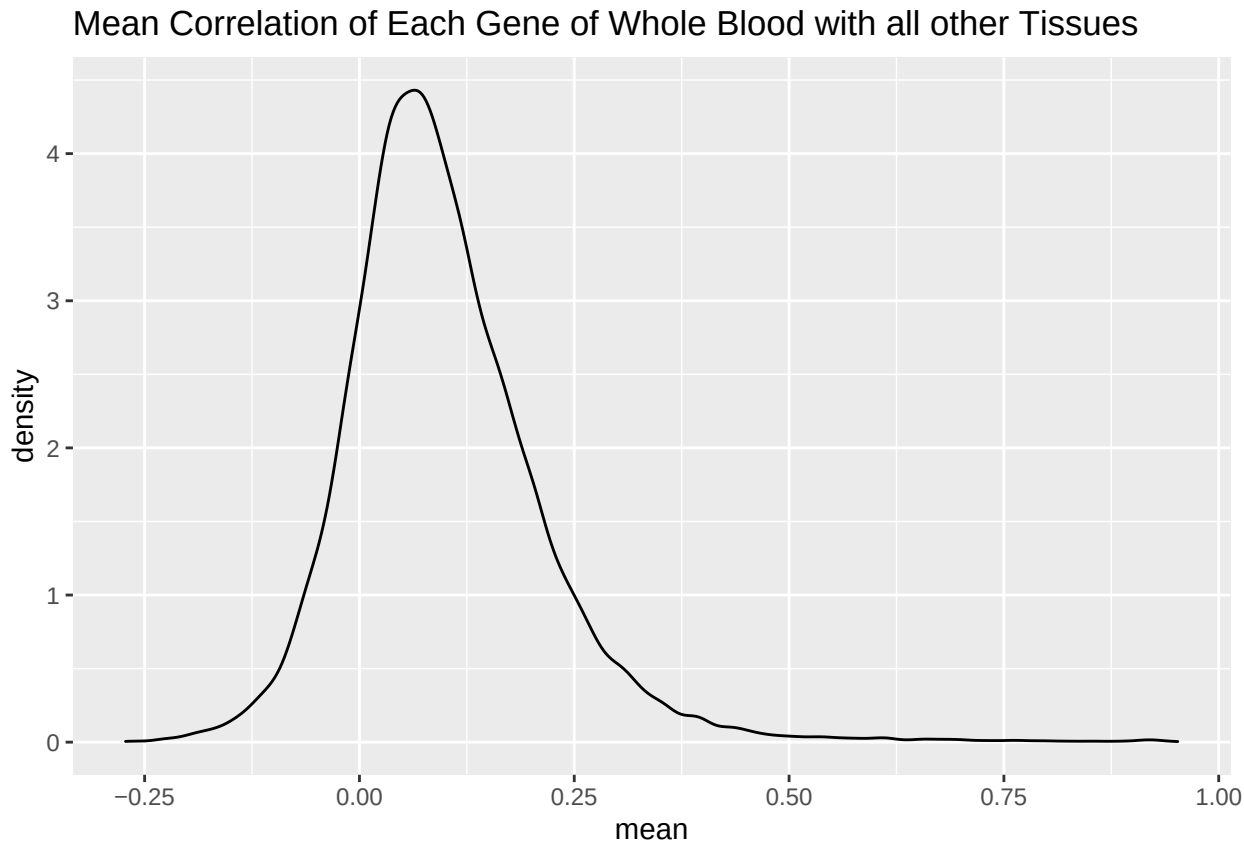

Figure 16: Mean Correlation of Each Gene of Whole Blood with all other Tissues

Figure 16 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 **Figure 17** considers absolute correlation of each gene in Whole Blood with the same gene found in all other tissues, the x axis has a lower bound at zero. The majority of genes in other tissues have a correlation of about 0.10, but a very small number of them have a strong correlation.

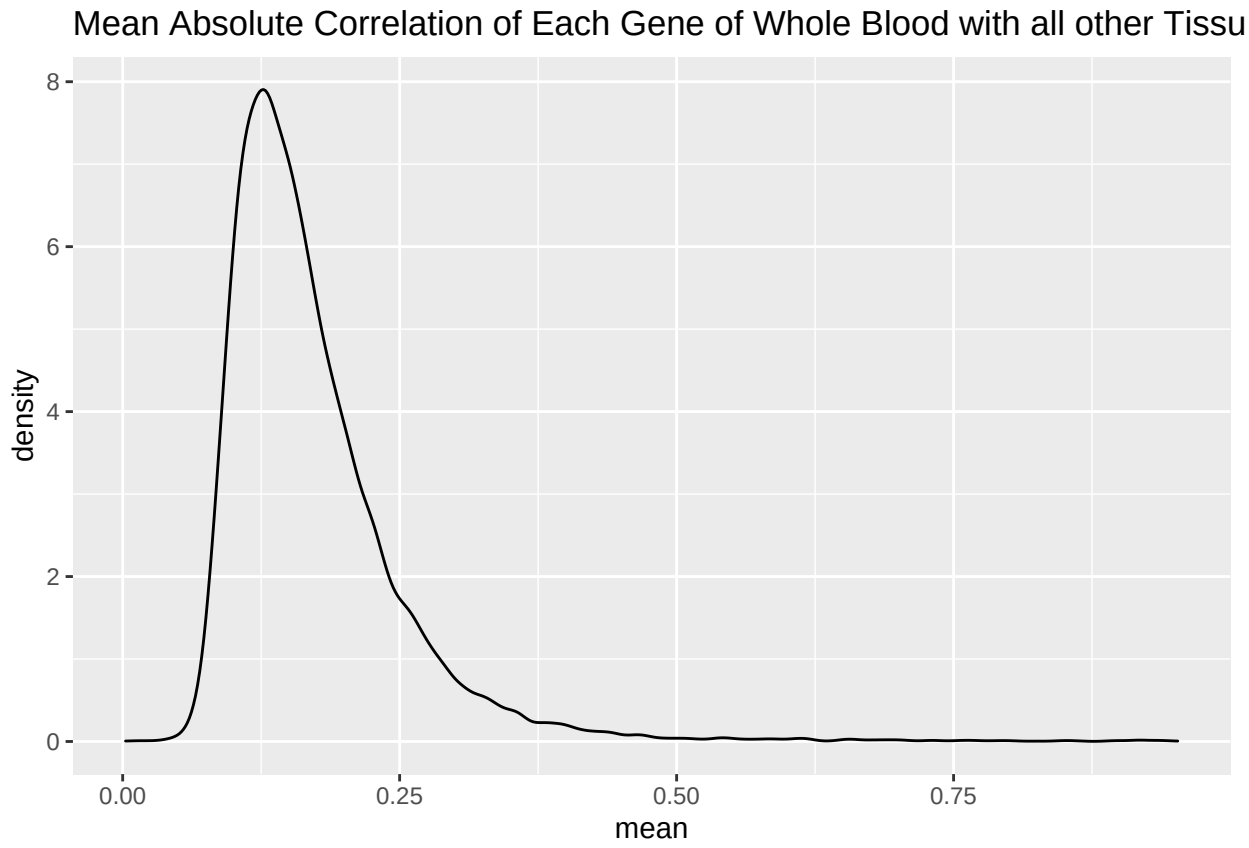

Figure 17: Mean Absolute Correlation of Each Gene of Whole Blood with all other Tissues

3.2.3 Ranking tissues based on previously observed measures

Finally, to acquire selection of the best tissues across all measures, we rank tissues based on three different measures: proportion of overlapping donors, proportion of shared genes and correlation between expression levels. The rank per tissue per measure is averaged out to give us the final rank it holds.

The following list and **Figure 18** 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.

3.2.4 Conclusion

Here we explored how shared genes across all tissues are correlated to Whole Blood. Each correlation value measured in this section helps us understand the kind of relationship every tissue shares with Whole Blood, based on the genes they share.

Not necessarily having the lowest rank, some tissues are chosen as the starting point for further analysis. The tissues are Lung, Skin - Not sun exposed, and Nerve - Tibial. These tissues are chosen because they have high overlapping number of donors with whole blood, taking into account its low correlations with whole blood. Contradicting the initial reason of using the lowest ranked tissues as the starting point, it is decided to use tissues with low correlation with whole blood so that the starting point could also be considered as the worst case scenario.

4 Proposal

Using Lung as the starting point, we want to use the gene expression from Whole Blood as the input for the models and the gene expression from Lung as the output. Below we will discuss the sample size and the methods.

4.1 Sample Size

This section aims at selecting appropriate sample size for training, validation and test sets defined for the chosen tissues i.e. Lung, Skin - Not sun exposed, and Nerve - Tibial. The type of sets involved are:

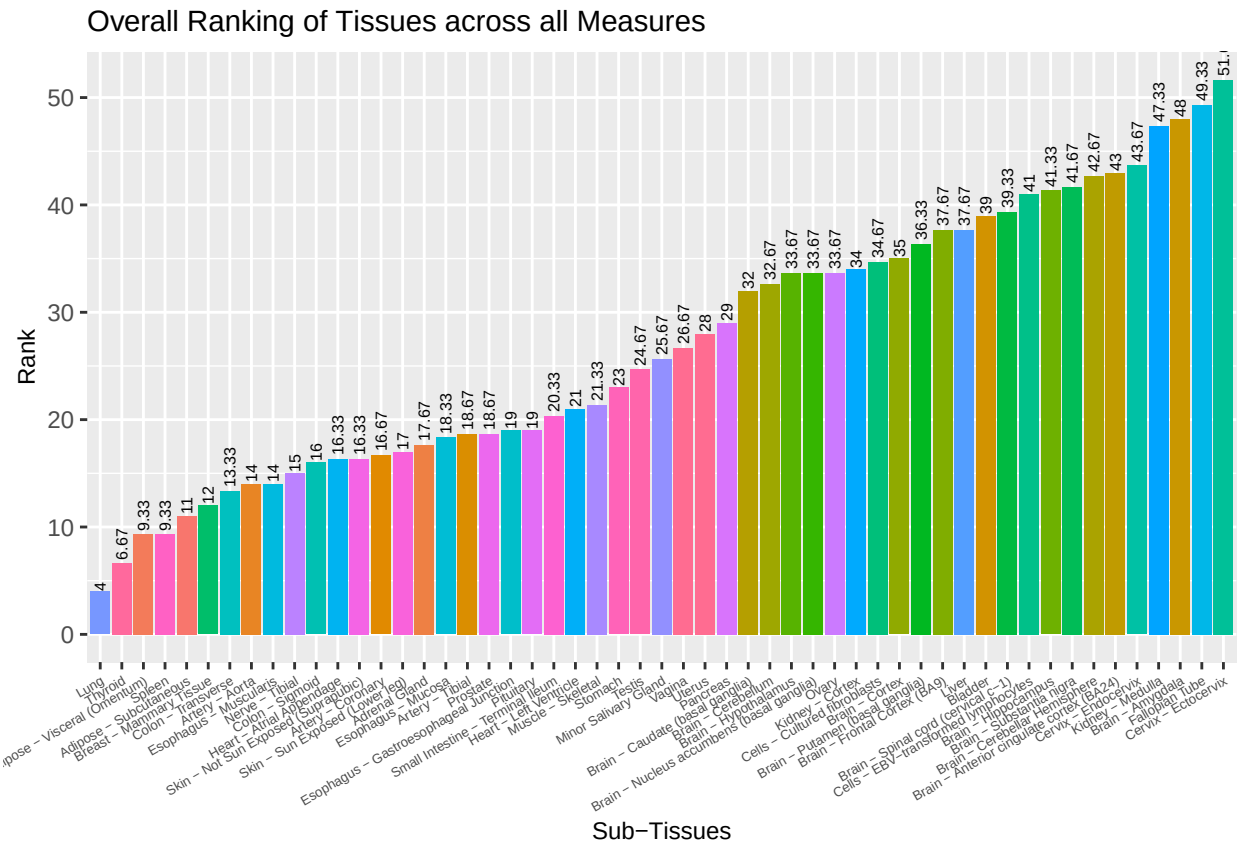

Figure 18: Overall Ranking of Tissues across all Measures

1. Training set: the samples belonging to this set will be used to create the model. It is typically the set with the largest sample size.
2. Validation set: this set will be used to tune the parameters of the model and to further optimise it. Furthermore, this set will also be used to choose the best model between the 2 models (Random Forest and Neural Network) which are going to be further discussed in the next subsection.
3. Test set: the model is finally run on this set, which it has never seen. It is used to assess the performance of the model and the accuracy of results.

We could not find a strong agreement after referring various sources and published papers on selecting appropriate proportion to define Training, Validation and Test sets. Hence, our approach to choosing the correct proportions for the sets was to ensure that each one of them has an ample amount of donors. Choosing an arbitrary number of 30 samples, our main aim is to ensure that the test set (the smallest across the three) must have a sample size of at least 30 donors overlapping with Whole Blood.

By the method of trial and error, random sets of proportions were tried and tested on the chosen tissues. The proportions that work the best for modelling the three chosen tissues are:

1. Training set: 0.7 or 70%
2. Validation set: 0.2 or 20%
3. Test set: 0.1 or 10%

In **Figure 19**, we can see that for each tissue, the test set comfortably exceeds the count of 30 donors. Hence, we can conclude that 0.7-0.2-0.1 is a suitable proportion for the sets.

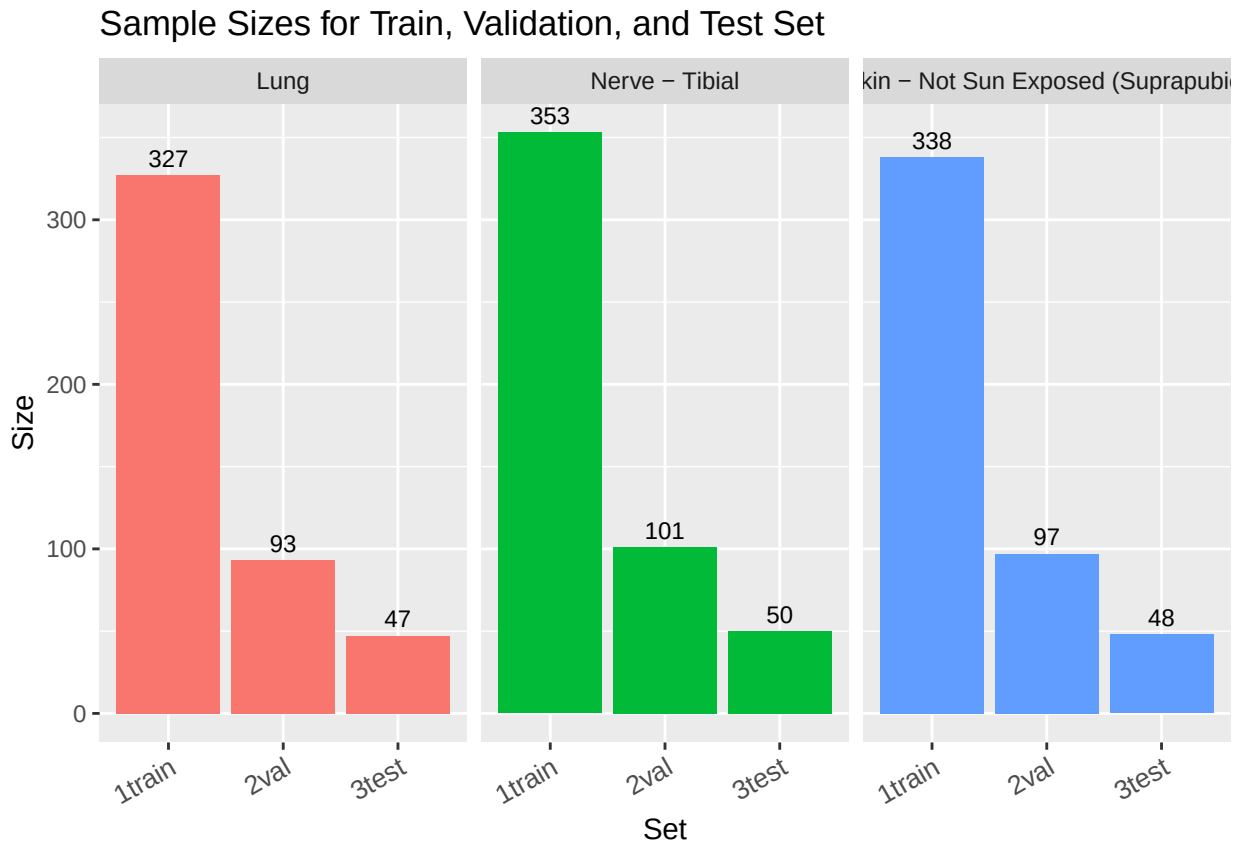

Figure 19: Count of samples per set for chosen tissues

4.2 Methods

4.2.1 Model Deployment

Neural Network

Using the keras_model_sequential() function from 'keras' library, we will create and compile the model. The input will be a matrix of the gene expression of whole blood and the output will be a matrix of the gene expression prediction of lung. The visualisation of the expected neural network model is shown in **Figure 20**. Note that the number of hidden layers is yet to be decided, and it is for visualisation purpose only. Same thing applies for the number of input and output nodes. The number of input and output nodes in the real model will depend on the number of shared genes between whole blood and the other tissue.

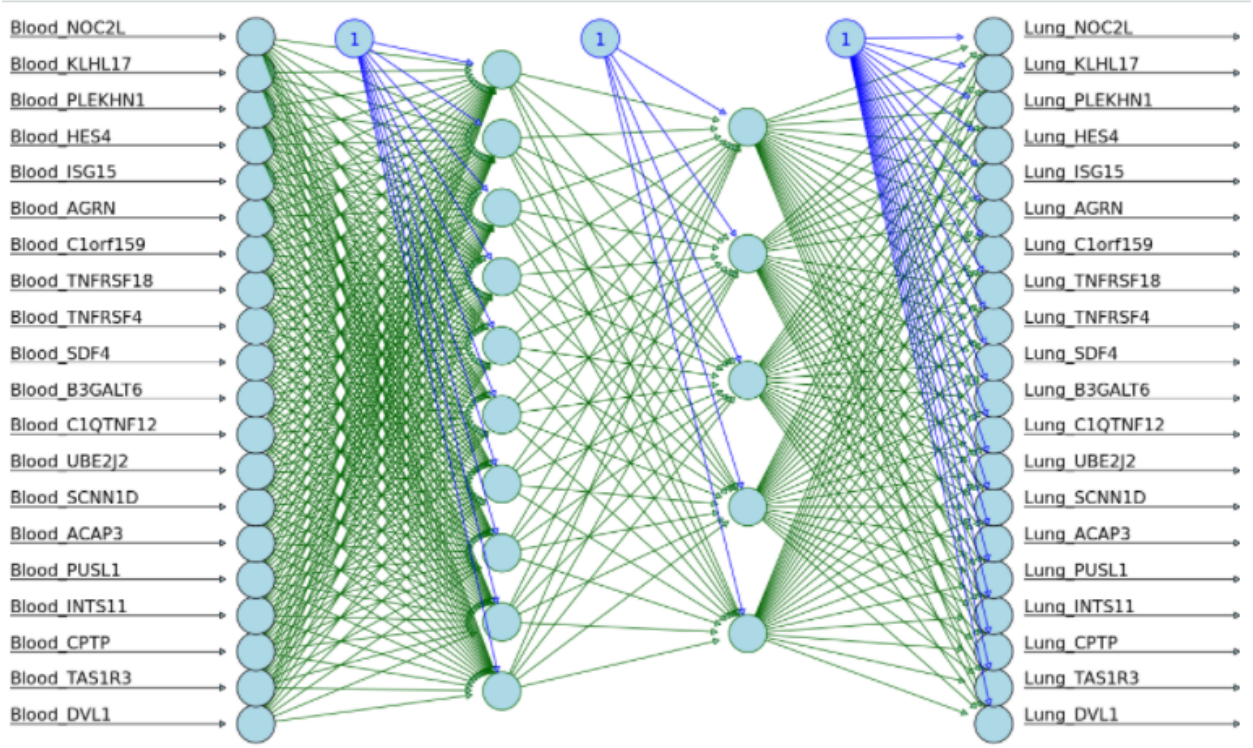

Figure 20: An snapshot of GTex data as a neural network

Random Forest

Here, we will implement eXtreme gradient boosting or XGBoost (XGB). It is easily available and installable as a software package on R, in 'caret' and 'xgboost' packages. The input will be a matrix of the gene expression of Whole Blood and the expected output is a random forest of the best predictors for gene expression in the other tissue. The model will be built with the XGB package, using the training set and tuned for better results over the validation set.

4.2.2 Model Fine Tuning

For both models, we will tune the models' parameters using the validation set. If possible (given the time and computational power limitation), we are planning to do an n-fold cross validation to get a more accurate and non-biased results.

4.2.3 Final Model Testing

From the previous section's results, we will take the best model from each proposed model (i.e. one best neural network model and one best random forest model), and test it on the test set. The best model among the two models will be chosen to move forward to further process.

5 Timeline

5.1 Achieved

1. **Week 1 (26 Apr 2021 - 30 Apr 2021)**: Id matching & batch cleaning

- Understanding what links information from multiple tissues to the tissue donors.
- Exploring gene expression profiles in all tissues.
- Understanding whether clustering based on information available is occurring and how to solve it.

2. **Week 2-3 (3 May 2021 - 14 May 2021)**: Exploration of tissues similarity and gene expression correlation in GTEx V8

- Most tissue similarity and correlation has been performed using previous versions of GTEx so this step helped understand the extent and structure of the dataset.
- Similarity scores, correlation, hierarchical clustering, plotting are performed here.

3. **Week 4 (17 May 2021 - 21 May 2021)**: Selection of tissues and sample sizes for training, validation and test sets

- In GTEx, sample sizes vary according to tissue type and blood is the largest. sample size. The appropriate sample size for training, validation and test sets is defined for each tissue.

4. **Week 5 (24 May 2021 - 28 May 2021)**: Proposal of a number of methods and discussion of such methods

- Selecting 2 methods and justifying the selected methods in the context of the project.

5.2 Plan

1. **Week 6-9 (26 Jul 2021 - 20 Aug 2021)**: Deployment of all the proposed methods in different tissues

- Each member will then deploy their method on a few selected tissues (approx. 10 tissues with the largest sample sizes). We will then tune the models to each of the tissues on the validation sets and produce comparable results.

2. **Week 10 (23 Aug 2021 - 27 Aug 2021)**: Benchmarking of the methods and selection of a final methodology

- Compare the methods against each other and describe which one should be used for the final deployment.

3. **Week 11-13 (30 Aug 2021 - 17 Sep 2021)**: Deployment of the final methodology and fine tuning on validation

- The chosen method will be deployed on all tissues and finely tuned to each tissue.

4. **Week 14 (27 Sep 2021 - 1 Oct 2021)**: Assessment of the final methodology and the test set

- The final tuned models will be deployed to the test sets and the final conclusion will be drawn.

6. **Week 15-16 (4 Oct 2021 - 15 Oct 2021)**: Exploration and description of the results

- Collection, summary, reportage and presentation of the final results will be performed. The students will present at both CSL and WEHI seminar as a team.

7. **Week 17-18 (18 Oct 2021 - 29 Oct 2021)**: Development of a user friendly algorithm

- The student will clean the final algorithm script and make it as user friendly as possible for future deployment.

Bibliography

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. IEEE. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Al Bkhetan, Z. (2020). Optimisation of phasing: towards improved haplotype-based genetic investigations (Doctoral dissertation, pp. 10-13). http://hdl.handle.net/11343/258861

Bakar, N. M. A., & Tahir, I. M. (2009). Applying multiple linear regression and neural network to predict bank performance. International Business Research, 2(4), (pp.176-183). https://doi.org/10.5539/ibr.v2n4p176

Basu, M., Wang, K., Ruppin, E., & Hannenhalli, S. (2021). Predicting tissue-specific gene expression from whole blood transcriptome. Science Advances, 7(14), eabd6991. https://doi.org/10.1101/2020.05.10.086942

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4).

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In Ensemble machine learning (pp. 157-175). Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-9326-7_5

Hsueh, H. M., Zhou, D. W., & Tsai, C. A. (2013). Random forests-based differential analysis of gene sets for gene expression data. Gene, 518(1), 179-186. https://doi.org/10.1016/j.gene.2012.11.034

Qi, Y., Oja, M., Weston, J., & Noble, W. S. (2012). A unified multitask architecture for predicting local protein properties. PloS one, 7(3). https://doi.org/10.1371/journal.pone.0032235

Qi, Y. (2012). Random forest for bioinformatics. In Ensemble machine learning (pp. 307-323). Springer, Boston, MA.https://doi.org/10.1007/978-1-4419-9326-7_11

Singh, R., Lanchantin, J., Robins, G., & Qi, Y. (2016). DeepChrome: deep-learning for predicting gene expression from histone modifications. In Bioinformatics (Vol. 32, pp. i639–i648). https://doi.org/10.1093/bioinformatics/btw427