Block 1: Introduction and data exploration

--- Huizhi Lin (881125518130)

--- Jan Orsel (950608630010)

November 2, 2017

Task 1

First we loaded in the data from file "get_normal_vs_tumor2_RAW_Skin.out" into dataframe "skin.df". This is shown in the following code block. By opening the file manually we found that as a separator a space should be used.

```
skin.df <- read.table("get_normal_vs_tumor2_RAW_Skin.out",sep=' ',
header=TRUE)
```

Secondly we created another dataframe "re.skin.df" with samples in column and genes in rows. For this we used an adapted version of the code given to us in the assignment handout.

```
re.skin.df <- data.frame(t(skin.df[,-2562]))
# Add sample type as column name
colnames(re.skin.df)<-paste0(skin.df$tissue,1:72)
```

We chose tumor samples tumor1,tumor2 and tumor3, and normal samples normal70, normal71 and normal72. Then we calculated the range of expressions for those samples using the range function.

```
range(re.skin.df$tumor1)

## [1]    10.40 18651.45

range(re.skin.df$tumor2)

## [1]    10.32 18504.65

range(re.skin.df$tumor3)

## [1]    10.39 20513.70

range(re.skin.df$normal70)

## [1]    10.81 19013.02

range(re.skin.df$normal71)

## [1]    10.47 11436.83

range(re.skin.df$normal72)
```

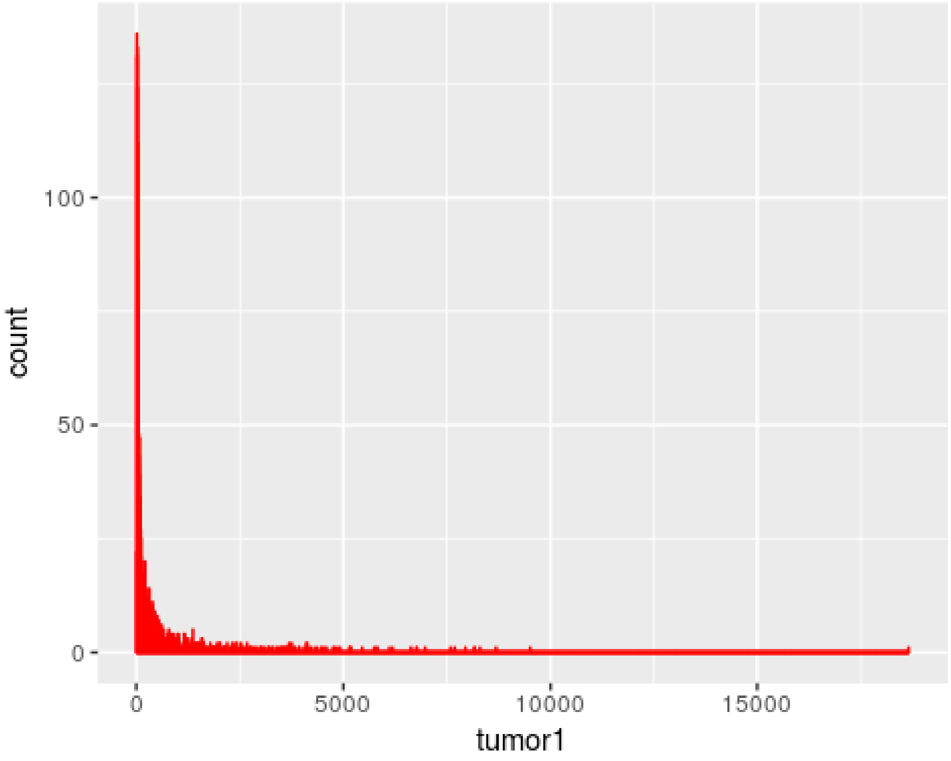
```
## [1]     9.96 14932.83
```

The range of the expression for all tumor samples combined and all the normal samples combined was calculated too.

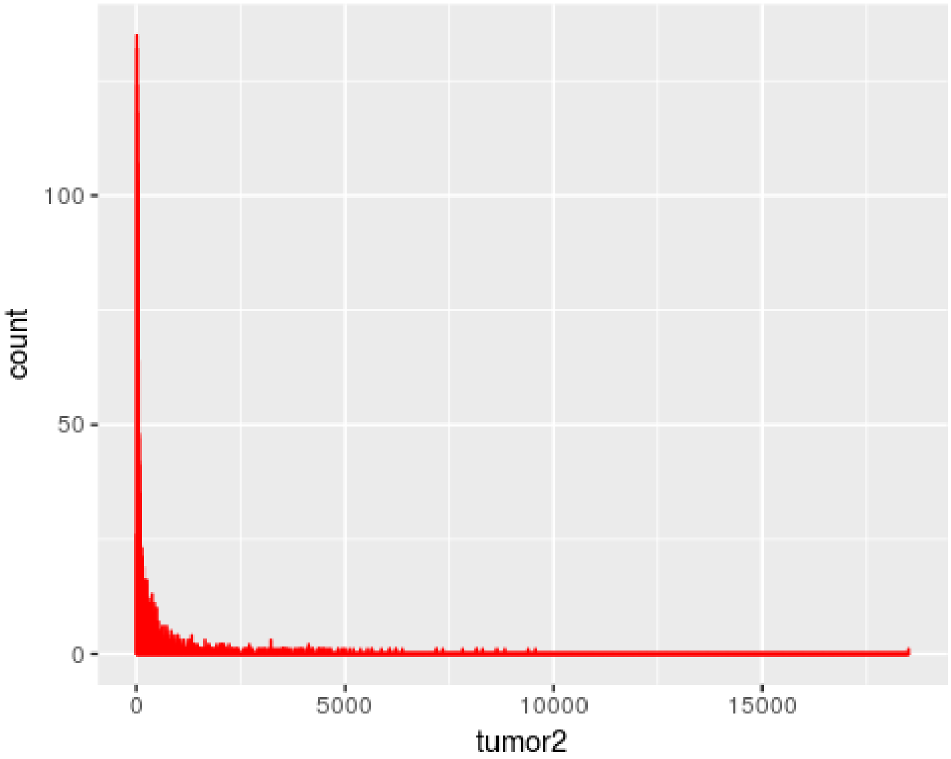
```
# over all tumor smples
range(re.skin.df[1:43])

## [1]     9.70 21891.76

# over all normal smples
range(re.skin.df[44:72])

## [1]     9.62 19013.02
```

For the six data sets we chose histograms were made using ggplot. In order to use this the ggplot library was addressed.

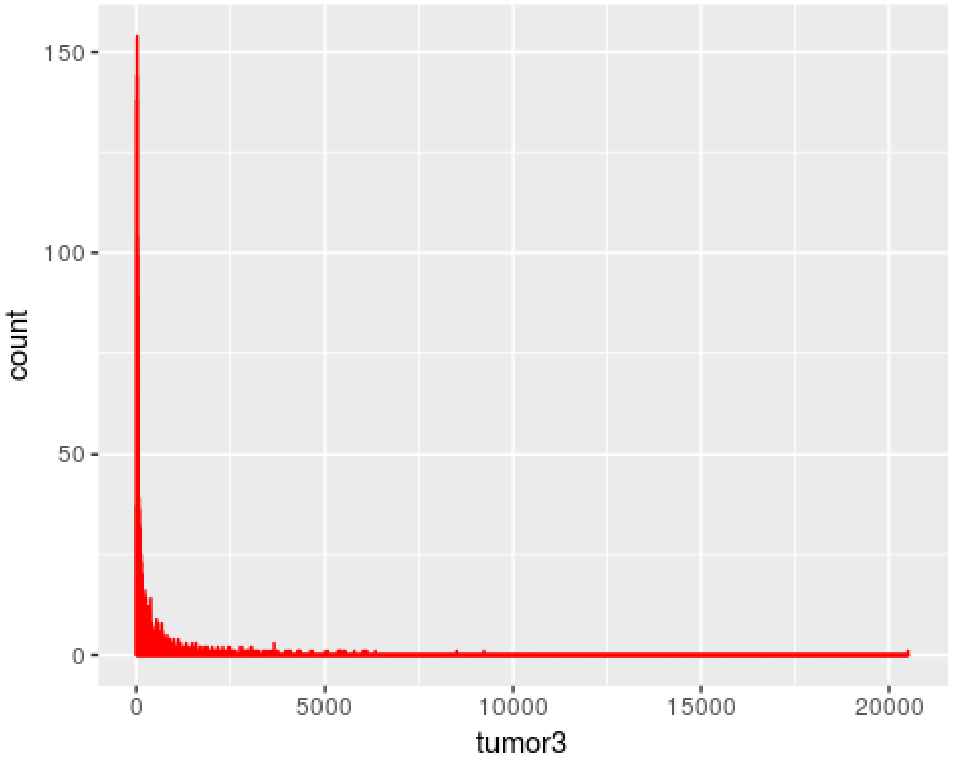
```
# load ggplt2
library(ggplot2)
# Create histogram for tumor1
pl.tumor1 <- ggplot(data=re.skin.df, aes(x=re.skin.df$tumor1))
pl.tumor1 <- pl.tumor1 + geom_histogram(binwidth=5,color="red",fill="red",
alpha=0.8) + xlab("tumor1")
print(pl.tumor1)
```



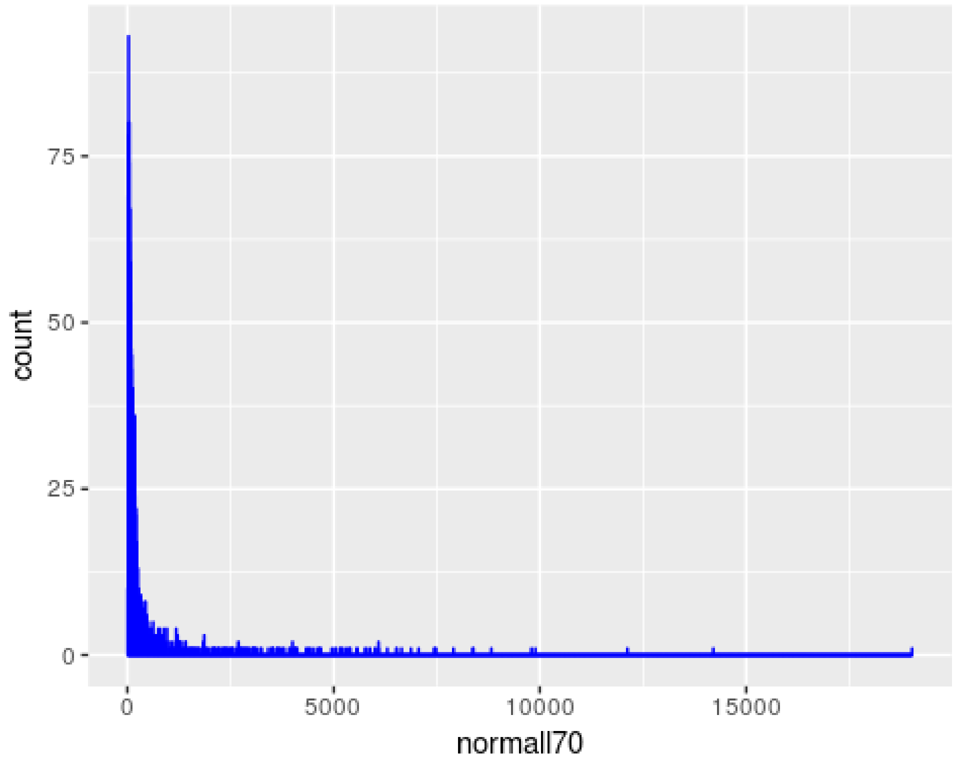
```
# Create histogram for tumor2
pl.tumor2 <- ggplot(data=re.skin.df, aes(x=re.skin.df$tumor2))
pl.tumor2 <- pl.tumor2 + geom_histogram(binwidth=5,color="red",fill="red",
alpha=0.8) + xlab("tumor2")
print(pl.tumor2)
```



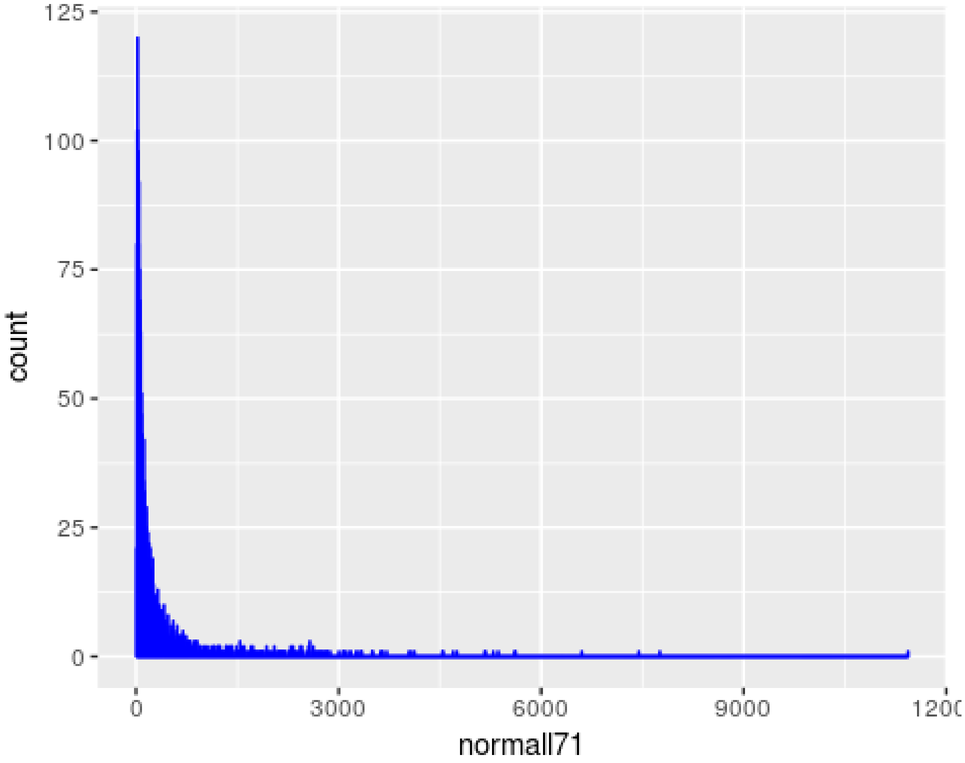
```
# Create histogram for tumor3
pl.tumor3 <- ggplot(data=re.skin.df, aes(x=re.skin.df$tumor3))
pl.tumor3 <- pl.tumor3 + geom_histogram(binwidth=5,color="red",fill="red",
alpha=0.8) + xlab("tumor3")
print(pl.tumor3)
```



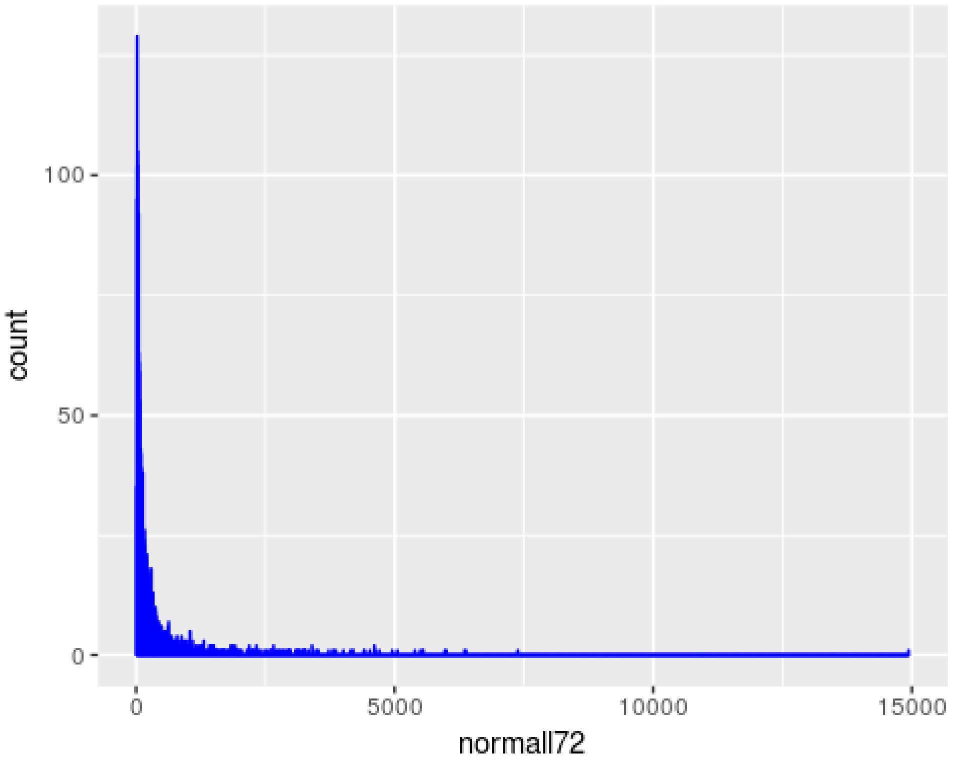
```r
# Create histogram for normall70
pl.normall70 <- ggplot(data=re.skin.df, aes(x=re.skin.df$normal70))
pl.normall70 <- pl.normall70 +
geom_histogram(binwidth=5,color="blue",fill="blue", alpha=0.8) +
xlab("normall70")
print(pl.normall70)
```



```
# Create histogram for normall71
pl.normall71 <- ggplot(data=re.skin.df, aes(x=re.skin.df$normal71))
pl.normall71 <- pl.normall71 +
geom_histogram(binwidth=5,color="blue",fill="blue", alpha=0.8) +
xlab("normall71")
print(pl.normall71)
```

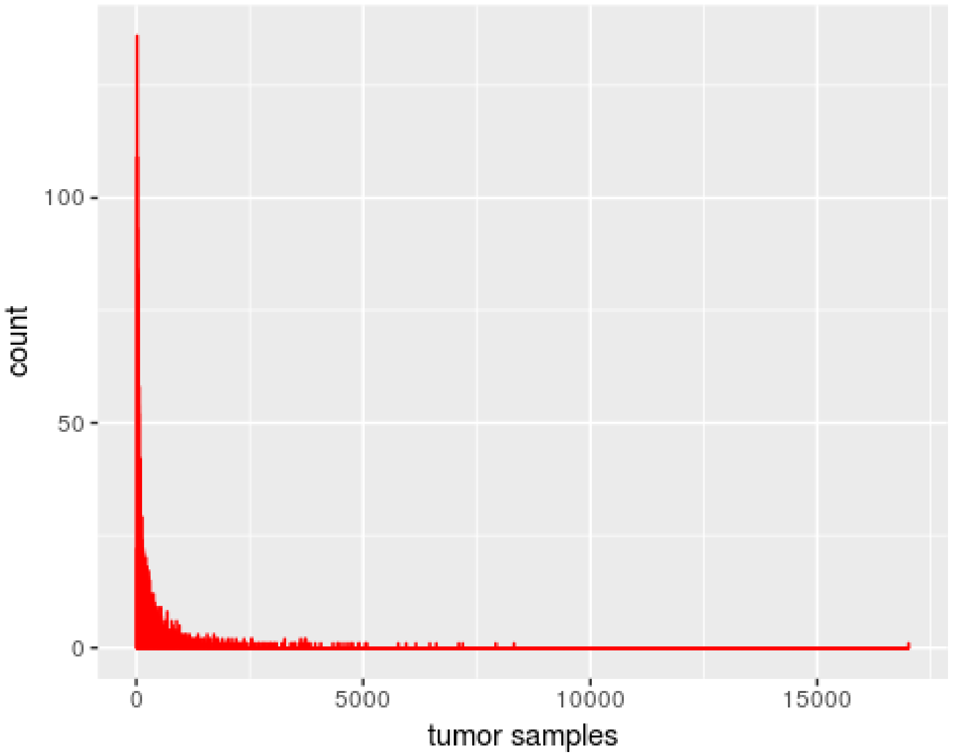


```
# Create histogram for normall72
pl.normall72 <- ggplot(data=re.skin.df, aes(x=re.skin.df$normal72))
pl.normall72 <- pl.normall72 +
geom_histogram(binwidth=5,color="blue",fill="blue", alpha=0.8) +
xlab("normall72")
print(pl.normall72)
```

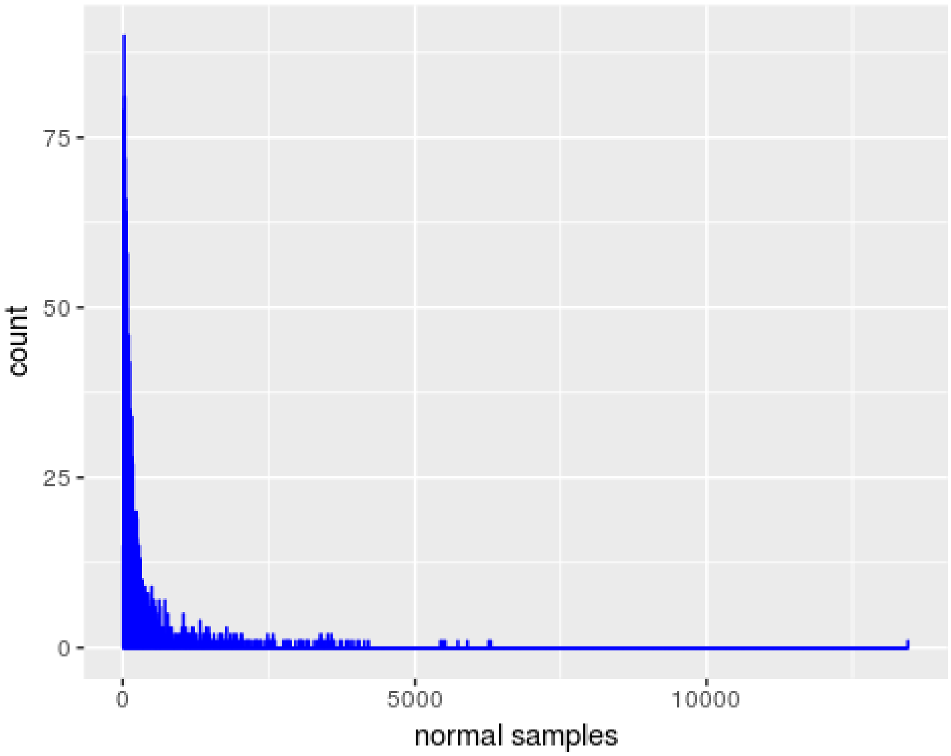


The data of all tumor samples was used to create a red colored histogram. All the normal samples were used to create a blue histogram.

```r
# Create histogram for tumor samples
pl.tumor <- ggplot(data=re.skin.df, aes(x=rowMeans(re.skin.df[1:43])))
pl.tumor <- pl.tumor + geom_histogram(binwidth=5,color="red",fill="red",
alpha=0.8) + xlab("tumor samples")
print(pl.tumor)
```



```
# Create histogram for normal samples
pl.normal <- ggplot(data=re.skin.df, aes(x=rowMeans(re.skin.df[44:72])))
pl.normal <- pl.normal + geom_histogram(binwidth=5,color="blue",fill="blue",
alpha=0.8) + xlab("normal samples")
print(pl.normal)
```

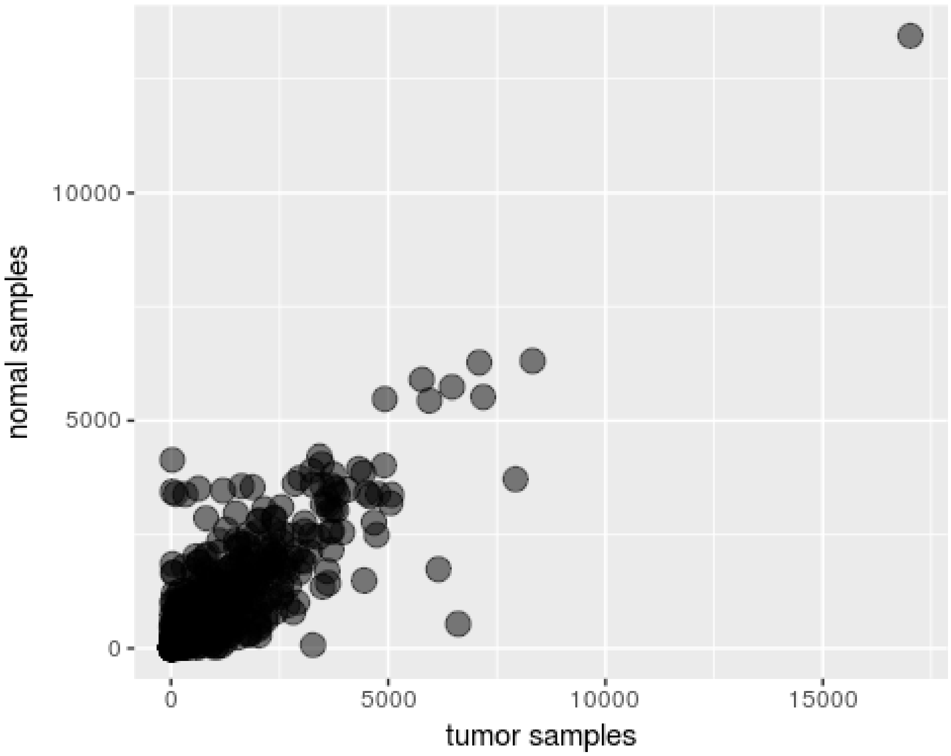


When looking at the gene expression for the three chosen tumor samples it becomes clear that most genes have a low rate of expression with some outliers. The same can be stated on the three samples chosen for the normal samples. In order to compare all the tumor samples to all the normal samples two hisograms were created, this is shown in the block above. When comparing the two hisograms that this code creates no clear differences are shown. It is clear that there are a few genes that are heavily upregulated but a scatterplot is needed to better visualise this.

Task 2

We made a scatterplot of the average expression level over all tumor samples vs. the average of all normal samples. Ggplot was used to create the graph using the geom_point function.

```
pl.scatter <- ggplot(data=re.skin.df,
aes(x=rowMeans(re.skin.df[1:43]),y=rowMeans(re.skin.df[44:72])))
pl.scatter <- pl.scatter + geom_point(alpha=0.5,size=4)
pl.scatter <- pl.scatter + xlab("tumor samples") + ylab("nomal samples")
print(pl.scatter)
```

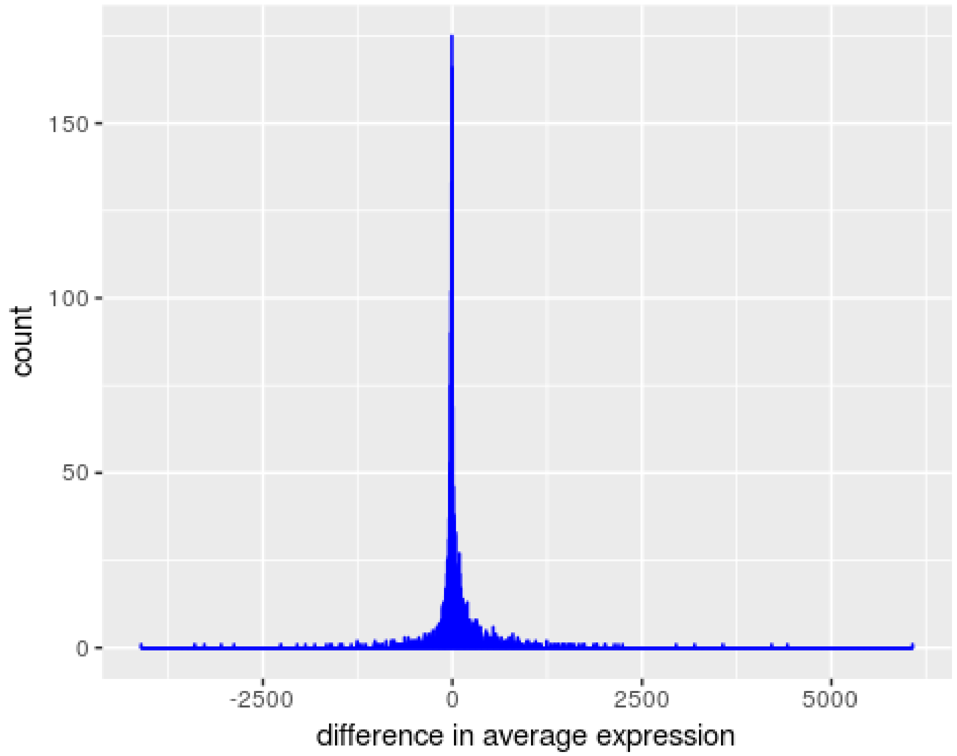


here we see the avarage expression level of tumor samples compared to all normal samples. If there would have been no change in expression all samples would line up in a diagonal line over the middle of the plot. This plot shows a concentraion of dots on this line indicating that for a lot of genes the change in expression is minimal. There are however some dots that clearly show that some genes are over or underexpressed.

Task 3

A histogram of the difference between the average expression level over all tumor samples and that over all normal samples was made using the ggplot geom_histogram function.

```
pl.differ <- ggplot(data=re.skin.df, aes(x=rowMeans(re.skin.df[1:43]) -
rowMeans(re.skin.df[44:72])))
pl.differ <- pl.differ + geom_histogram(binwidth=5,color="blue",fill="blue",
alpha=0.8)
pl.differ <- pl.differ + xlab("difference in average expression")
print(pl.differ)
```



This histogram shows that the numbers of over- and underexpressed genes. The graph skyrockets around zero. This indicates that a lot of genes have little or no average difference in expression. It must be noted though that the few outliers indicate that for some genes the difference in expression is big. The rowMeans function is used to create an average of expression.

```
# Calculate the differences
differ <- rowMeans(re.skin.df[1:43]) - rowMeans(re.skin.df[44:72])
# Count the numbers of over- and underexpressed genes with a for loop
overexpressed.gene <- 0
underexpressed.gene <- 0
for (dif in differ){
  if (dif>0){overexpressed.gene <- overexpressed.gene+1}
  if (dif<0){underexpressed.gene <- underexpressed.gene+1}
}
# print out the result
print(overexpressed.gene)
```

```
## [1] 1218
```

```
print(underexpressed.gene)
```

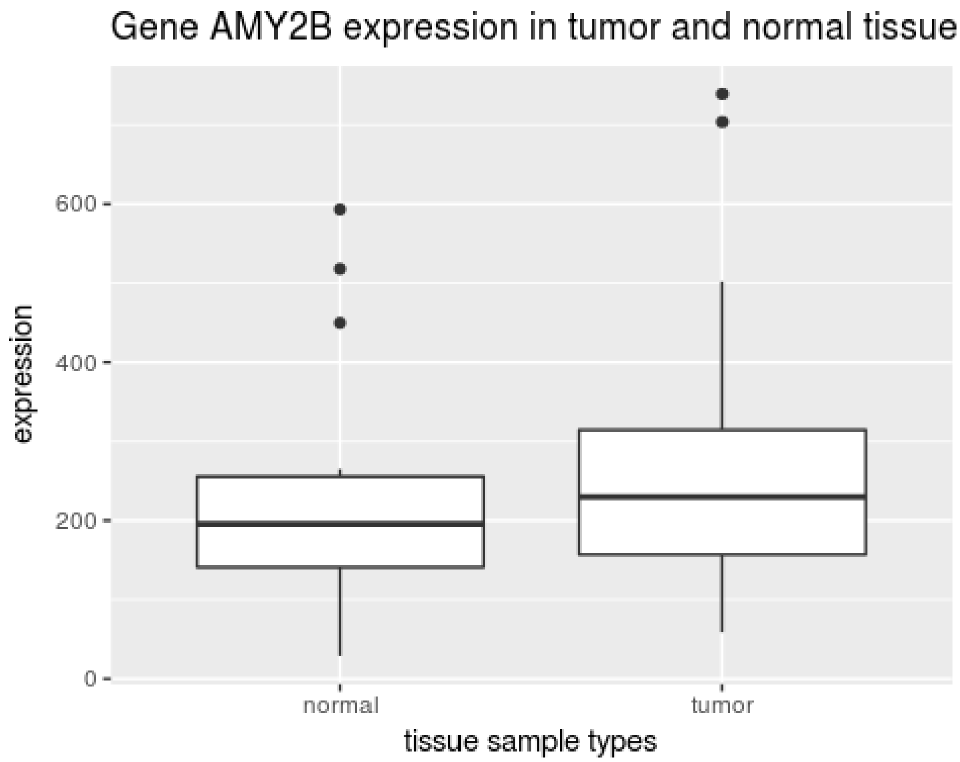
```
## [1] 1343
```

The result shows that there are 1218 overexpressed genes and 1343 underexpressed genes. This devides the data in two camps but it must be noted that for most of these data points the difference in expression is small. For one gene there is no difference in expression at all.

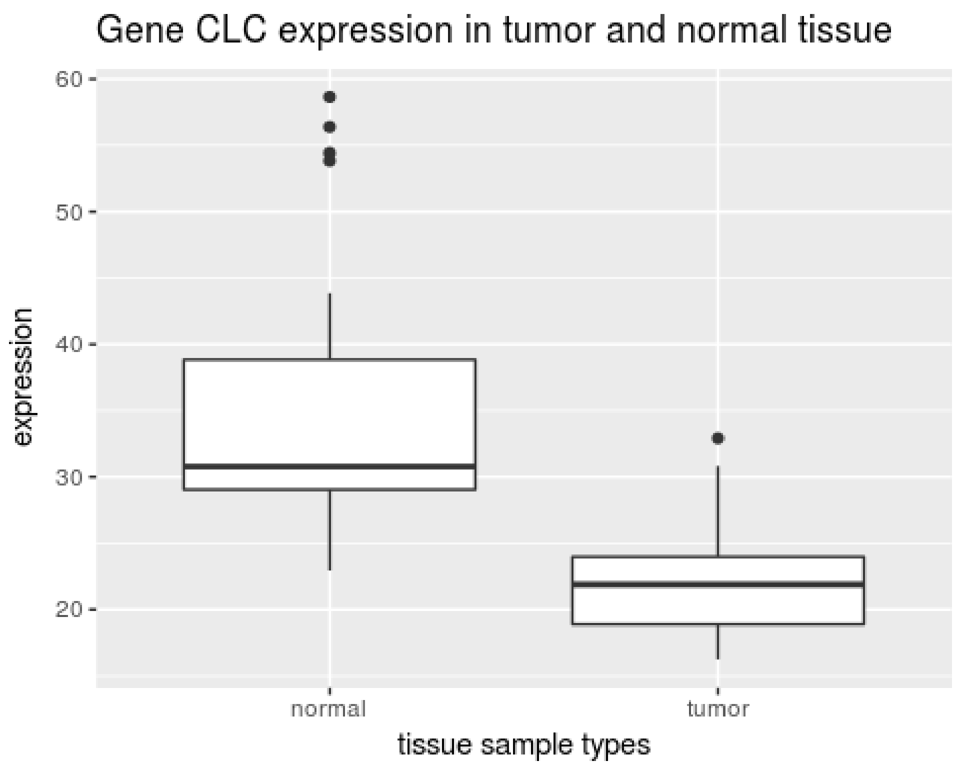
Task 4

For the genes AMY2B, CLC and NAT1 boxplots are created using the block of code below. The geom_boxplot function is used to create these graphs. In these graphs the gene expression in normal and tumor tissue is compared to eachother.

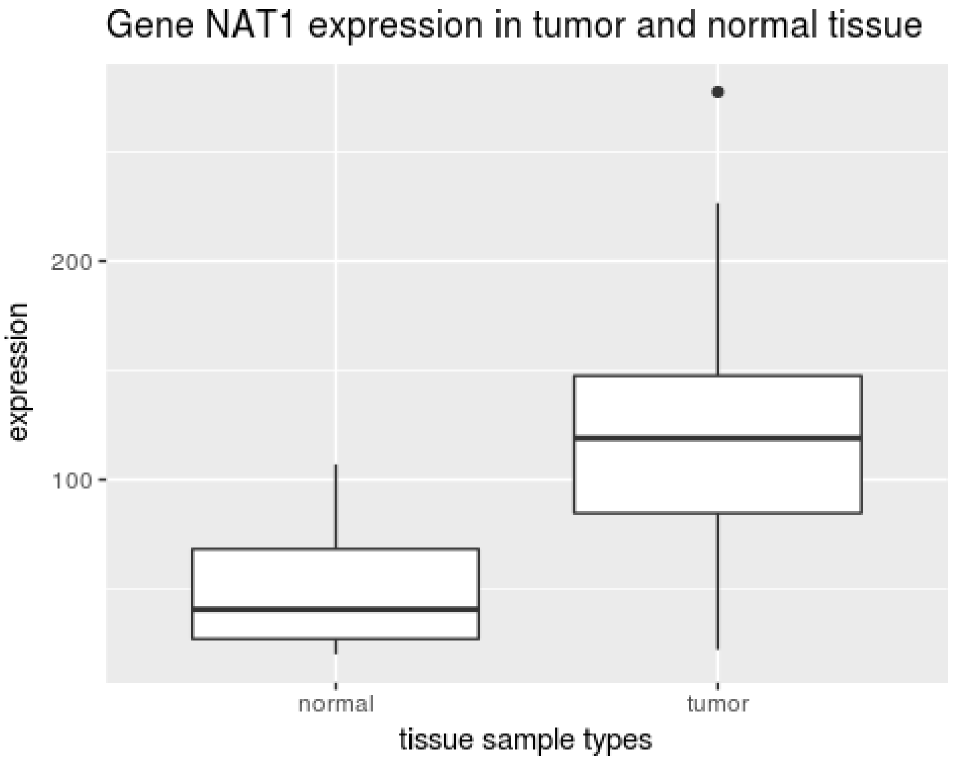
```
# Creat a boxplot for gene AMY2B
pl.box1 <- ggplot(data=skin.df,
aes(x=factor(skin.df$tissue),y=skin.df$AMY2B))
pl.box1 <- pl.box1 + geom_boxplot() + xlab("tissue sample types") +
ylab("expression")
pl.box1 <- pl.box1 + ggtitle("Gene AMY2B expression in tumor and normal
tissue")
print(pl.box1)
```



```
# Creat a boxplot for gene CLC
pl.box2 <- ggplot(data=skin.df, aes(x=factor(skin.df$tissue),y=skin.df$CLC))
pl.box2 <- pl.box2 + geom_boxplot() + xlab("tissue sample types") +
ylab("expression")
pl.box2 <- pl.box2 + ggtitle("Gene CLC expression in tumor and normal
tissue")
print(pl.box2)
```



```
# Creat a boxplot for gene NAT1
pl.box3 <- ggplot(data=skin.df, aes(x=factor(skin.df$tissue),y=skin.df$NAT1))
pl.box3 <- pl.box3 + geom_boxplot() + xlab("tissue sample types") +
ylab("expression")
pl.box3 <- pl.box3 + ggtitle("Gene NAT1 expression in tumor and normal
tissue")
print(pl.box3)
```



Based on these boxplots, there is no big difference in gene AMY2B expression between normal and tumor samples. However the CLC-gene expression is up regulated in normal samples compared to the tumor samples, this could be an indication that CLC is a tumor suppressor gene. On the contrary, the NAT1 gene-expression is up regulated in tumor tissue compared to normal tissue.