Brain Lower Grade Glioma (LGG) dataset of TCGA for data analysis

Aman

10/04/2021

Assigment 3 - Aman kumar - 17025

You are provided with gene expression data from the TCGA (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcg (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcg)

a. project for Brain Lower Grade Glioma. The data is available in a text format known as “gct”. You will need to read into R and convert it to suitable format before using it for your assignment. To do this, you will need to understand the “gct” format. This is a important part of the assignment. Most large data projects use or invent custom formats to manage the unique meta data associated with their study. Hence, it is important to know how to become familiar with new dataformats and use it appropriately.

1. Provide all the steps used in R to convert the gct file into the format that you decide to use.
2. Make use of PCA, correlation analysis, linear regression and other statistical tests to explore the data. Identify the genes or other factors from the metadata that can help explain the heterogeneity between samples. Based on the exploratory data and your knowledge of biology, propose hypothesis that can be tested using this data.
3. Provide details of how these hypothesis are tested and what you find.

```
getwd()
```

```
## [1] "D:/IISER BHOPAL/SEM 8/DSE 401/Assignment 4"
```

```
setwd("D:/IISER BHOPAL/SEM 8/DSE 401/Assignment 4")
getwd()
```

```
## [1] "D:/IISER BHOPAL/SEM 8/DSE 401/Assignment 4"
```

```
#install.packages("limma")
```

```
library(ggplot2)
library(tidyverse)
library(plotly)
library(gplots)
require(ggiraph)
require(ggiraphExtra)
require(plyr)
library(PCAtools)
library("factoextra")
```

Part 1 : gct file to Data-frame

```
my_data <- read.delim("Brain Lower Grade Glioma.gct",skip = 2,check.names = F)
```

We can observe that we get gene count data for various genes in every patient after row number 97. So we will divide our dataframe into two parts

- one of which gives information about the patient we will call it **metadata**
- second one gives gene count for the particular gene we will call it **gene_count**

```
# geting subset as metadata:

metadata <- my_data[my_data$Source == "na",]
metadata <- t(metadata[,-2])
colnames(metadata) <- metadata[1,]
metadata <- as.data.frame(metadata[-1,])
```

```
# geting subset as gene_count:

gene_count <- my_data[my_data$Source != "na",]
colnames(gene_count) <- my_data[1,]
rownames(gene_count) <- make.names(gene_count[,1], unique=TRUE)
gene_count <- gene_count[,-c(1,2)]
```

```
# changing character entries to numeric

gene_count[,1:530] <- lapply(gene_count[,1:530], as.numeric)
```

Dividing metadata in categories

Dividing metadata in categories so we could better handle the data, we will be dividing data in following categories

On the basis of gender

- Dead male
- Dead Female
- Alive Male
- Alive Female

On the basis of Histology

- Astrocytoma Male
- Astrocytoma Female
- Oligodendroglioma Male
- Oligodendroglioma Female
- Oligoastrocytoma Male
- Oligoastrocytoma Female

Based on Age

- Young Male (Below 45 years of age)
- Young Female
- Old Male (Above 45 years of age)
- Old Female

```
Male <- apply(as_tibble(lapply(gene_count[,metadata[metadata$gender == "male",1]],as.numeric)),1
,mean)

Female <- apply(as_tibble(lapply(gene_count[,metadata[metadata$gender == "female",1]],as.numeri
c)),1,mean)

young_a <-  apply(as_tibble(lapply(gene_count[,metadata[metadata$vital_status== "alive" & metada
ta$age_at_initial_pathologic_diagnosis <= 45,1]],as.numeric)),1,mean)

young_d <- apply(as_tibble(lapply(gene_count[,metadata[metadata$vital_status== "dead" & metadata
$age_at_initial_pathologic_diagnosis <= 45,1]],as.numeric)),1,mean)

old_a <- apply(as_tibble(lapply(gene_count[,metadata[metadata$vital_status== "alive" & metadata
$age_at_initial_pathologic_diagnosis >= 45,1]],as.numeric)),1,mean)

old_d <- apply(as_tibble(lapply(gene_count[,metadata[metadata$vital_status== "dead" & metadata$a
ge_at_initial_pathologic_diagnosis >= 45,1]],as.numeric)),1,mean)


Dead_m <- apply(as_tibble(gene_count[,metadata[metadata$vital_status== "dead" & metadata$gender
 == "male",1]]),1,mean)

Dead_f <- apply(as_tibble(gene_count[,metadata[metadata$vital_status== "dead" & metadata$gender
 == "female",1]]),1,mean)

Alive_m <- apply(as_tibble(gene_count[,metadata[metadata$vital_status== "alive" & metadata$gende
r == "male",1]]),1,mean)

Alive_f <- apply(as_tibble(gene_count[,metadata[metadata$vital_status== "alive" & metadata$gende
r == "female",1]]),1,mean)

Astro_m <- apply(as_tibble(gene_count[,metadata[metadata$gender == "male" & metadata$histologica
l_type == "astrocytoma",1]]),1,mean)

Astro_f <- apply(as_tibble(gene_count[,metadata[metadata$gender == "female" & metadata$histologi
cal_type == "astrocytoma",1]]),1,mean)

Oligoa_m <- apply(as_tibble(gene_count[,metadata[metadata$gender == "male" & metadata$histologic
al_type == "oligoastrocytoma",1]]),1,mean)

Oligoa_f <- apply(as_tibble(gene_count[,metadata[metadata$gender == "female" & metadata$histolog
ical_type == "oligoastrocytoma",1]]),1,mean)

Oligod_m <- apply(as_tibble(gene_count[,metadata[metadata$gender == "male" & metadata$histologic
al_type == "oligodendroglioma",1]]),1,mean)

Oligod_f <- apply(as_tibble(gene_count[,metadata[metadata$gender == "female" & metadata$histolog
ical_type == "oligodendroglioma",1]]),1,mean)

young_m <-  apply(as_tibble(gene_count[,metadata[metadata$gender == "male" & metadata$age_at_ini
tial_pathologic_diagnosis <= 45,1]]),1,mean)

young_f <- apply(as_tibble(gene_count[,metadata[metadata$gender == "female" & metadata$age_at_in
```

```
itial_pathologic_diagnosis <= 45,1]]),1,mean)

old_m <- apply(as_tibble(gene_count[,metadata[metadata$gender == "male" & metadata$age_at_initia
l_pathologic_diagnosis >= 45,1]]),1,mean)

old_f <- apply(as_tibble(gene_count[,metadata[metadata$gender == "female" & metadata$age_at_init
ial_pathologic_diagnosis >= 45,1]]),1,mean)
```

Making Dataframe

```
new_data <- data.frame(Dead_Male = Dead_m, Dead_Female= Dead_f, Alive_Male= Alive_m, Alive_Femal
e= Alive_f, Astro_m= Astro_m, Astro_f=Astro_f, Oligoa_m=Oligoa_m, Oligoa_f= Oligoa_f, Oligod_m=
Oligod_m, Oligod_f= Oligod_f, young_m= young_m, young_f=young_f, old_m= old_m, old_f= old_f)
rownames(new_data) <- make.names(rownames(gene_count), unique=TRUE)
```

```
my_df <- new_data
```

```
# replacing na values with 0 for performing pca:

for (k in 1:14)
{
  m = mean(na.omit(new_data[,k]))
  new_data[,k][is.na(new_data[,k])] <- 0
  new_data[,k][new_data[,k] == 0] <- m
}
```

Part 2 : use of PCA, correlation analysis, linear regression and other statistical tests

PCA male vs female patients:

Finding Genes with most variance:

Since we have data for more than 18,000 genes, it will be impossible to cluster all those genes using PCA. So, we will instead choose genes with highest variance. High variance in data will be able to give us more information about the factors affecting the gene count to vary.

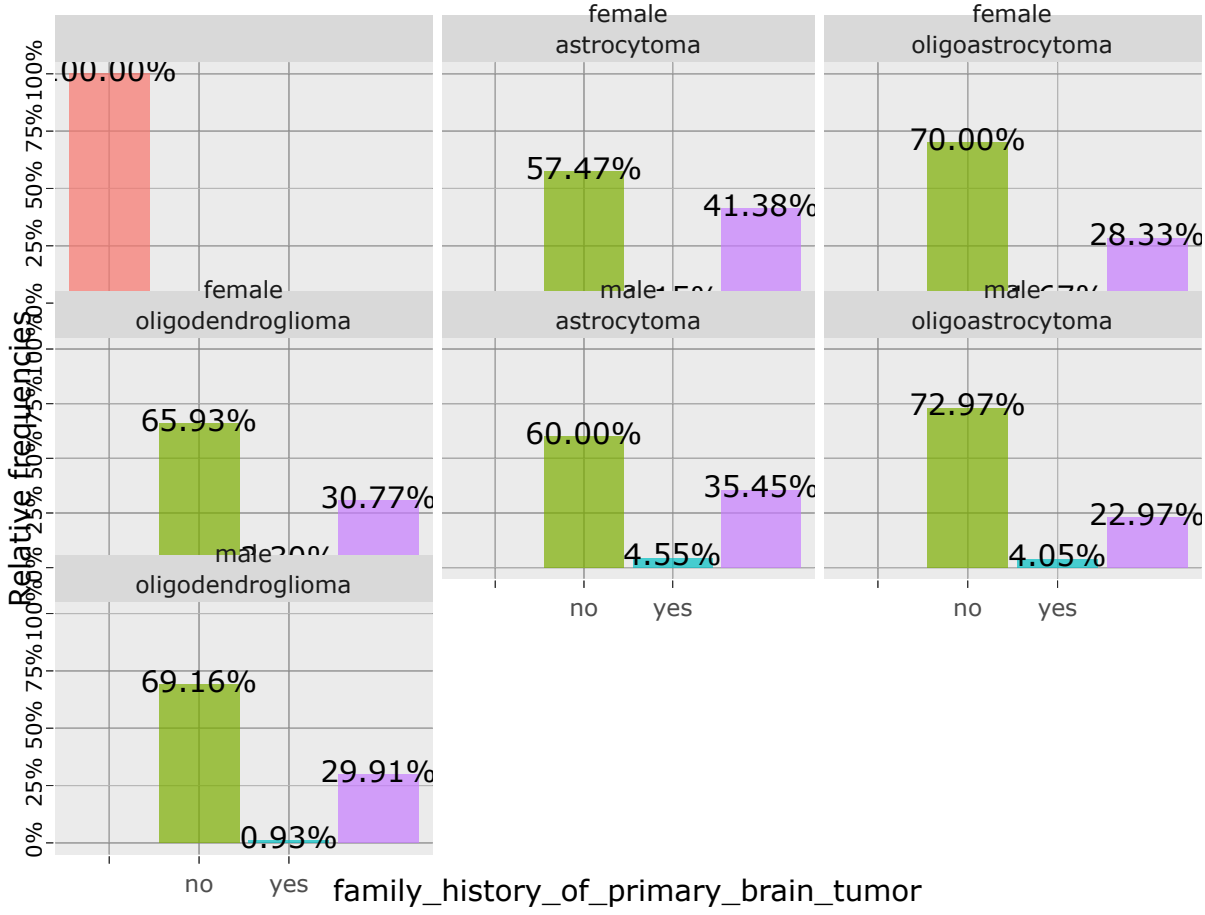
```
# subsetting the new_data based on genes with highest variance:

var_genes<-apply(my_df,1 ,var)
```

```
top_20_genes<- names(sort(var_genes,decreasing = T))[1:20]
```

```
new_data_pca<- new_data[top_20_genes,]
```

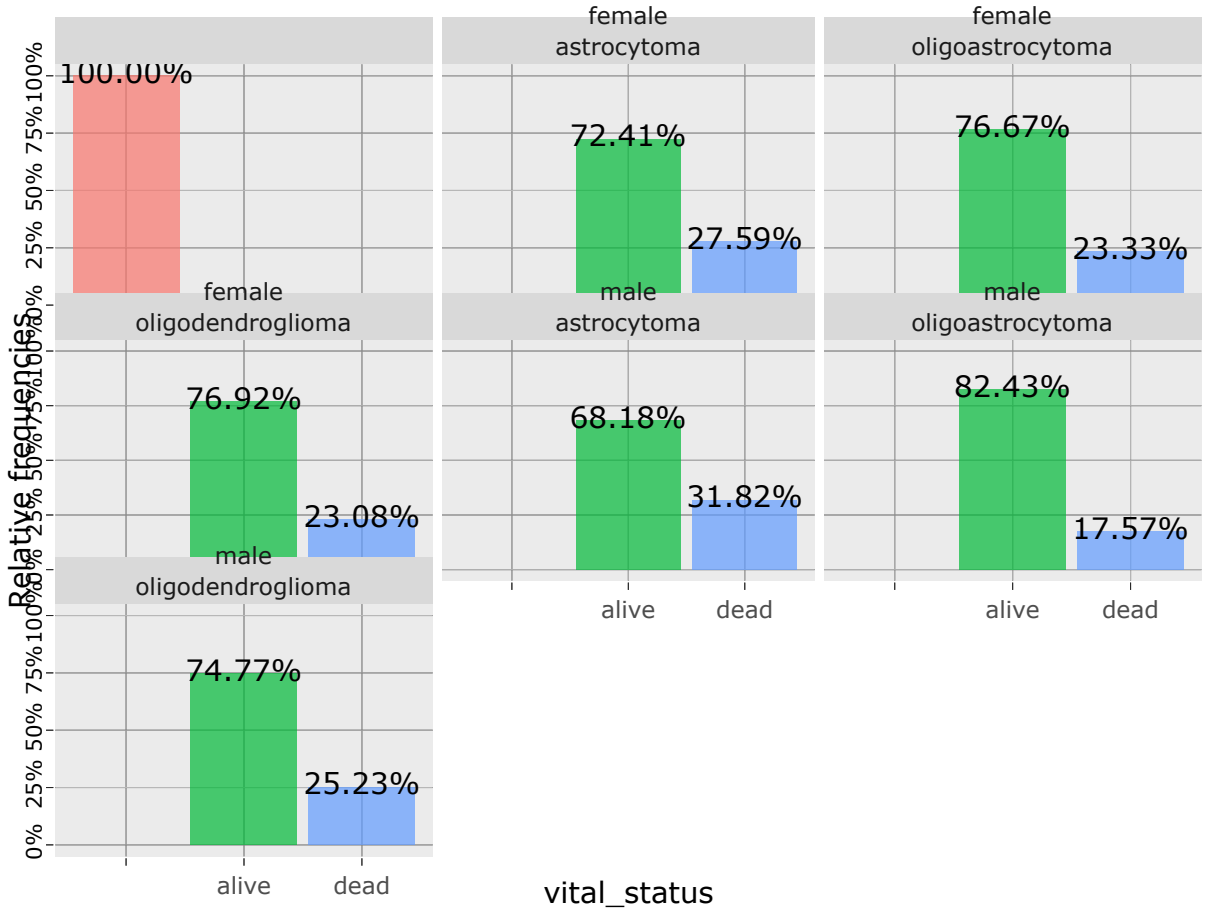
Family History and Histotype

```
ggplotly(ggplot(metadata,aes(family_history_of_primary_brain_tumor,group = gender))+
  geom_bar(aes(y=..prop..,fill = factor(..x..)),stat = "count",alpha=0.7)+
  scale_y_continuous(labels=scales::percent)+
    geom_text(aes( label = scales::percent(..prop..),
                   y= ..prop.. ), stat= "count", vjust = 0)+
  ylab("Relative frequencies")+
  theme(axis.text.y = element_text(colour = "black", size = 8, angle = 90,  hjust = 1, vjust = 1
))+
  facet_wrap(vars(gender,histological_type))+
    theme(legend.position = "none"))
```



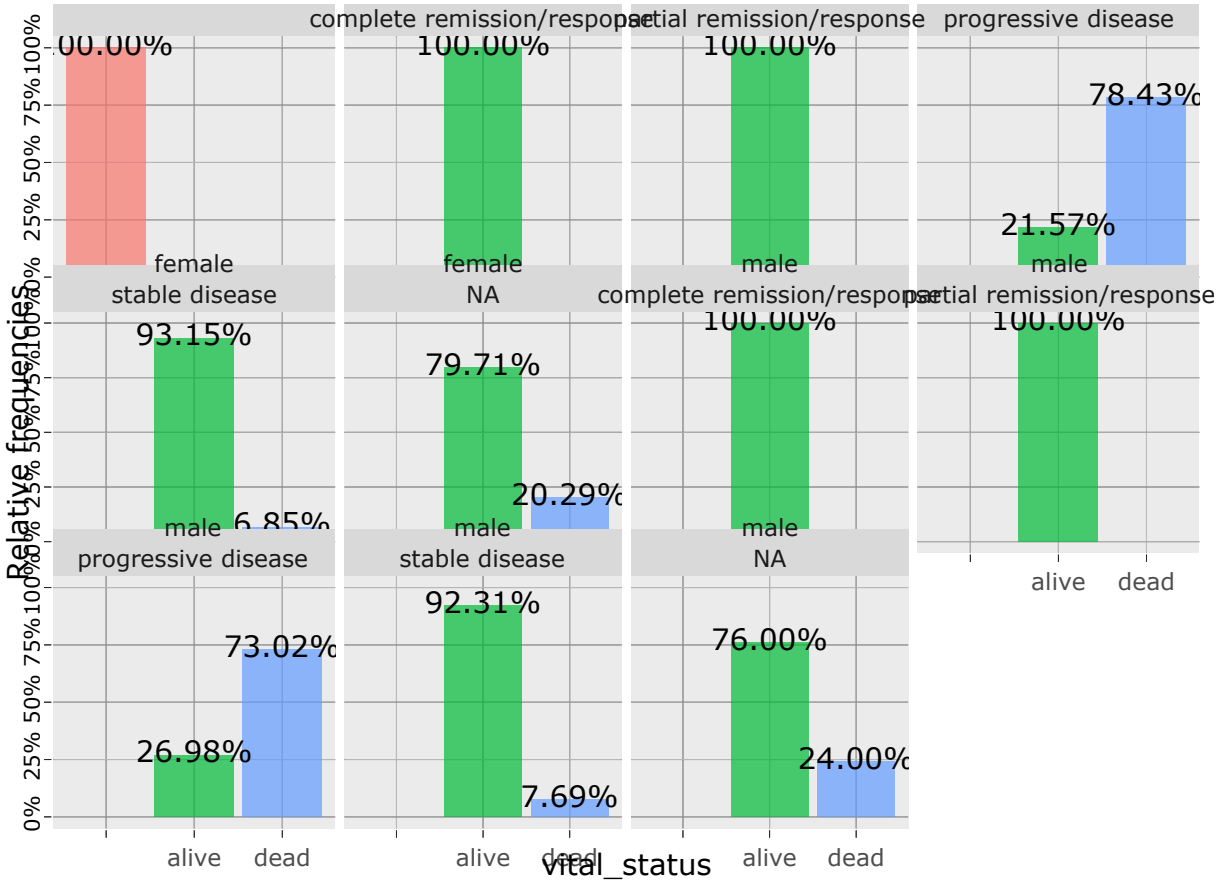
Vital Status as per histological type

```
ggplotly(ggplot(metadata,aes(vital_status,group = gender))+
  geom_bar(aes(y=..prop..,fill = factor(..x..)),stat = "count",alpha=0.7)+
  scale_y_continuous(labels=scales::percent)+
    geom_text(aes( label = scales::percent(..prop..),
                   y= ..prop.. ), stat= "count", vjust = 0)+
  ylab("Relative frequencies")+
  theme(axis.text.y = element_text(colour = "black", size = 8, angle = 90,  hjust = 1, vjust = 1
))+
  facet_wrap(vars(gender,histological_type))+
    theme(legend.position = "none"))
```



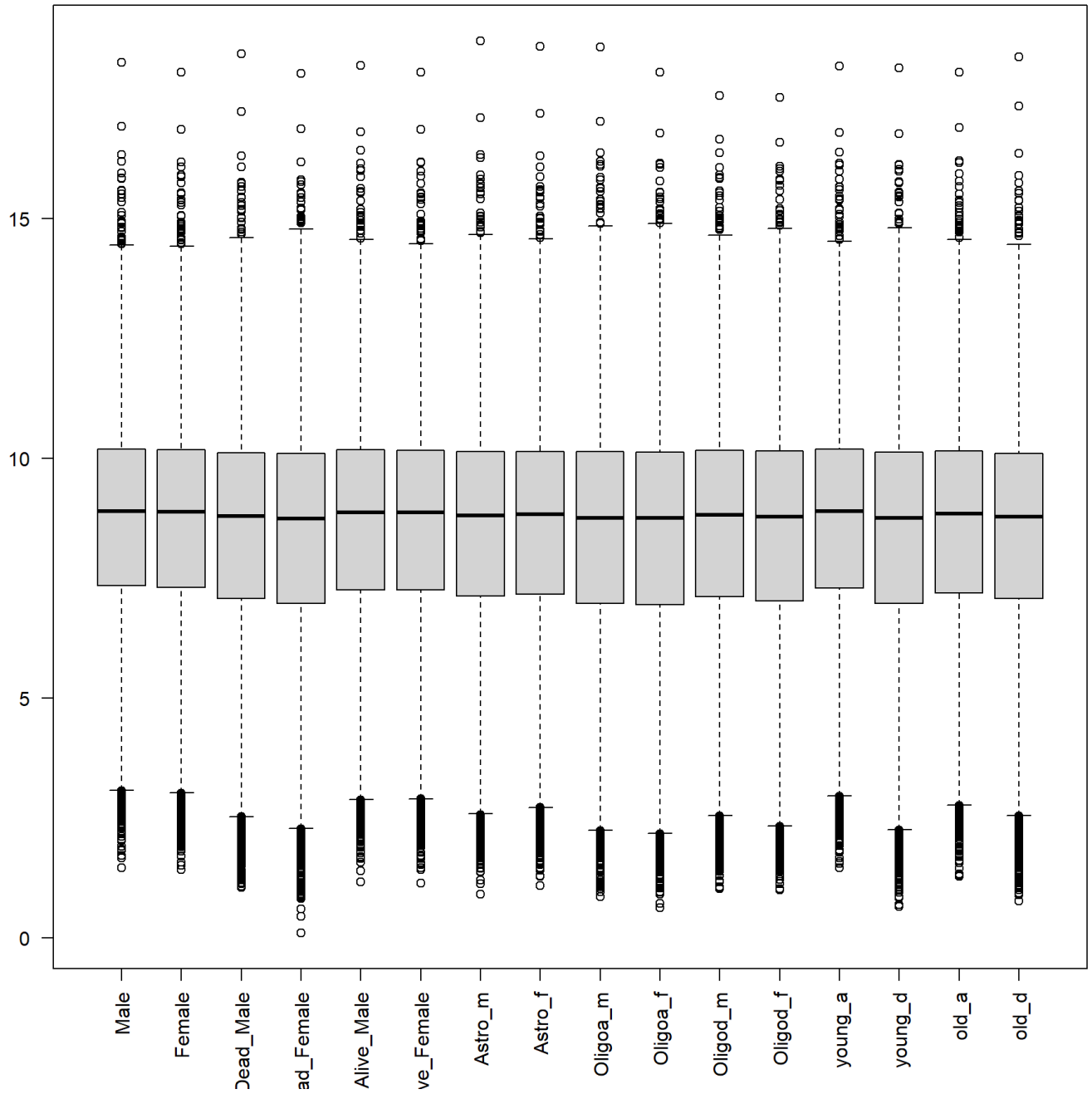
Progress death

```
ggplotly(ggplot(metadata,aes(vital_status,group = gender))+
  geom_bar(aes(y=..prop..,fill = factor(..x..)),stat = "count",alpha=0.7)+
  scale_y_continuous(labels=scales::percent)+
    geom_text(aes( label = scales::percent(..prop..),
                   y= ..prop.. ), stat= "count", vjust = 0)+
  ylab("Relative frequencies")+
  theme(axis.text.y = element_text(colour = "black", size = 8, angle = 90,  hjust = 1, vjust = 1
))+
  facet_wrap(vars(gender,followup_treatment_success))+
    theme(legend.position = "none"))
```




```
grouprna <- data.frame(Male= Male, Female= Female, Dead_Male = Dead_m, Dead_Female= Dead_f, Alive_Male= Alive_m, Alive_Female= Alive_f, Astro_m= Astro_m, Astro_f=Astro_f, Oligoa_m=Oligoa_m, Oligoa_f= Oligoa_f, Oligod_m= Oligod_m, Oligod_f= Oligod_f, young_a= young_a, young_d=young_d, old_a= old_a, old_d= old_d)

rownames(grouprna)<-rownames(gene_count)
```

```
boxplot(grouprna, las=2)
```

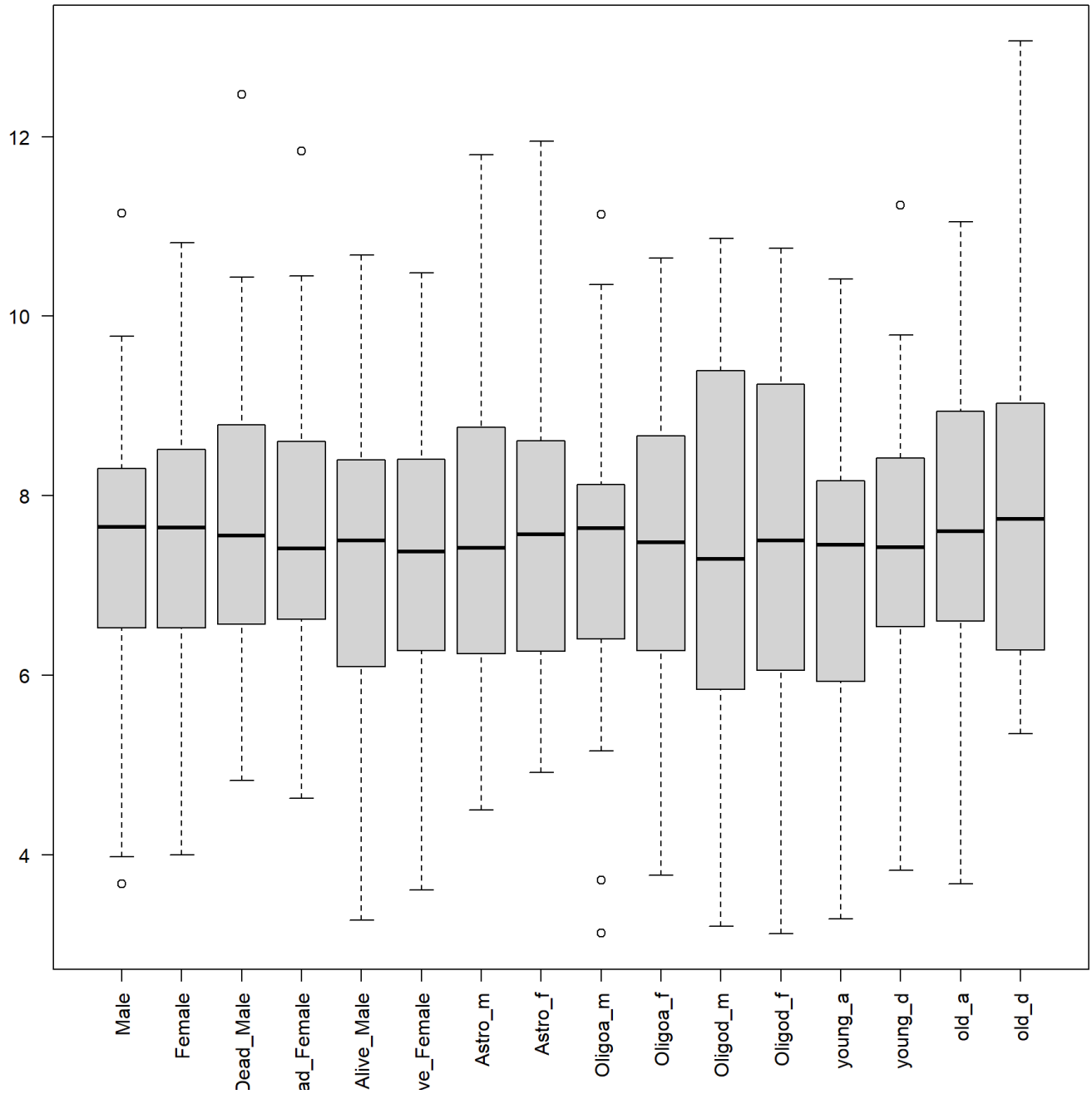


Taking maximum variable genes from the rna reads

```
vargenes <- apply(grouprna,1, var) # saving variables in a vector
top20genes <- names(sort(vargenes,decreasing = T))[1:20] # Extracting Top 20 genes
```

```
grouprna[apply(grouprna,1,function(x)all(x>=0)),]->grouprna
# Removing the rows with negative values of expression

grouprnagender <- grouprna



topvargenes<- grouprna[top20genes,]
```

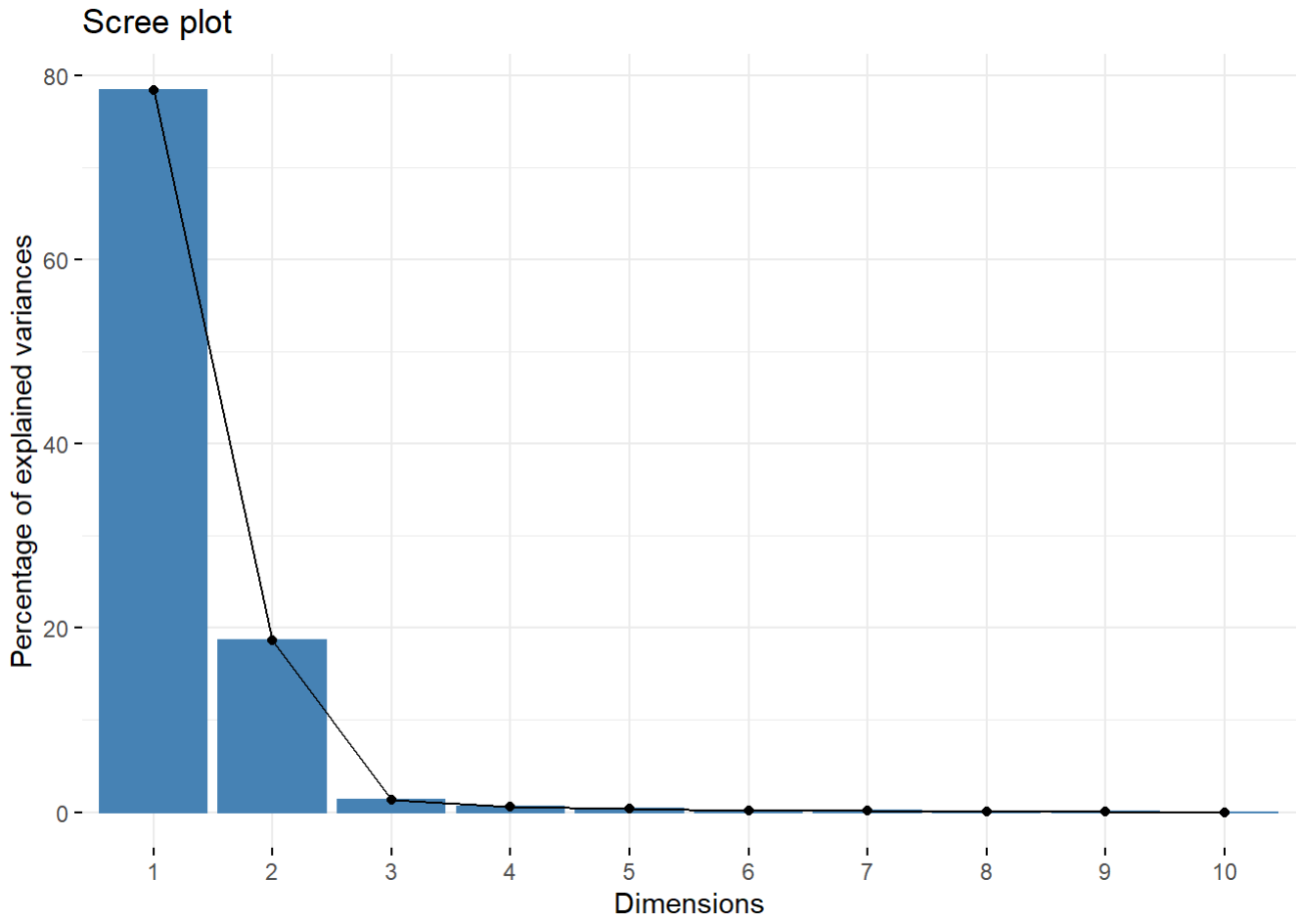
```
boxplot(topvargenes, las=2)
```



PCA on top 20 highest variable genes

Pca by prcomp function and factoextra

```
pca.res<-prcomp(t(topvargenes),scale. = T)
fviz_eig(pca.res)
```



Calculating the covariance matrix for the above data:

```
res <- cor(t(new_data_pca))
corrplot(res, type = "lower", order = "hclust",
         tl.col = "blue", tl.srt = 45)
```

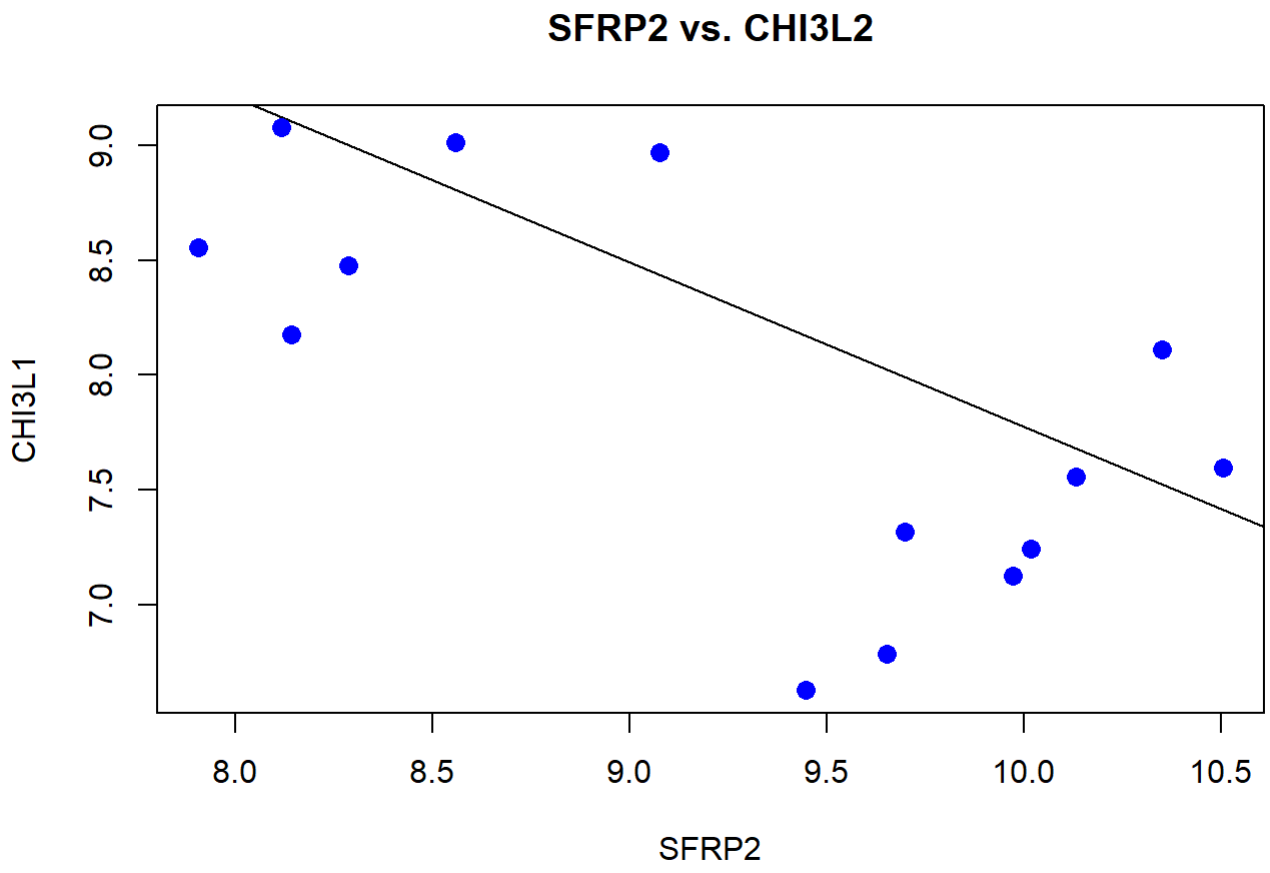
```
## Error in corrplot(res, type = "lower", order = "hclust", tl.col = "blue", : could not find fu
nction "corrplot"
```

Part 3: Hypothesis And Testing

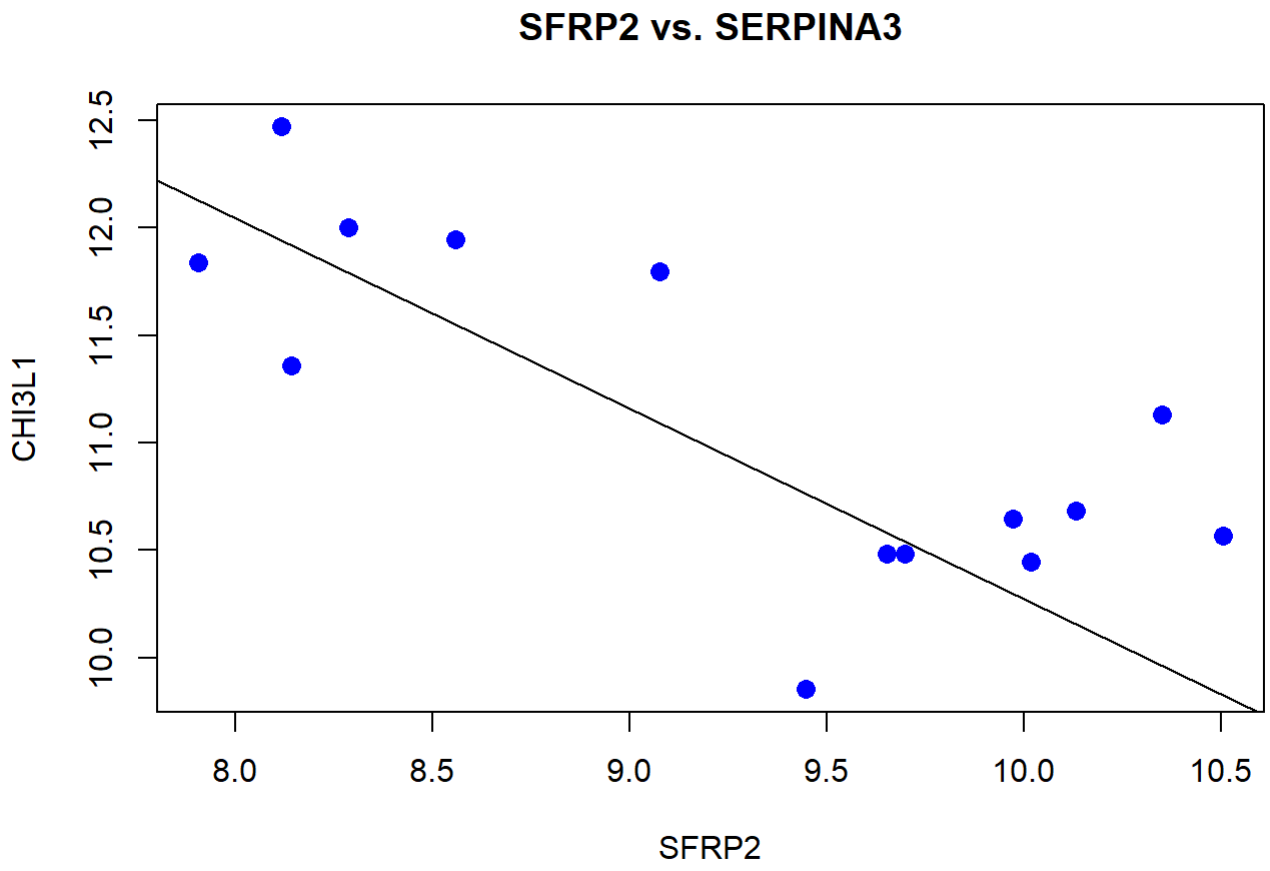
Hypothesis 1:

Hypothesis: : SFRP2 has high negative correlation with most of the other high variance genes
To test this visually, we can fit a linear regression model between SFRP2 and the genes it shows negative correlation with.

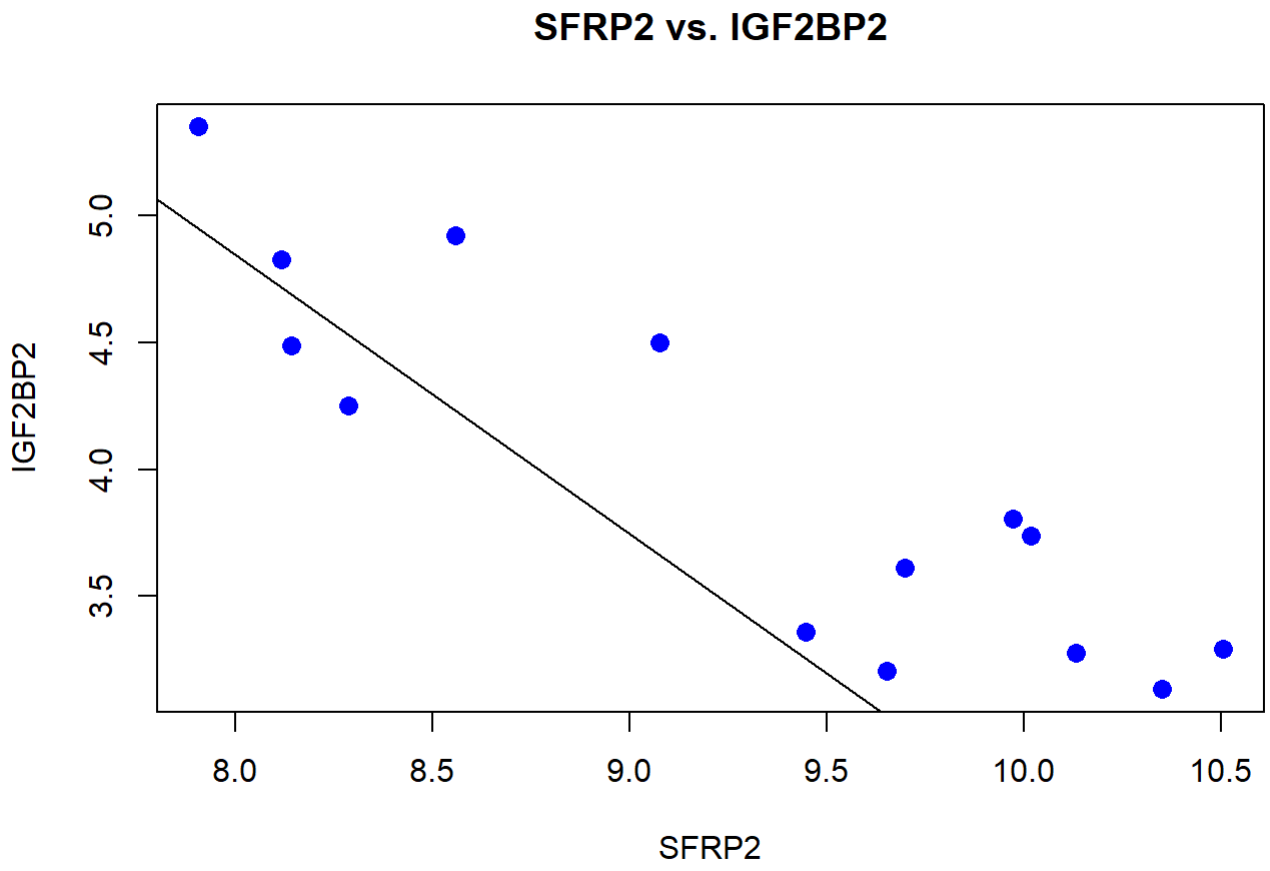
```
test = simplify2array(new_data_pca['SFRP2',])
test2 = simplify2array(new_data['CHI3L2',])
relation <- lm(test~test2)
plot(test,test2,col = "blue",main = "SFRP2 vs. CHI3L2", abline(lm(test~test2)),cex = 1.3,pch = 1
6,xlab = "SFRP2",ylab = "CHI3L1")
```



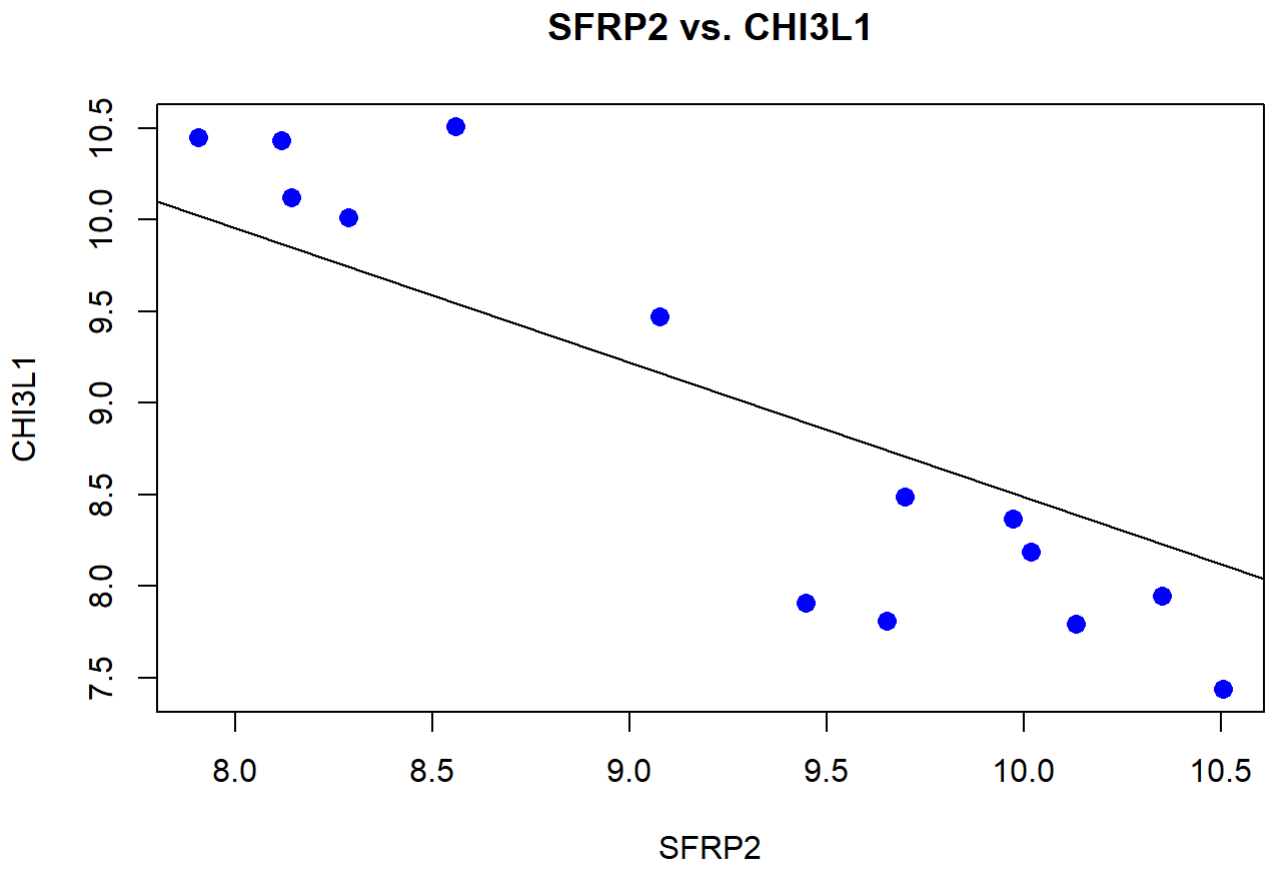

```
test = simplify2array(new_data_pca['SFRP2',])
test2 = simplify2array(new_data['SERPINA3',])
relation <- lm(test~test2)
plot(test,test2,col = "blue",main = "SFRP2 vs. SERPINA3", abline(lm(test~test2)),cex = 1.3,pch =
16,xlab = "SFRP2",ylab = "CHI3L1")
```



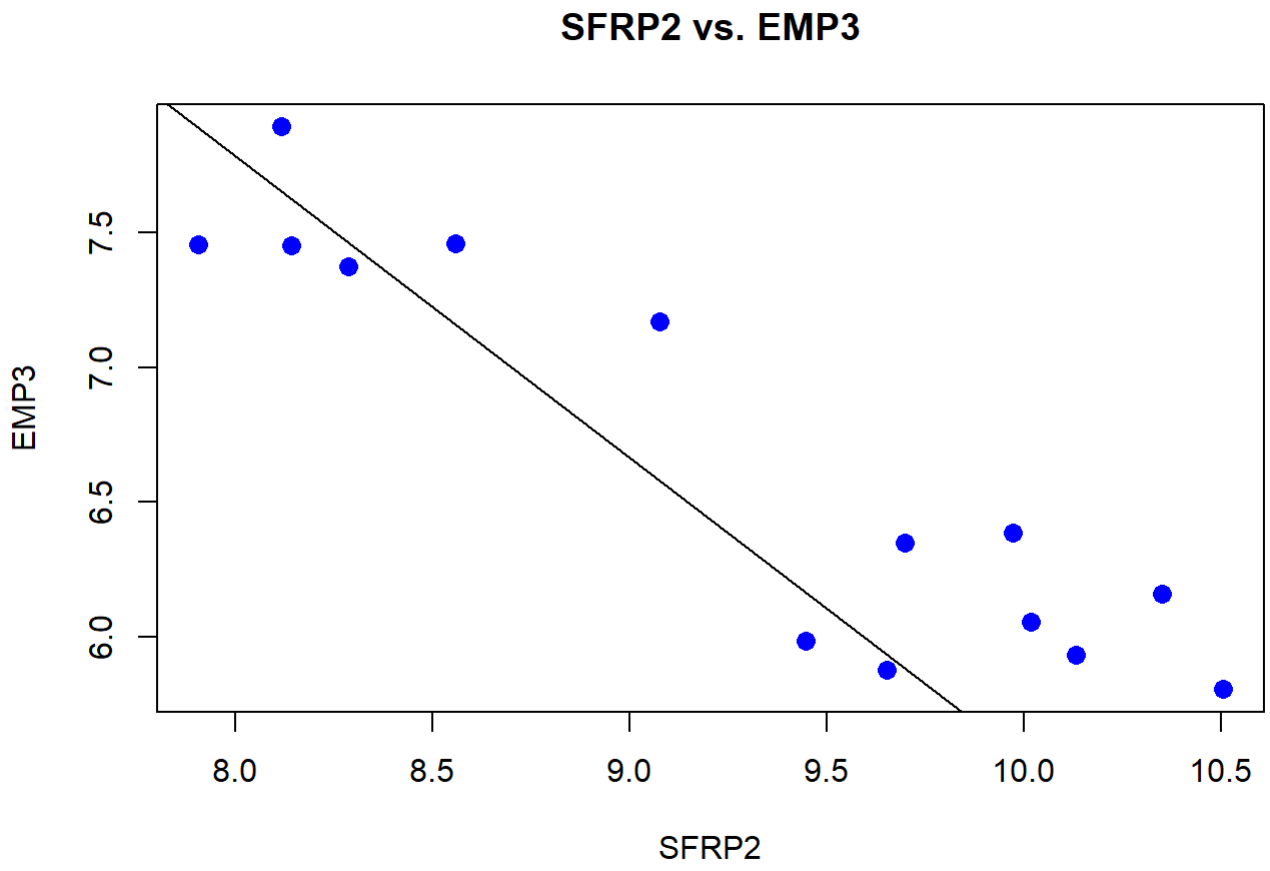

```
test = simplify2array(new_data_pca['SFRP2',])
test2 = simplify2array(new_data['IGF2BP2',])
relation <- lm(test~test2)
plot(test,test2,col = "blue",main = "SFRP2 vs. IGF2BP2", abline(lm(test~test2)),cex = 1.3,pch =
16,xlab = "SFRP2",ylab = "IGF2BP2")
```




```
test = simplify2array(new_data_pca['SFRP2',])
test2 = simplify2array(new_data['CHI3L1',])
relation <- lm(test~test2)
plot(test,test2,col = "blue",main = "SFRP2 vs. CHI3L1", abline(lm(test~test2)),cex = 1.3,pch = 1
6,xlab = "SFRP2",ylab = "CHI3L1")
```




```
test = simplify2array(new_data_pca['SFRP2',])
test2 = simplify2array(new_data['EMP3',])
relation <- lm(test~test2)
plot(test,test2,col = "blue",main = "SFRP2 vs. EMP3", abline(lm(test~test2)),cex = 1.3,pch = 16,
xlab = "SFRP2",ylab = "EMP3")
```

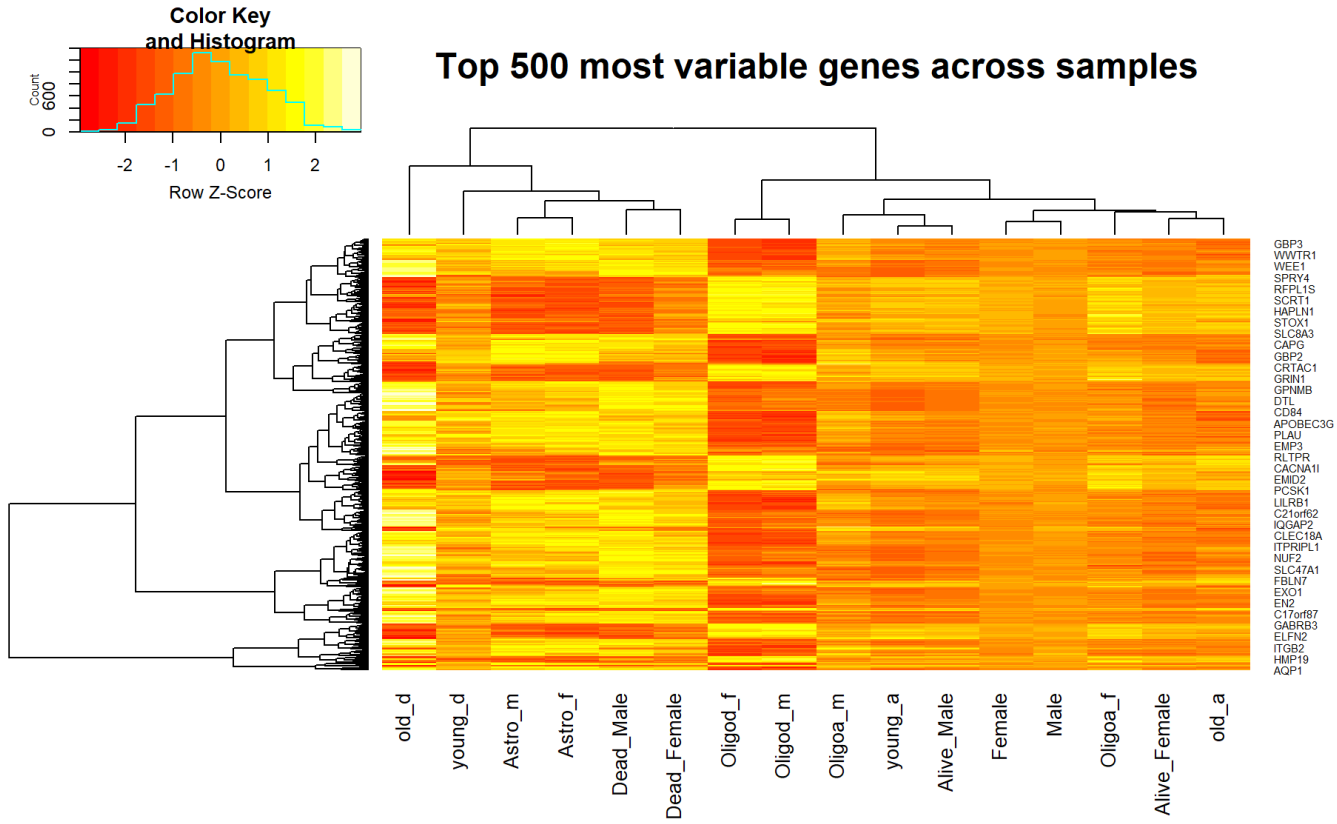


Hence the above graphs prove the negative correlation of the SFRP2 gene with most of the other high variance genes

Heatmap on Highy Expressed Genes

```
top500genes<- names(sort(vargenes,decreasing = T))[1:500]
top500highvar<-grouprna[top500genes,]
```

```
heatmap.2(as.matrix(top500highvar),trace="none",
          main="Top 500 most variable genes across samples",
          scale="row",margins = c(8, 7))->a
```



The Above Heatmap denotes expression of genes in different categories we have divided, color key is as following - The more positively it is more highly expressed it is - The more red it is, it denotes downregulated geens

Also, Phylogenetically closely placed columns are similar in genes expression, for example **Oligod_f** and **Oligod_m** have similar gene expressions, similarily **Astro_m** and **Astro_f** has same level of expressions.

Z Score tells the standard deviation of the expression values from the mean

In the bottom genes in the heatmap, for example **TMSB4Y** gene is a **Y-chromosomal** gene so it is specifically expressed in Males.

Subsetting New metadata into male and female part:

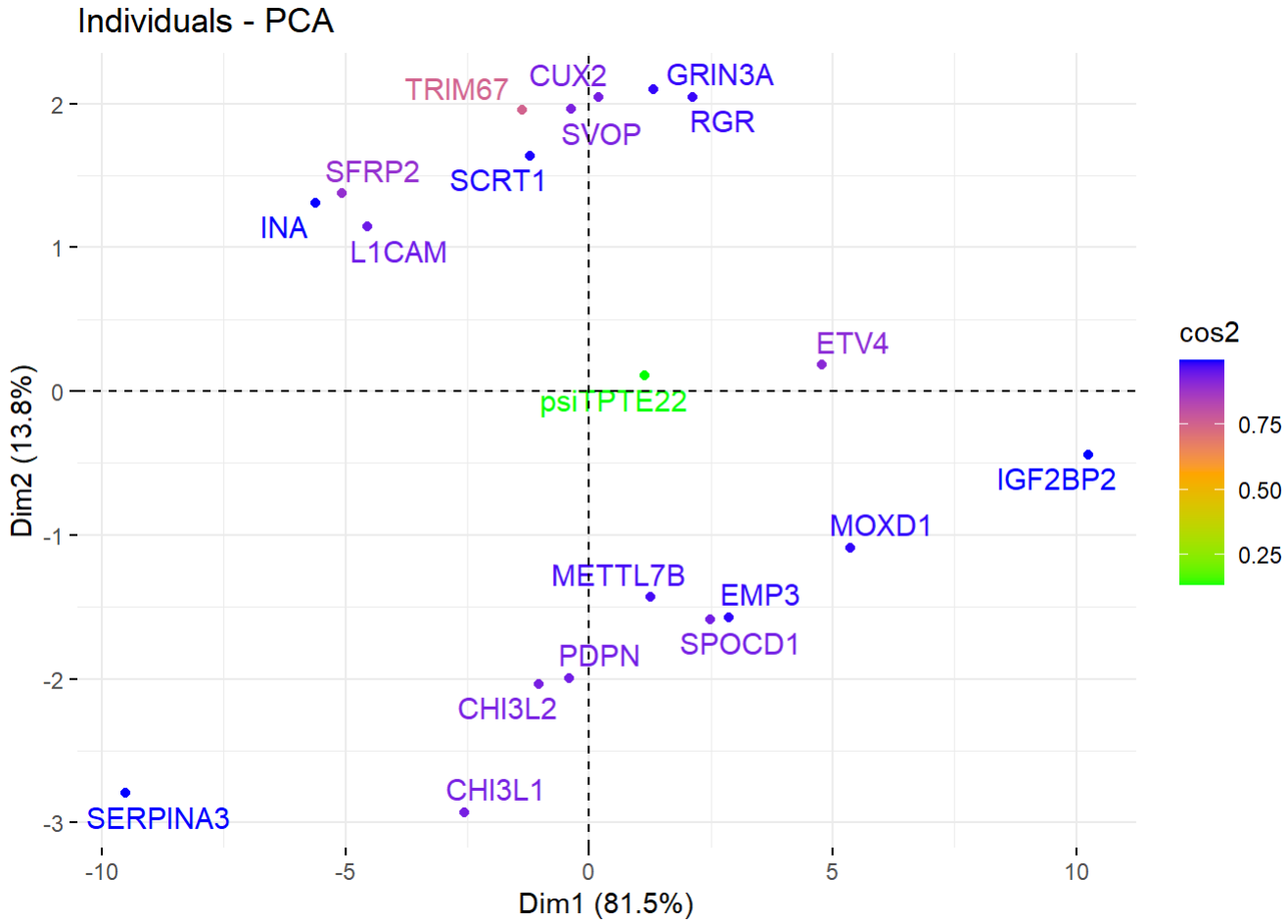
```
new_data_pca_male <- new_data_pca[c("Dead_Male","Alive_Male","Astro_m","Oligoa_m","Oligod_m","yo
ung_m","old_m")]
```

```
new_data_pca_female <- new_data_pca[c("Dead_Female","Alive_Female","Astro_f","Oligoa_f","Oligod_
f","young_f","old_f")]
```

PCA on Male data:

```
pca_male <- prcomp(new_data_pca_male)
```

```
#Graph of individuals. Individuals with a similar profile are grouped together.
fviz_pca_ind(pca_male,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("green", "orange", "blue"),
             repel = TRUE # Avoid text overlapping
)
```



Based on features that we extracted from metadata for Male patients suffereing from Brain Lower Grade Glioma, we can find clear clustering of high variance genes in above graphs as follows.

Cluster 1: TRIM67, GRIN3A, SVOP, CUX2, SCRT1, RGR

Cluster 2: ETV4, psiTPTE22

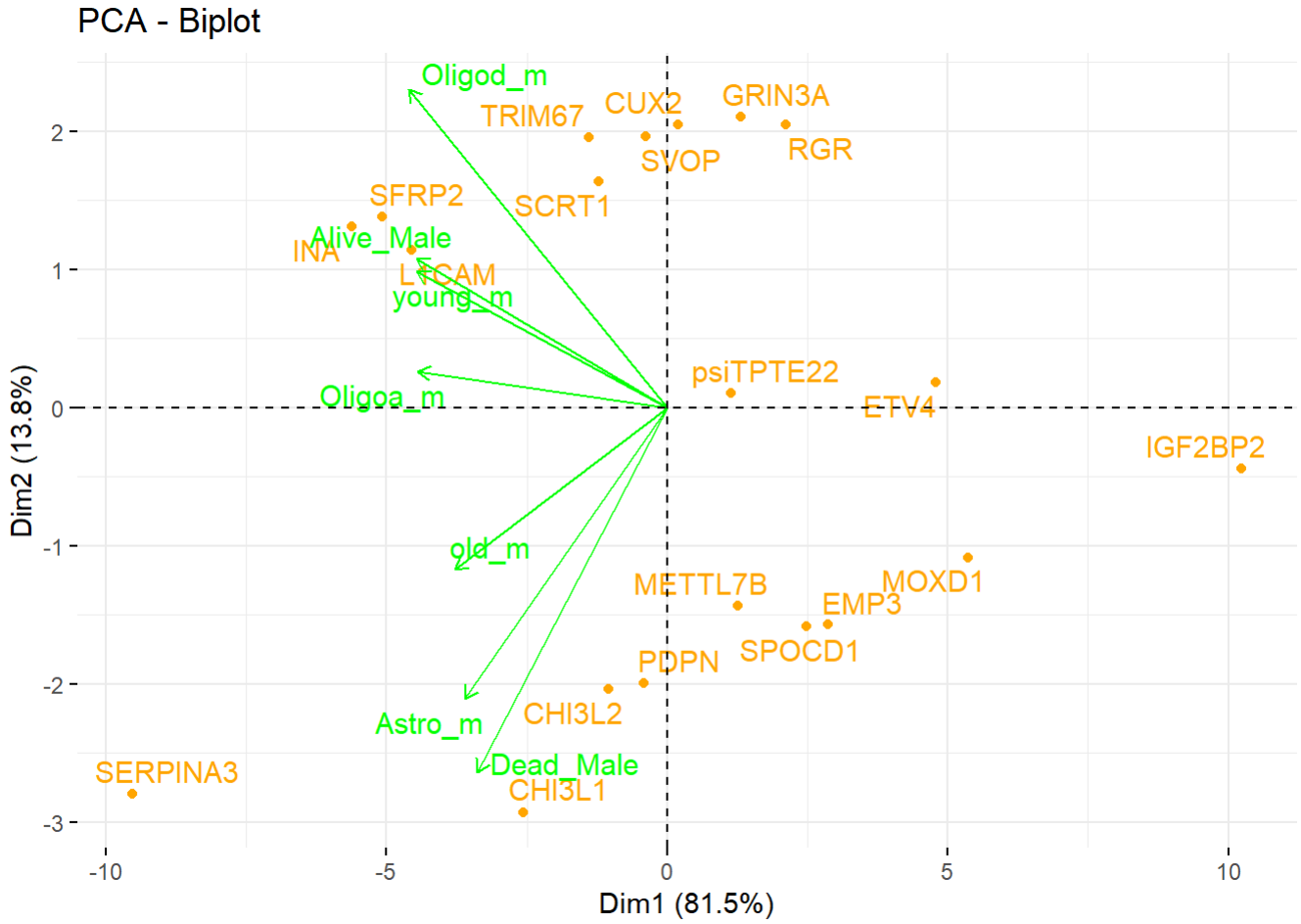
Cluster 3: L1CAM, INA, SFRP2

Cluster 4: SERPINA3

Cluster 5: CHI2L1, CHI3L2, PDPN, SPOCD1, EMP3, METTL7B, MOXD1

Cluster 6: IGF2BP2

```
#Graph of variables and individuals. Positive correlated variables point to the same side of the
plot. Negative correlated variables point to opposite sides of the graph.
fviz_pca_biplot(pca_male, repel = TRUE,
                col.var = "green", # Variables color
                col.ind = "orange" # Individuals color
)
```



From the PCA biplot above we can see that the eigenvectors for Astro_m and Dead_male point in the same direction and the eigenvectors for Alive_male, young_m and Oligoa_m point in same direction. So we can say that the vectors that point in same direction are positively correlated. Since we are looking for a small sets of genes, we test these observations for the bigger set by checking for following hypotheses.

Hypothesis 2:

Hypothesis: Male patients with Astrocytoma histology type tumor are more likely to die.

To check this above hypothesis we will do following steps:

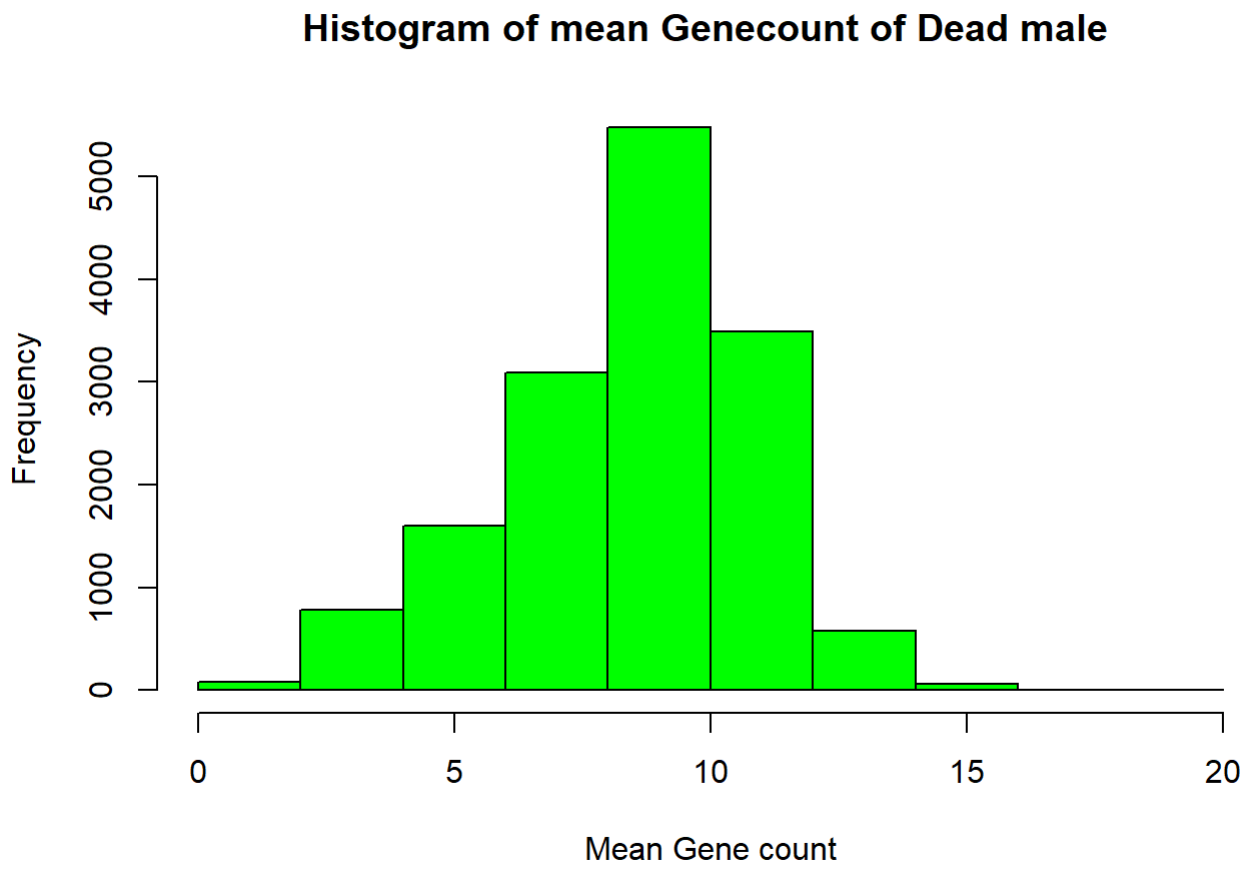
- Check if gene count for Dead male and Astrocytoma male is normally distributed or not.
- Using linear regression, visualizing correlation between gene count for Dead male and Astrocytoma male.
- Correlation test between gene count for Dead male and Astrocytoma male.

Note that: to perform all the above data analysis we will using original data (ie. my_df) where we haven't replaced NA values with mean and we will be considering gene_count for top 5000 hight variance genes.

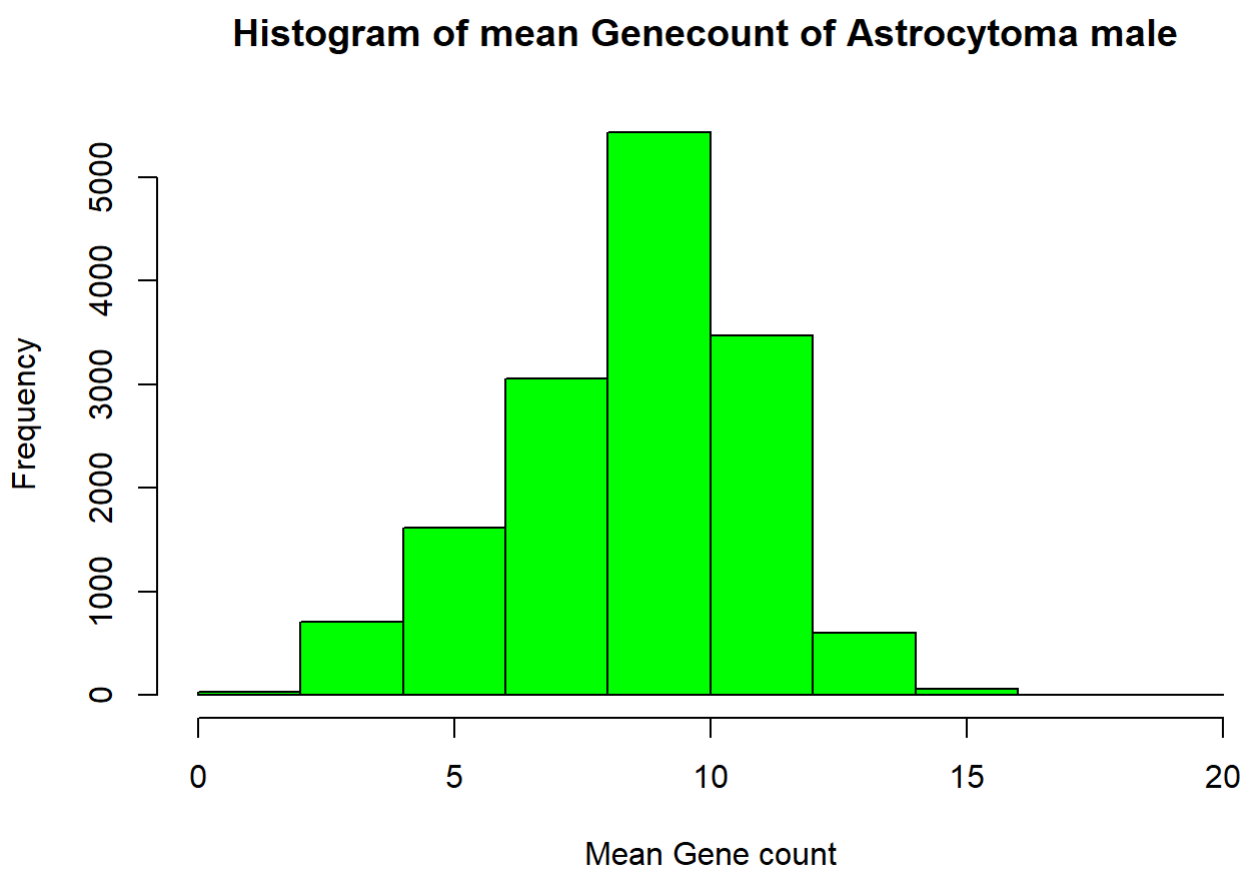
```
top_5000_genes<- names(sort(var_genes,decreasing = T))[1:5000]

new_data_ana<- my_df[top_5000_genes,]
```

Normality Test:

```
# plotting histogram:

hist(my_df$Dead_Male,breaks = 10,col = c("green"),xlab = "Mean Gene count",main = "Histogram of
 mean Genecount of Dead male ")
```




```
hist(my_df$Astro_m,breaks = 10,col = c("green"),xlab = "Mean Gene count",main = "Histogram of me
an Genecount of Astrocytoma male ")
```



From the histograms it seems that both gene count for Dead male and Astrocytoma male are distributed normally. Although we will confirm this result using Shapiro-Wilk normality test. We define Shapiro-Wilk test formally as,

Null Hypothesis: The data sample is Normally distributed. **Alternate Hypotheis:** The data sample is not Normally distributed.

```
shapiro.test(new_data_ana$Dead_Male)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_data_ana$Dead_Male
## W = 0.99682, p-value = 7.877e-09
```

```
shapiro.test(new_data_ana$Astro_m)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_data_ana$Astro_m
## W = 0.99609, p-value = 2.923e-10
```

since p-value for data samples it much less than critical value 0.05. Thus we must reject our Null hypothesis and get that gene count for Dead male and Astrocytoma male are **not** distributed normally.

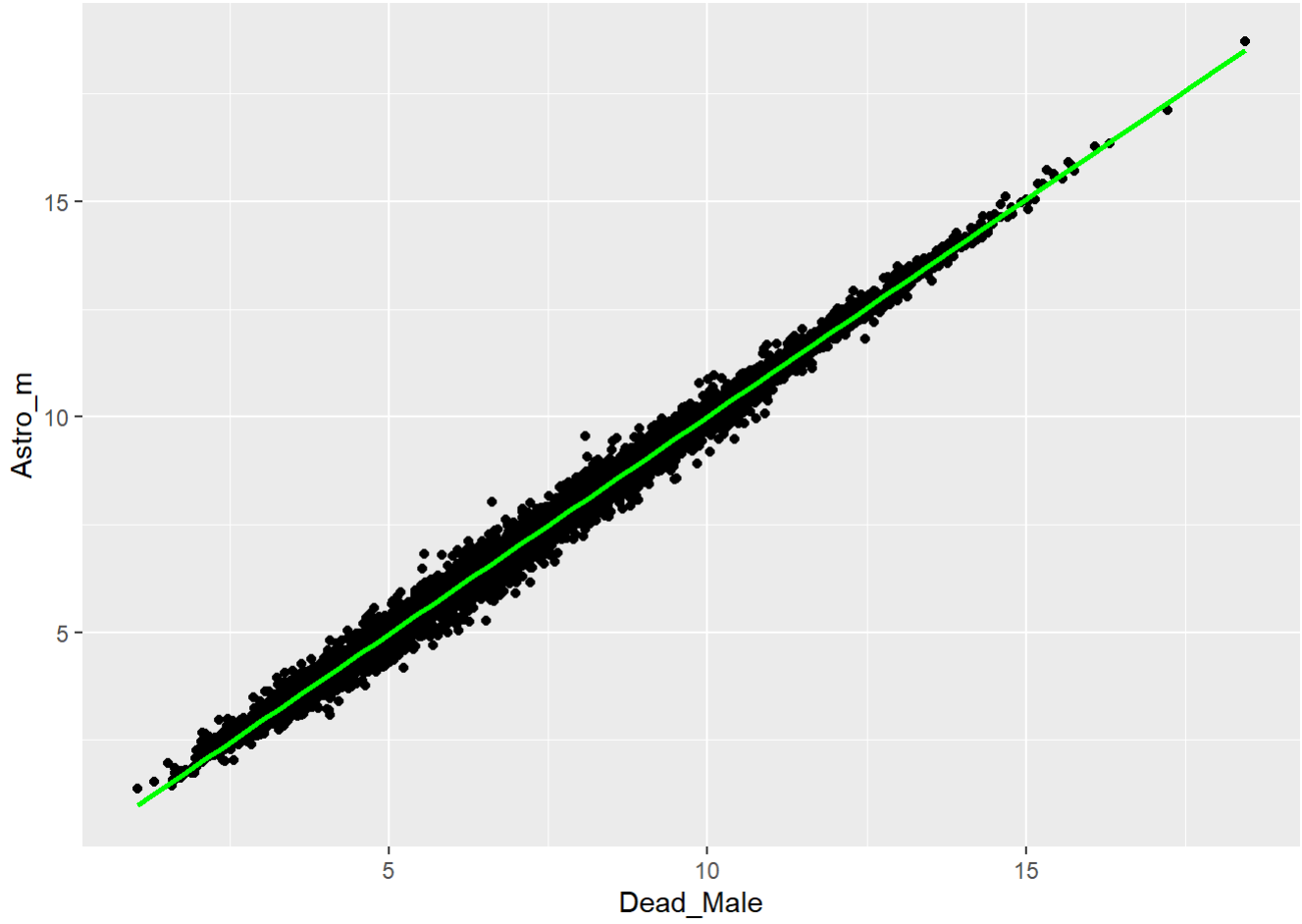
Linear Regression:

```
p1 <- ggplot(my_df, aes(x=Dead_Male, y=Astro_m)) +
  geom_point() +
  geom_smooth(method=lm , color="green", fill="blue", se=TRUE)

p1
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 3496 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3496 rows containing missing values (geom_point).
```



One can observe that mean gene count in Male patients that die and male patients with Astrocytoma histology type tumor is linearly related. Hence we can say they are positivey correlated. For further analysis we can perform correlation test on these two samples.

Correlation test:

Since the gene count for Dead male and Astrocytoma male are **not** distributed normally we will hae to use non parametric correlation test. Here, we will use Spearman test. To perform Spearman Correlation test we will use cor.test() function available in R. Formally we define cor.test as:

Null Hypothesis: The X population vector in not correlated to Y.

Alternate Hypothesis: The X population vector is correlated to Y with the given correlation index.

```
cor.test(my_df$Dead_Male,my_df$Astro_m ,method = 'spearman', use = 'pairwise.complete.obs',exact
= F)
```

```
##
##  Spearman's rank correlation rho
##
## data:  my_df$Dead_Male and my_df$Astro_m
## S = 1688459415, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9968851
```

Since the p-value is mush less than cricital value 0.05. Thus we reject our Null hypothesis and get that gene count for Dead male and Astrocytoma male is positively correlated with high correlation index.

Conclusion:

Looking at the above correlation test and other ananlysis we observe that mean gene count in Male patients that die and male patients with Astrocytoma histology type tumor is positively correlated i.e. when a particular gene is expressed more in tumor of dead male, it is highly likely that the same gene is expressed more in tumor cell of male with Astrocytoma histology type tumor. Hence, we caonclude that male with Astrocytoma histology type tumor are more likely to die.

Hypothesis 3:

Hypothesis: Young male patients with Oligoastrocytoma histology type tumor are more likely to live.

To check this above hypothesis we will do following steps:

- Check if mean gene count for Alive male and Young male patients with Oligoastrocytoma histology type tumor is normally distributed or not.
- Using linear regression, visualizing correlation between gene count for Dead male and Young male patients with Oligoastrocytoma histology type tumor.
- Correlation test between gene count for Dead male and Young male patients with Oligoastrocytoma histology type tumor.

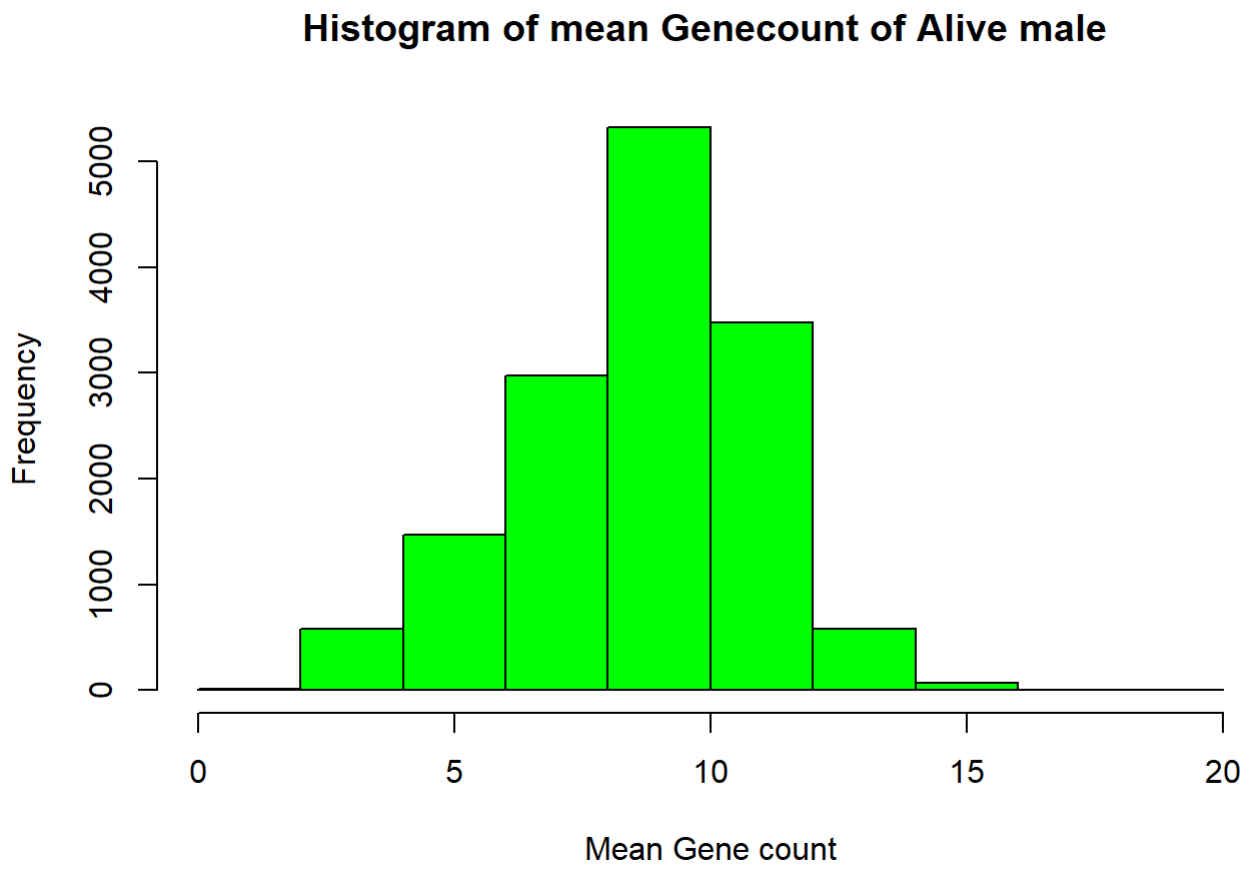
Note that: to perform all the above data analysis we will using original data (ie. my_df) where we haven't replaced NA values with mean and we will be considering gene_count for top 5000 hight variance genes.

```
# getting mean of gene_count for young male patients with Oligoastrocytoma histology type tumor

my_df$Oligod_m_young <- apply(as_tibble(gene_count[,metadata[metadata$gender == "male" & metadat
a$histological_type == "oligodendroglioma" & metadata$age_at_initial_pathologic_diagnosis <= 45,
1]]),1,mean)
```

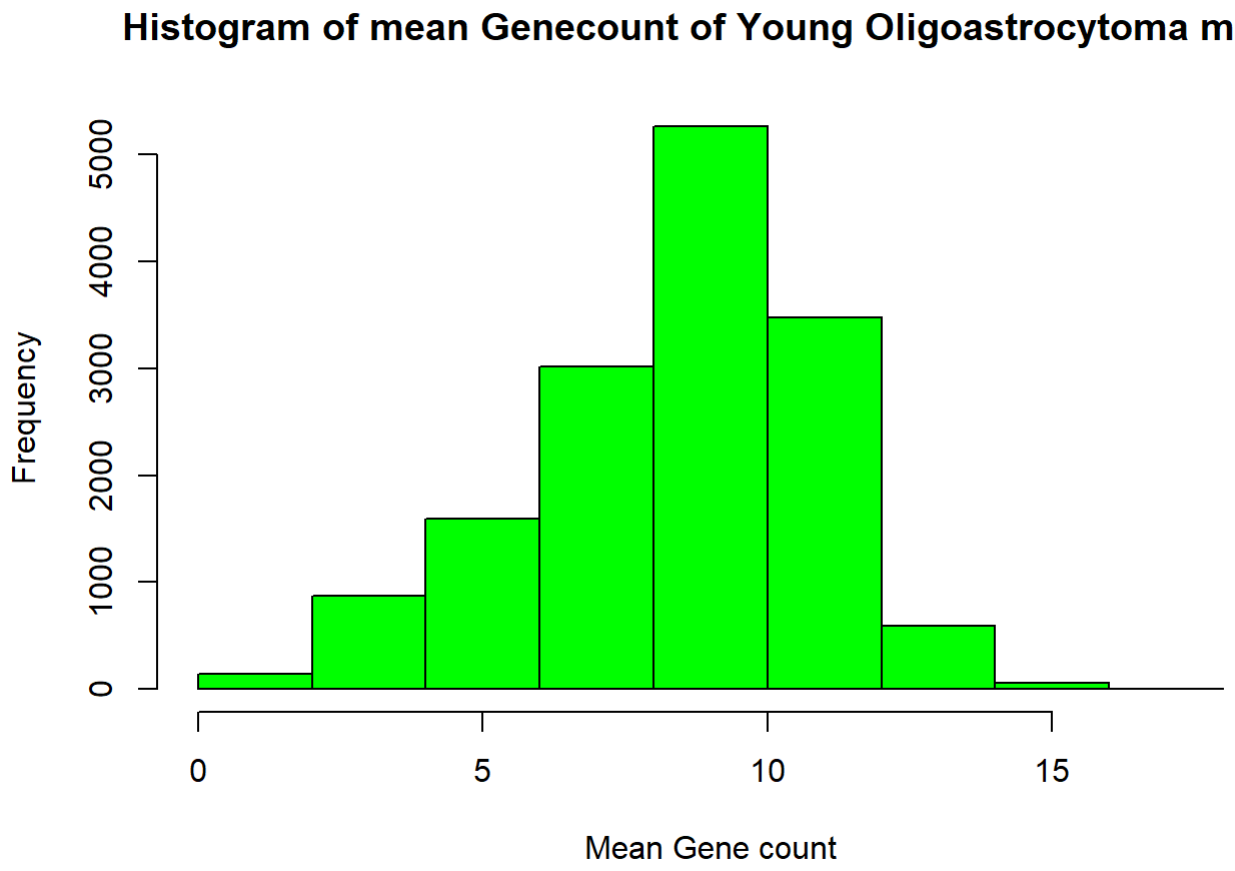
```
new_data_ana1<- my_df[top_5000_genes,]
```

Normality Test:

```
# plotting histogram:

hist(my_df$Alive_Male,breaks = 10,col = c("green"),xlab = "Mean Gene count",main = "Histogram of mean Genecount of Alive male ")
```




```
hist(my_df$Oligod_m_young,breaks = 10,col = c("green"),xlab = "Mean Gene count",main = "Histogram of mean Genecount of Young Oligoastrocytoma male ")
```



From the histograms it seems that both gene count for Alive male and Young male patients with Oligoastrocytoma histology type tumor aren't distributed normally. Although we will confirm this result using Shapiro-Wilk normality test. We define Shapiro-Wilk test formally as,

Null Hypothesis: The data sample is Normally ditributed. **Alternate Hypotheis:** The data sample is not Normally ditributed.

```
shapiro.test(new_data_ana1$Alive_Male)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_data_ana1$Alive_Male
## W = 0.99576, p-value = 7.589e-11
```

```
shapiro.test(new_data_ana1$Oligod_m_young)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_data_ana1$Oligod_m_young
## W = 0.99625, p-value = 5.824e-10
```

since p-value for data samples it much less than critical value 0.05. Thus we must reject our Null hypothesis and get that gene count for Alive male and Young Oligoastrocytoma male are **not** distributed normally.

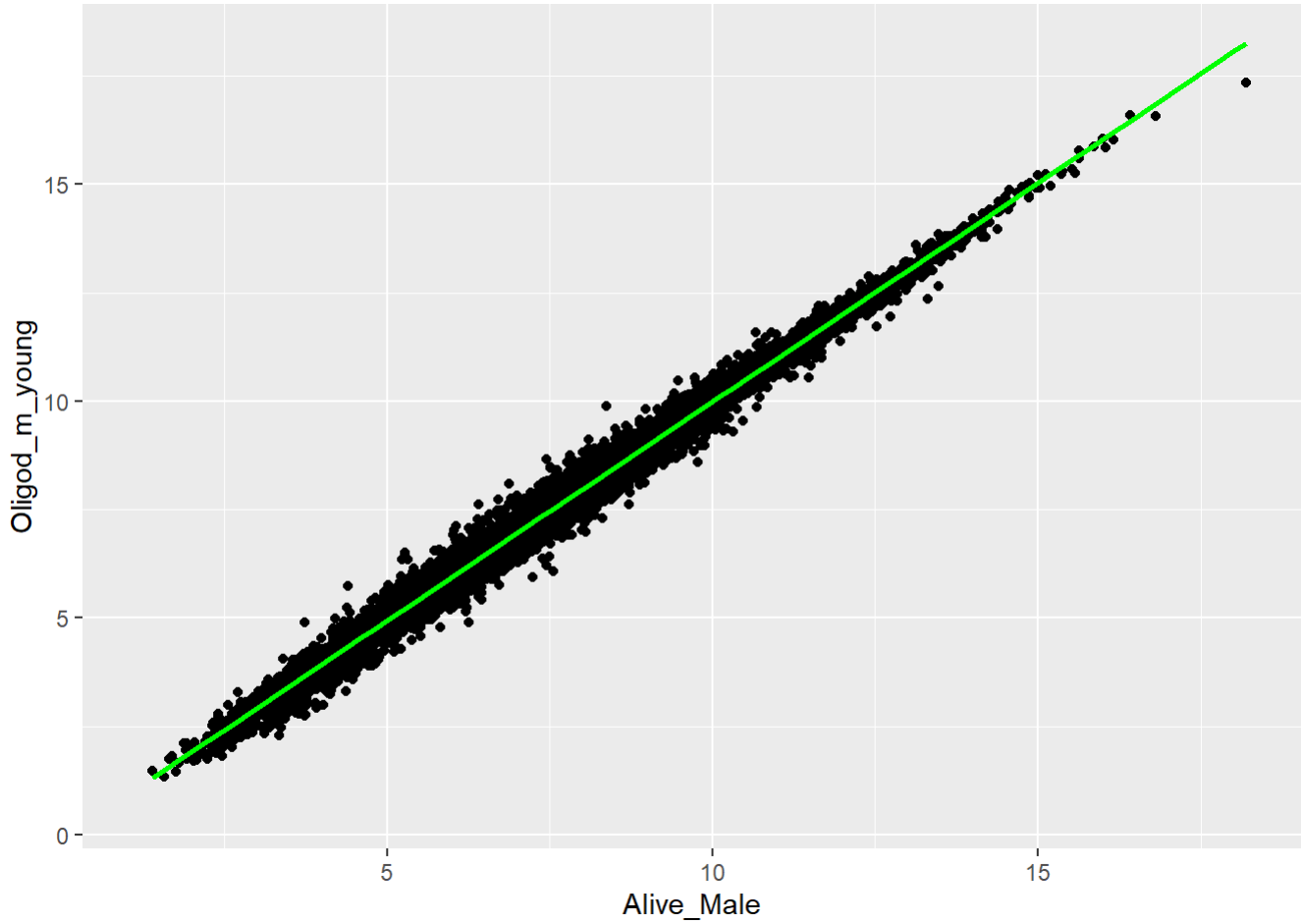
Linear Regression:

```
p2 <- ggplot(my_df, aes(x=Alive_Male, y=Oligod_m_young)) +
  geom_point() +
  geom_smooth(method=lm , color="green", fill="#69b3a2", se=TRUE)

p2
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 3846 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3846 rows containing missing values (geom_point).
```



One can observe that mean gene count in Male patients that are alive and Young male patients with Oligoastrocytoma histology type tumor is linearly related. Hence we can say they are positively correlated. For further analysis we can perform correlation test on these two samples.

Correlation test:

Since both the samples in consideration are **not** distributed normally we will have to use non parametric correlation test. Here, we will use Spearman test. To perform Spearman Correlation test we will use cor.test() function available in R. Formally we define cor.test as:

Null Hypothesis: The X population vector in not correlated to Y.

Alternate Hypothesis: The X population vector is correlated to Y with the given correlation index.

```
cor.test(my_df$Alive_Male,my_df$Oligod_m_young ,method = 'spearman', use = 'pairwise.complete.ob
s',exact = F)
```

```
##
##  Spearman's rank correlation rho
##
## data:  my_df$Alive_Male and my_df$Oligod_m_young
## S = 2313329846, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.995415
```

Since the p-value is mush less than critical value 0.05. Thus we reject our Null hypothesis and get that gene count for Alive male and Young Oligoastrocytoma male is positively correlated with high correlation index.

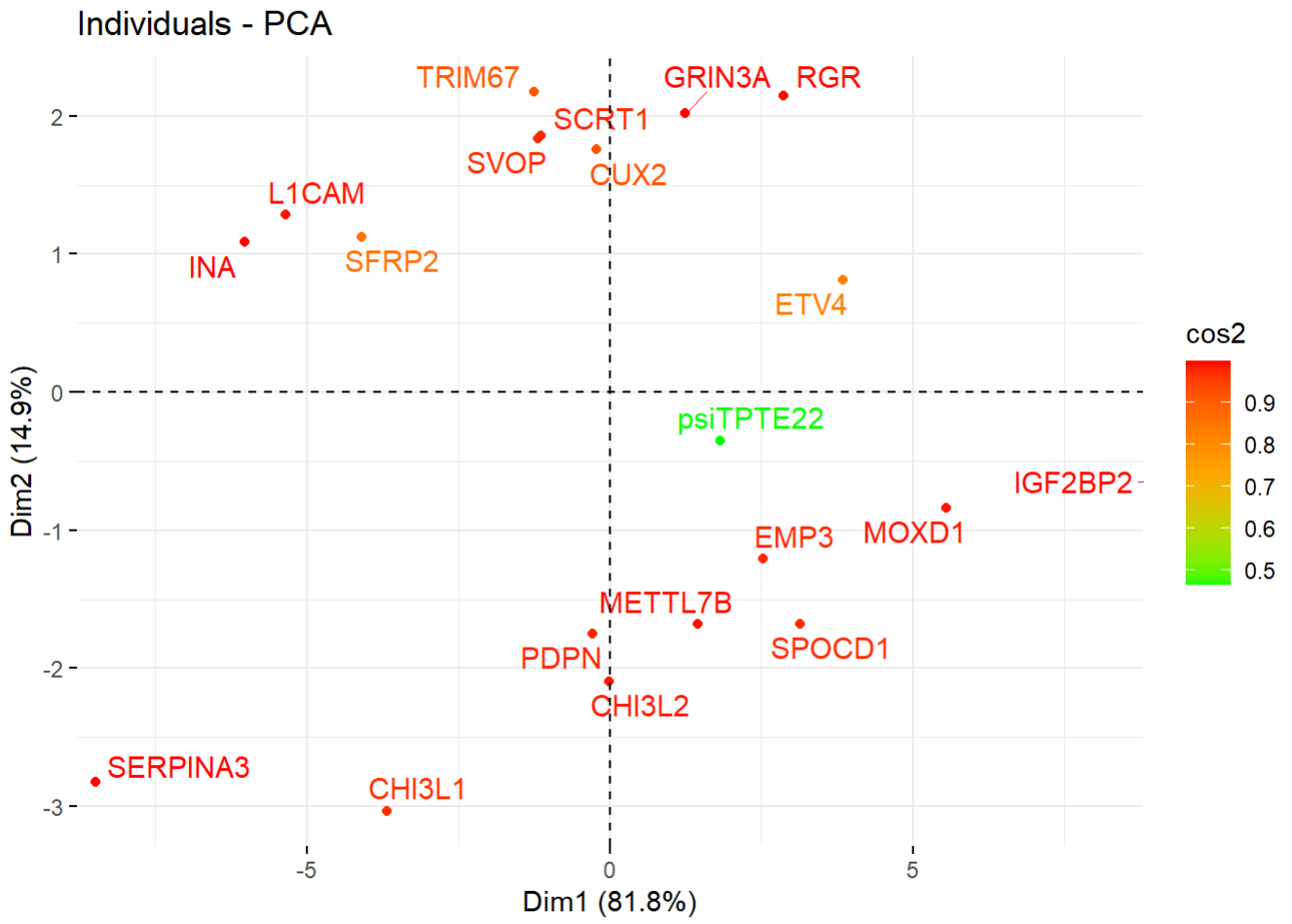
Conclusion:

Looking at the above correlation test and other analysis we observe that mean gene count in Male patients that are alive and Young male patients with Oligoastrocytoma histology type tumor is positively correlated i.e. when a particular gene is expressed more in tumor of alive male, it is highly likely that the same gene is expressed more in tumor cell of young male with Oligoastrocytoma histology type tumor. Hence, we conclude that young male with Oligoastrocytoma histology type tumor are more likely to stay alive.

PCA on Female data:

```
pca_female <- prcomp(new_data_pca_female)
```

```
#Graph of individuals. Individuals with a similar profile are grouped together.
fviz_pca_ind(pca_female,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("green", "orange", "red"),
             repel = TRUE # Avoid text overlapping
             ,xlim = c(-8,8)
)
```



Based on features that we extracted from metadata for Female patients suffering from Brain Lower Grade Glioma, we can find clear clustering of high variance genes in above graphs as follows.

Cluster 1: TRIM67, SVOP, GRIN3A, SCRT1, CUX2, RGR

Cluster 2: L1CAM, INA, SFRP2

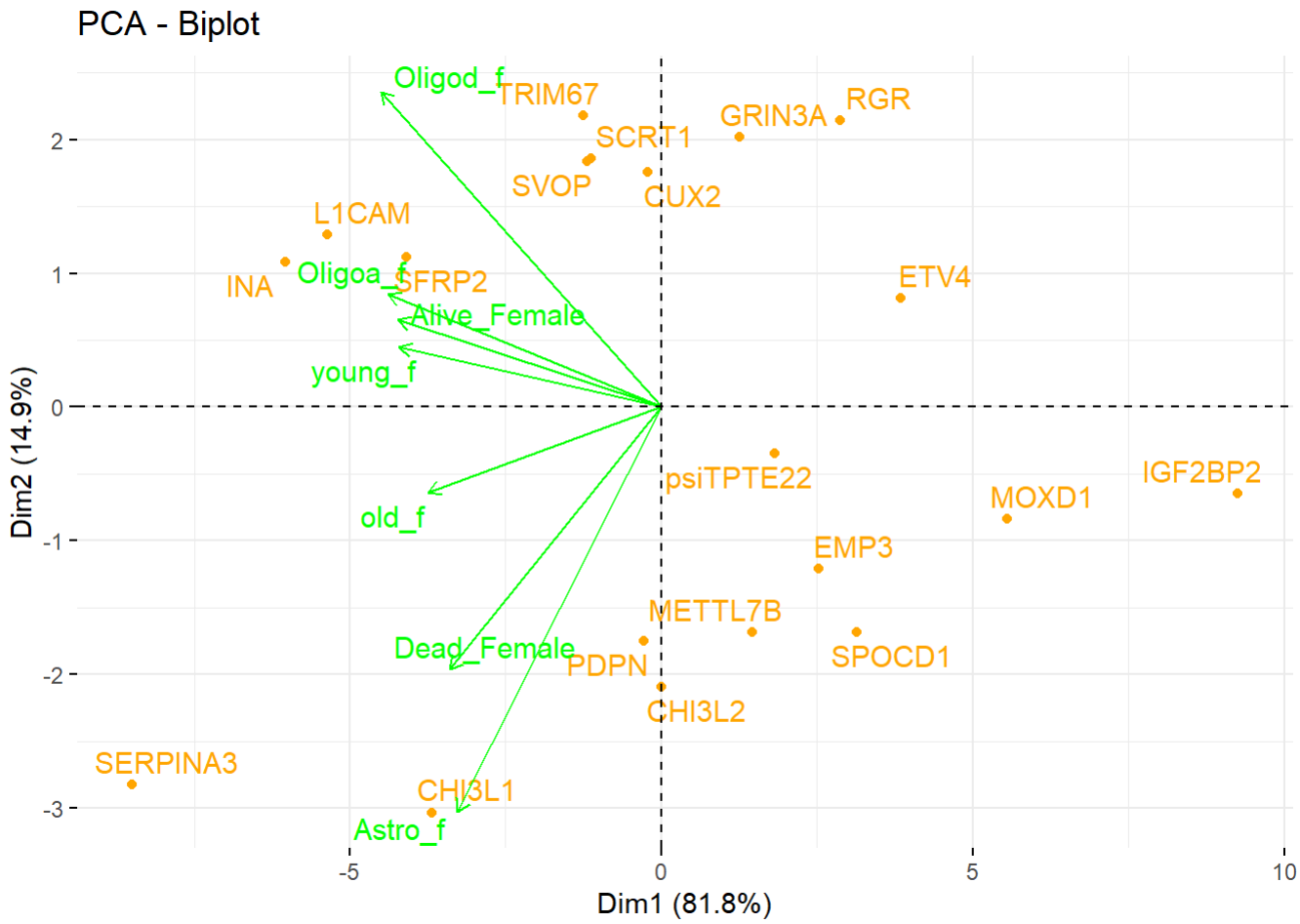
Cluster 3: ETV4

Cluster 4: psiTPTE22, EMP3, MOXD1, METTL7B, SPOCD1, PDPN, CHI3L2, IGF2BP2

Cluster 5: SERPINA3

Cluster 6: CHI3L1

```
#Graph of variables and individuals. Positive correlated variables point to the same side of the
plot. Negative correlated variables point to opposite sides of the graph.
fviz_pca_biplot(pca_female, repel = TRUE,
                col.var = "green", # Variables color
                col.ind = "orange" # Individuals color
)
```



From the PCA biplot above we can see that the eigenvectors for Astro_f and Dead_Female point in the same direction and the eigenvectors for Alive_Female, young_f and Oligoa_f point in same direction. So we can say that the vectors that point in same direction are positively correlated. Since we are looking for a small sets of genes, we test these observations for the bigger set by checking for following hypotheses.

Hypothesis 4:

Hypothesis: Female patients with Astrocytoma histology type tumor are more likely to die.

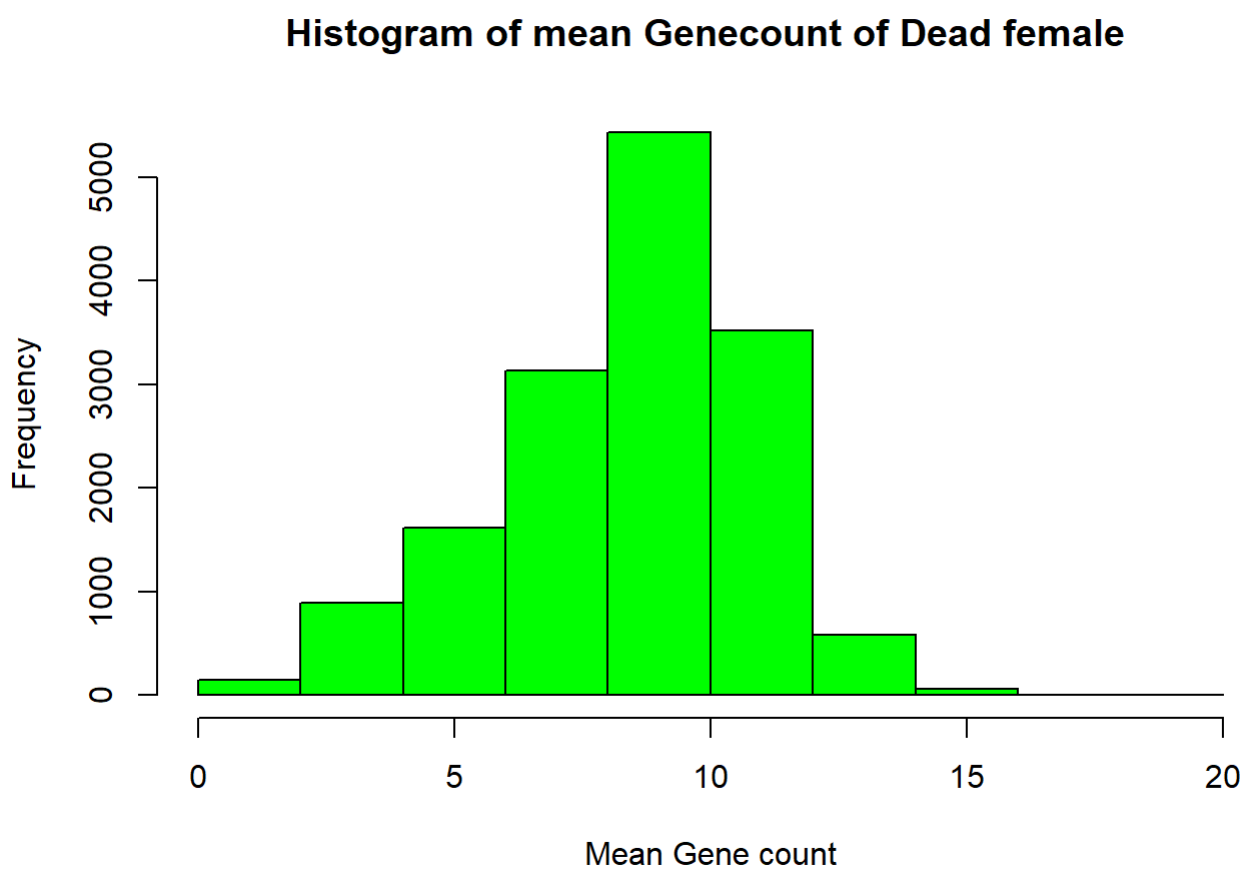
To check this above hypothesis we will do following steps:

- Check if gene count for Dead female and Astrocytoma female is normally distributed or not.
- Using linear regression, visualizing correlation between gene count for Dead female and Astrocytoma female.
- Correlation test between gene count for Dead female and Astrocytoma female.

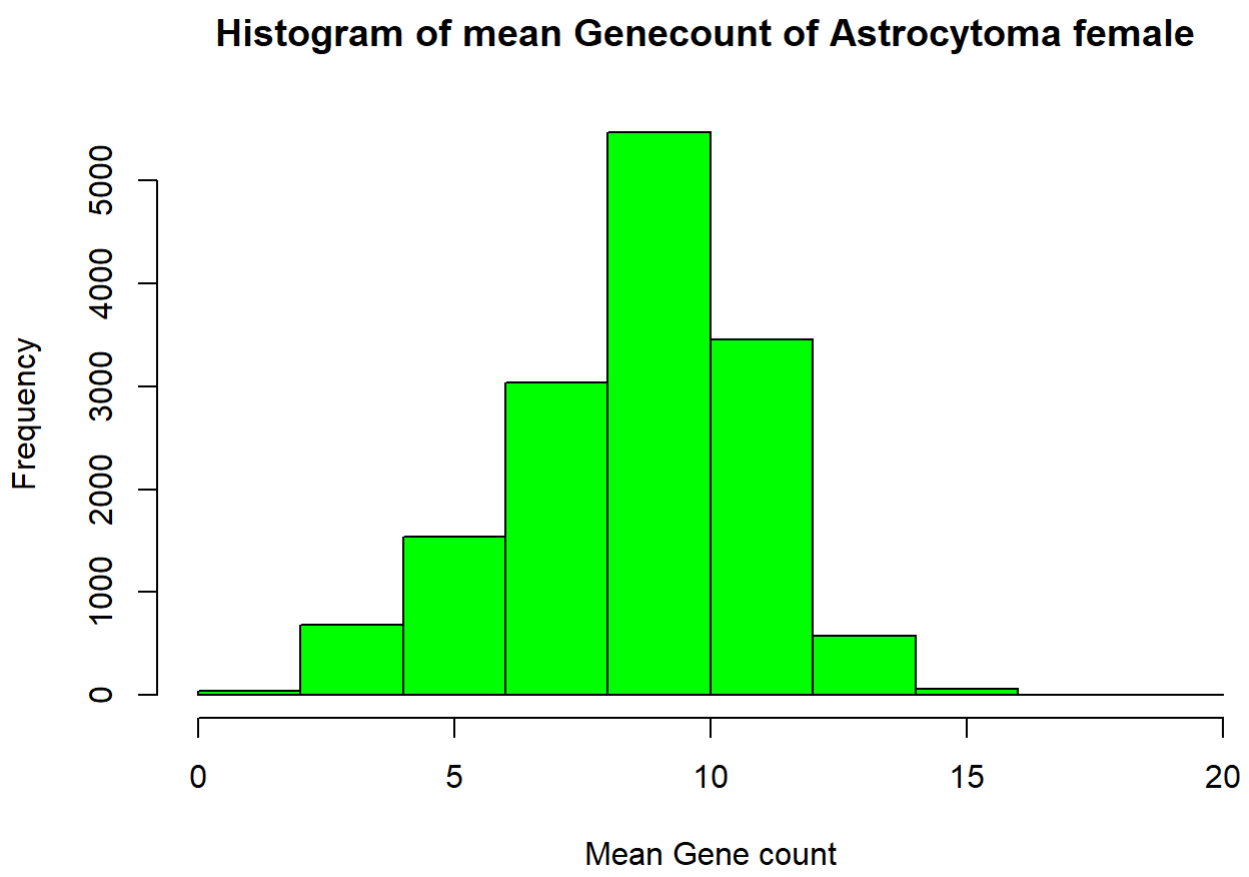
Note that: to perform all the above data analysis we will using original data (ie. my_df) where we haven't replaced NA values with mean and we will be considering gene_count for top 5000 hight variance genes.

Normality Test:

```
# plotting histogram:

hist(my_df$Dead_Female,breaks = 10,col = c("green"),xlab = "Mean Gene count",main = "Histogram o
f mean Genecount of Dead female ")
```




```
hist(my_df$Astro_f,breaks = 10,col = c("green"),xlab = "Mean Gene count",main = "Histogram of me
an Genecount of Astrocytoma female ")
```



From the histograms it seems that both gene count for Dead female and Astrocytoma female are distributed normally. Although we will confirm this result using Shapiro-Wilk normality test. We define Shapiro-Wilk test formally as,

Null Hypothesis: The data sample is Normally ditributed. **Alternate Hypotheis:** The data sample is not Normally ditributed.

```
shapiro.test(new_data_ana$Dead_Female)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_data_ana$Dead_Female
## W = 0.99757, p-value = 3.743e-07
```

```
shapiro.test(new_data_ana$Astro_f)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_data_ana$Astro_f
## W = 0.9965, p-value = 1.786e-09
```

since p-value for data samples it much less than critical value 0.05. Thus we must reject our Null hypothesis and get that gene count for Dead female and Astrocytoma female are **not** distributed normally.

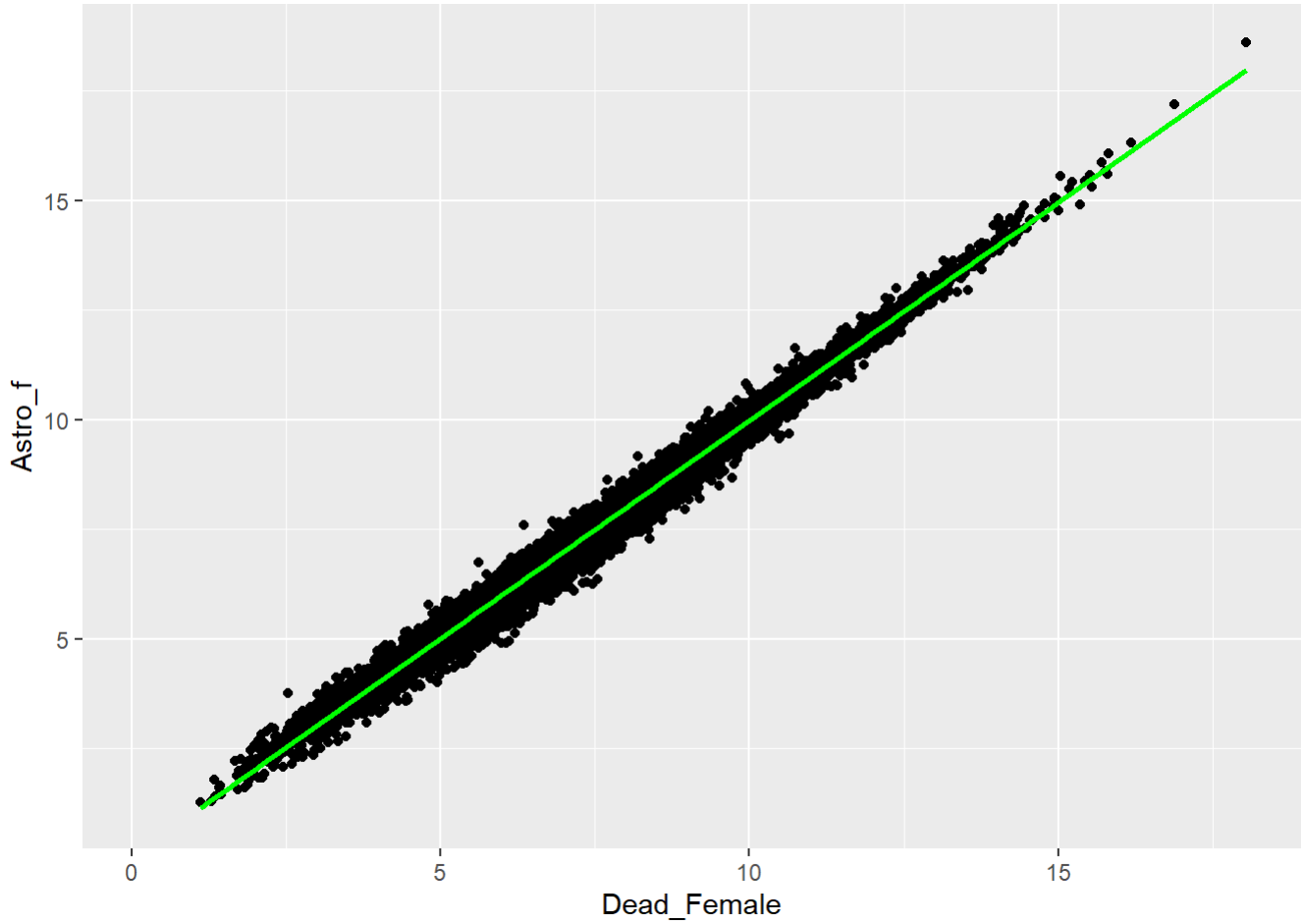
Linear Regression:

```
p3 <- ggplot(my_df, aes(x=Dead_Female, y=Astro_f)) +
  geom_point() +
  geom_smooth(method=lm , color="green", fill="#69b3a2", se=TRUE)

p3
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 3562 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3562 rows containing missing values (geom_point).
```



One can observe that mean gene count in Female patients that die and female patients with Astrocytoma histology type tumor is linearly related. Hence we can say they are positivey correlated. For further analysis we can perform correlation test on these two samples.

Correlation test:

Since the gene count for Dead female and Astrocytoma female are **not** distributed normally we will have to use non parametric correlation test. Here, we will use Spearman test. To perform Spearman Correlation test we will use cor.test() function available in R. Formally we define cor.test as:

Null Hypothesis: The X population vector in not correlated to Y.

Alternate Hypothesis: The X population vector is correlated to Y with the given correlation index.

```
cor.test(my_df$Dead_Female,my_df$Astro_f ,method = 'spearman', use = 'pairwise.complete.obs',exa
ct = F)
```

```
##
##  Spearman's rank correlation rho
##
## data:  my_df$Dead_Female and my_df$Astro_f
## S = 2695875906, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9949595
```

Since the p-value is mush less than critical value 0.05. Thus we reject our Null hypothesis and get that gene count for Dead female and Astrocytoma female is positively correlated with high correlation index.

Conclusion:

Looking at the above correlation test and other ananlysis we observe that mean gene count in Female patients that die and Female patients with Astrocytoma histology type tumor is positively correlated i.e. when a particular gene is expressed more in tumor of dead female, it is highly likely that the same gene is expressed more in tumor cell of female with Astrocytoma histology type tumor. Hence, we caonclude that female with Astrocytoma histology type tumor are more likely to die.

Hypothesis 5:

Hypothesis: Young female patients with Oligoastrocytoma histology type tumor are more likely to live.

To check this above hypothesis we will do following steps:

- Check if mean gene count for Alive female and Young female patients with Oligoastrocytoma histology type tumor is normally distributed or not.
- Using linear regression, visualizing correlation between gene count for Alive female and Young female patients with Oligoastrocytoma histology type tumor.
- Correlation test between gene count for Alive female and Young female patients with Oligoastrocytoma histology type tumor.

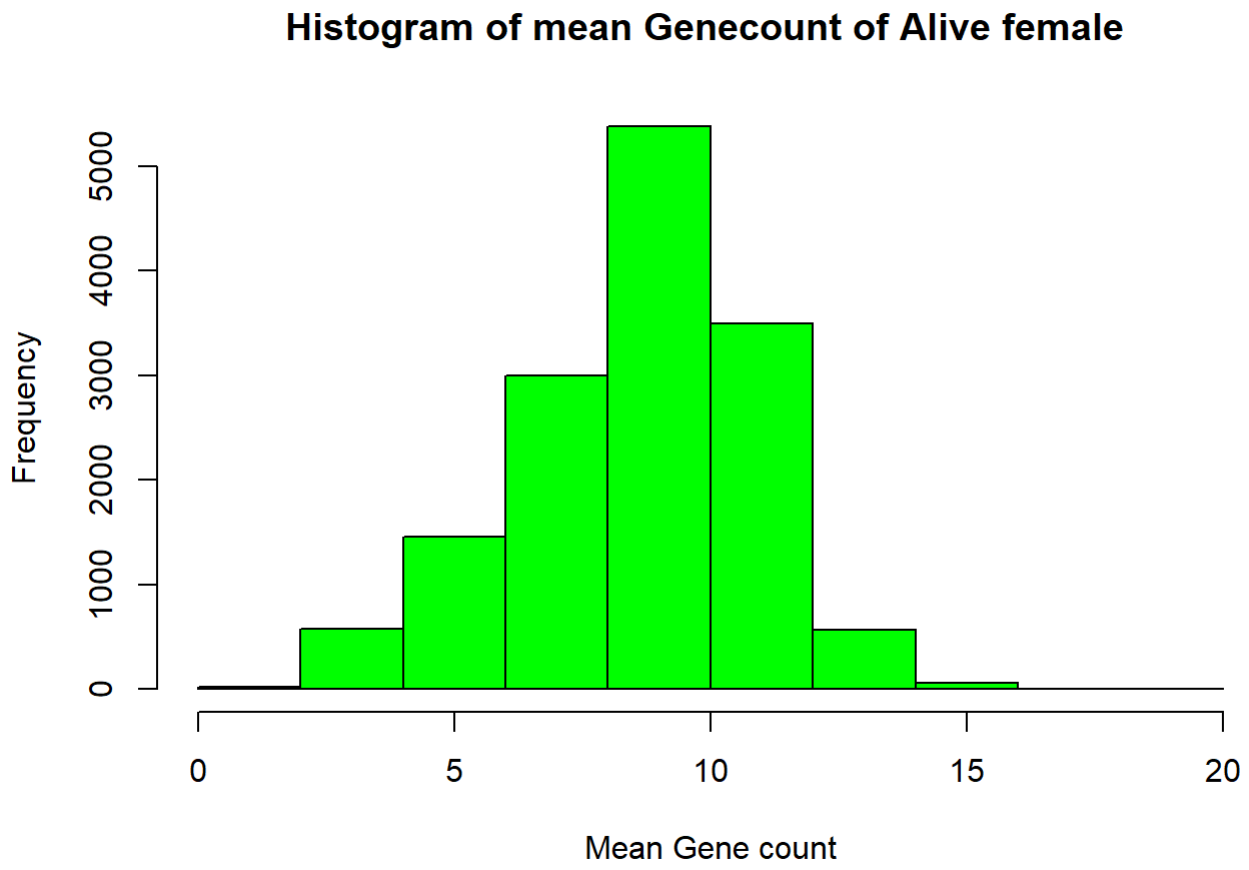
Note that: to perform all the above data analysis we will using original data (ie. my_df) where we haven't replaced NA values with mean and we will be considering gene_count for top 5000 hight variance genes.

```
# getting mean of gene_count for young male patients with Oligoastrocytoma histology type tumor

my_df$Oligod_f_young <- apply(as_tibble(gene_count[,metadata[metadata$gender == "female" & metad
ata$histological_type == "oligodendroglioma" & metadata$age_at_initial_pathologic_diagnosis <= 4
5,1]]),1,mean)
```

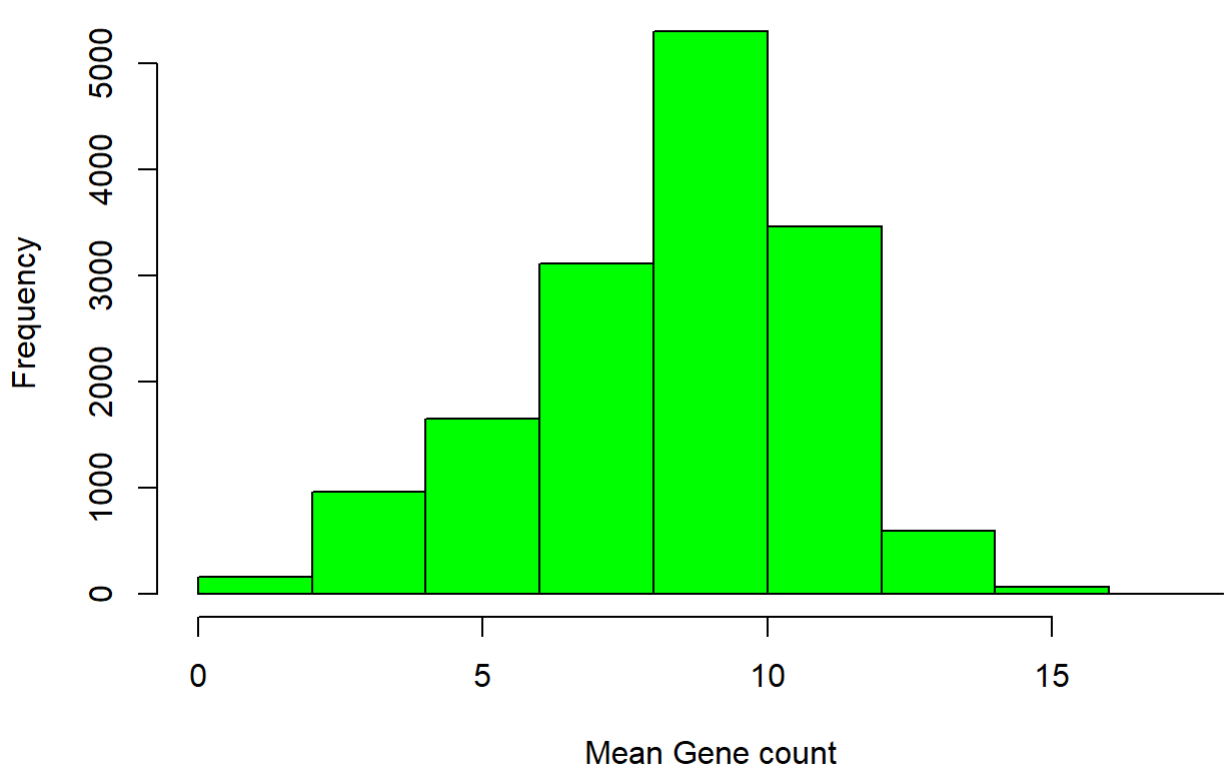
```
new_data_ana2<- my_df[top_5000_genes,]
```

Normality Test:

```
# plotting histogram:

hist(my_df$Alive_Female,breaks = 10,col = c("green"),xlab = "Mean Gene count",main = "Histogram
 of mean Genecount of Alive female ")
```




```
hist(my_df$Oligod_f_young,breaks = 10,col = c("green"),xlab = "Mean Gene count",main = "Histogra
m of mean Genecount of Young Oligoastrocytoma female ")
```



From the histograms it seems that both gene count for Alive female and Young female patients with Oligoastrocytoma histology type tumor aren't distributed normally. Although we will confirm this result using Shapiro-Wilk normality test. We define Shapiro-Wilk test formally as,

Null Hypothesis: The data sample is Normally ditributed. **Alternate Hypotheis:** The data sample is not Normally ditributed.

```
shapiro.test(new_data_ana2$Alive_Female)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_data_ana2$Alive_Female
## W = 0.99577, p-value = 7.724e-11
```

```
shapiro.test(new_data_ana2$Oligod_m_young)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_data_ana2$Oligod_m_young
## W = 0.99625, p-value = 5.824e-10
```

since p-value for data samples it much less than critical value 0.05. Thus we must reject our Null hypothesis and get that gene count for Alive female and Young Oligoastrocytoma female are **not** distributed normally.

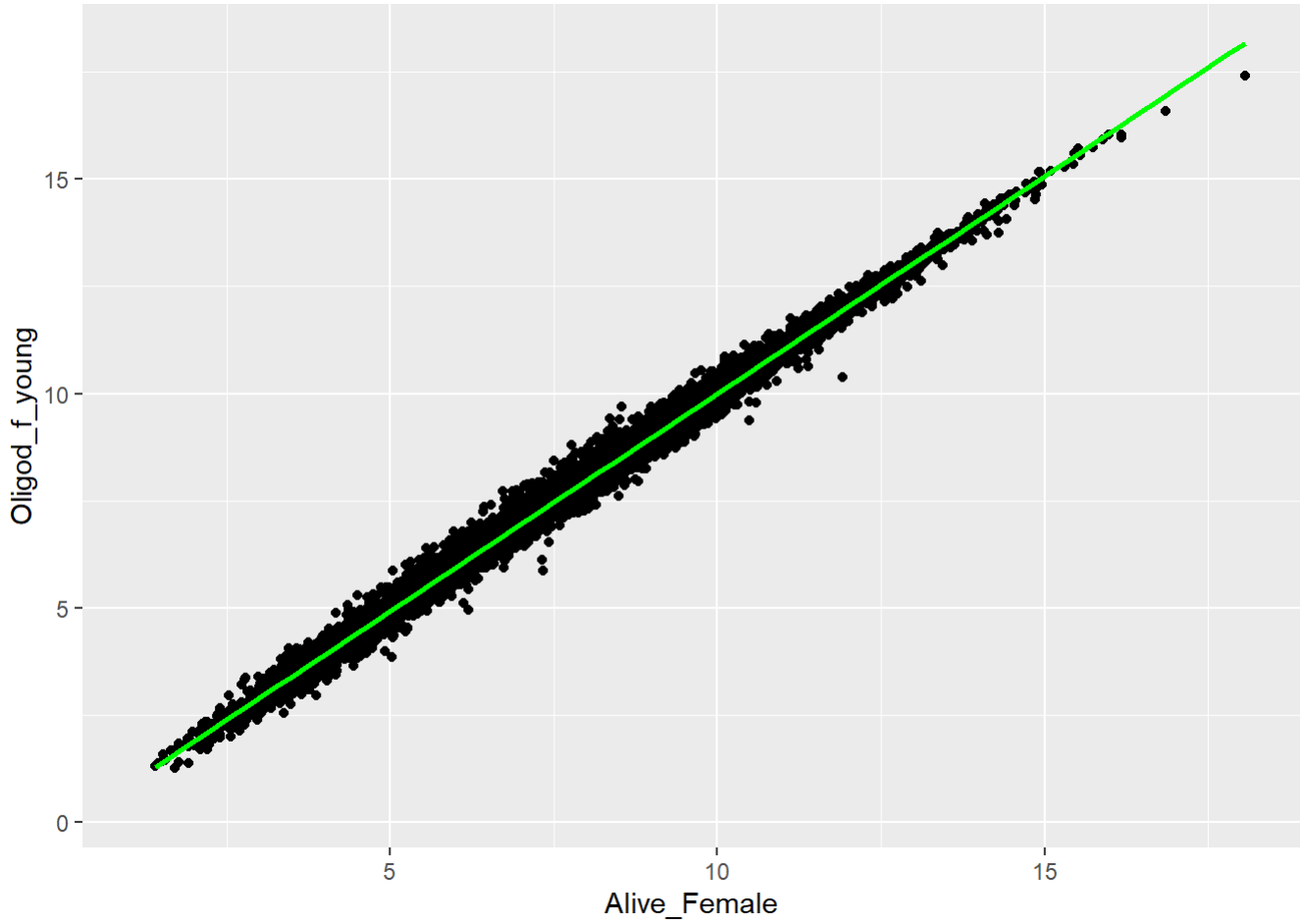
Linear Regression:

```
p4 <- ggplot(my_df, aes(x=Alive_Female, y=Oligod_f_young)) +
  geom_point() +
  geom_smooth(method=lm , color="green", fill="#69b3a2", se=TRUE)

p4
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 3789 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3789 rows containing missing values (geom_point).
```



One can observe that mean gene count in Female patients that are alive and Young Female patients with Oligoastrocytoma histology type tumor is linearly related. Hence we can say they are positively correlated. For further analysis we can perform correlation test on these two samples.

Correlation test:

Since both the samples in consideration are **not** distributed normally we will have to use non parametric correlation test. Here, we will use Spearman test. To perform Spearman Correlation test we will use cor.test() function available in R. Formally we define cor.test as:

Null Hypothesis: The X population vector in not correlated to Y.

Alternate Hypothesis: The X population vector is correlated to Y with the given correlation index.

```
cor.test(my_df$Alive_Female,my_df$Oligod_f_young ,method = 'spearman', use = 'pairwise.complete.
obs',exact = F)
```

```
##
##  Spearman's rank correlation rho
##
## data:  my_df$Alive_Female and my_df$Oligod_f_young
## S = 1837073909, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9964016
```

Since the p-value is mush less than critical value 0.05. Thus we reject our Null hypothesis and get that gene count for Alive female and Young Oligoastrocytoma female is positively correlated with high correlation index.

Conclusion:

Looking at the above correlation test and other analysis we observe that mean gene count in Female patients that are alive and Young female patients with Oligoastrocytoma histology type tumor is positively correlated i.e. when a particular gene is expressed more in tumor of alive female, it is highly likely that the same gene is expressed more in tumor cell of young female with Oligoastrocytoma histology type tumor. Hence, we conclude that young female with Oligoastrocytoma histology type tumor are more likely to stay alive.