TCGA Gene Expression

MOHANKumar

2024-01-08

Contents

Overview

This R markdown file is created for the purpose of this challenge to complete a full deferential gene expression analysis. The Tumor RNA gene expression raw counts data has been download from the GDC data portal. This workflow id for the Breast cancer gene expression. Th control tissue gene counts has been downloaded from the GTEX portal. After Downloading the data, the data preprocess steps has been done and the normalization has applied. Here, DESeq has performed to find the differential expressed genes and how they are correlated and the vaiable genes which are expressed in the volcono plot. Then the significant genes has been taken to find the gene enrichment analysis.

Importing the Packages

```
library(tidyverse)
library(SummarizedExperiment)
library(maftools)
library(pheatmap)
library(TCGAbiolinks)
library(gage)
library(biomaRt)
library(DESeq2)
library(dplyr)
library(EnhancedVolcano)
library(org.Hs.eg.db)
library(ggplot2)
library(clusterProfiler)
library(enrichplot)
library(DOSE)
library(ggridges)
library(pathview)
library(reshape)
library(RColorBrewer)
```

```
#setting the working directoiry

setwd("D:/Finished projects/TCGA_project")
```

Data Dowlaoding using TCGAbiolinks

Data Downloading from TCGA database for breasr cancer using TCGAbiolinks. The dataset is download from GDCdata

```
# get a list of projects
gdcprojects <- getGDCprojects()

getProjectSummary('TCGA-BRCA')


# building a query
query_TCGA <- GDCquery(project = 'TCGA-BRCA',
                       data.category = 'Transcriptome Profiling')
output_query_TCGA <- getResults(query_TCGA)


# build a query to retrieve gene expression data
query_TCGA <- GDCquery(project = "TCGA-BRCA",
                       data.category = 'Transcriptome Profiling',
                       experimental.strategy = 'RNA-Seq',
                       workflow.type = 'STAR - Counts',
                       access = 'open',
                       barcode=c('TCGA-LL-A73Y-01A-11R-A33J-07','TCGA-A2-A04R-01A-41R-A109-07','TCGA-AN
```

```
getResults(query_TCGA)

#dowload data

GDCdownload(query_TCGA)
```

```
# prepare data
tcga_brca_data <- GDCprepare(query_TCGA, summarizedExperiment = TRUE)

brca_matrix <- assay(tcga_brca_data)

brca_dataframe <- as.data.frame(brca_matrix)

colnames(brca_dataframe )
colnames(brca_dataframe)[1] = "A33J"
colnames(brca_dataframe)[2] = "A109"
colnames(brca_dataframe)[3] = "A034"
colnames(brca_dataframe)[4] = "A266"
colnames(brca_dataframe )
```

Getting the genename and gene id

Get the gene name form the ensemble version id

```
listEnsembl()
ensembl <- useEnsembl(biomart = "genes")
datasets <- listDatasets(ensembl)

ensembl.con <- useMart("ensembl", dataset = 'hsapiens_gene_ensembl')

attr <- listAttributes(ensembl.con)
filters <- listFilters(ensembl.con)

TCGA_gene<- getBM(attributes = c('ensembl_gene_id_version','ensembl_gene_id','external_gene_name'),
            filters = "ensembl_gene_id_version",
            values = row.names(brca_dataframe) ,
            mart = ensembl.con)

row.names(TCGA_gene) <- TCGA_gene$ensembl_gene_id_version

TCGA_df <- merge(brca_dataframe, TCGA_gene, by = 0, all = TRUE)

row.names(TCGA_df) = TCGA_df$Row.names
```

```
## [1] "TCGA data:"
```

```
##                            Row.names A33J A109 A034 A266
## ENSG00000000003.15 ENSG00000000003.15 7015 1557 6065 1982
## ENSG00000000005.6   ENSG00000000005.6   16    8    1    2
```

```
## ENSG00000000419.13 ENSG00000000419.13 2167 1825 4863 2444
## ENSG00000000457.14 ENSG00000000457.14 2505 1716 3842  761
## ENSG00000000460.17 ENSG00000000460.17  726 1013 2364  845
## ENSG00000000938.13 ENSG00000000938.13 1404  154  158  529
##                    ensembl_gene_id_version ensembl_gene_id external_gene_name
## ENSG00000000003.15                    <NA>            <NA>               <NA>
## ENSG00000000005.6        ENSG00000000005.6 ENSG00000000005               TNMD
## ENSG00000000419.13                    <NA>            <NA>               <NA>
## ENSG00000000457.14       ENSG00000000457.14 ENSG00000000457              SCYL3
## ENSG00000000460.17       ENSG00000000460.17 ENSG00000000460           C1orf112
## ENSG00000000938.13       ENSG00000000938.13 ENSG00000000938                FGR
```

Getting raw data counts from the reference tissue file

The refrence tissue data set is download form the GTEX Portal

```
gct_file_path <- "D:\\Finished projects\\TCGA_project\\GTEX\\gene_reads_2017-06-05_v8_breast_mammary_ti


dat.gct <- read.delim(file=gct_file_path, skip=2)

GTEX.df <- as.data.frame(dat.gct)


refrence_sample <- data.frame(row.names = GTEX.df$Name,GTEX.df$Description,GTEX.df$GTEX.1117F.2826.SM.5
colnames(refrence_sample)

colnames(refrence_sample)[1]= "gene_name"
colnames(refrence_sample)[2]= "5GZXL"
colnames(refrence_sample)[3]= "5GICC"
colnames(refrence_sample)[4]= "5H113"
colnames(refrence_sample)[5]= "5987X"
colnames(refrence_sample)
```

```
## [1] "refrence_sample:"

##                     gene_name 5GZXL 5GICC 5H113 5987X
## ENSG00000223972.5     DDX11L1     0     1     0     0
## ENSG00000227232.5      WASH7P   286   135   110   192
## ENSG00000278267.1   MIR6859-1     0     0     0     0
## ENSG00000243485.5 MIR1302-2HG     0     0     0     1
## ENSG00000237613.2     FAM138A     0     0     0     0
## ENSG00000268020.3      OR4G4P     1     3     0     0
```

```
# merging the two dataframe
merged_df1 <-  merge(brca_dataframe,refrence_sample, by = 0, all = FALSE)
row.names(merged_df1) = merged_df1$Row.names


selected_columns <- c("A034", "A109", "A33J","A266","5GZXL","5GICC","5H113","5987X")

new_df <- merged_df1[selected_columns]
```

```
## [1] "Raw data:"
```

```
##                     A034 A109 A33J A266 5GZXL 5GICC 5H113 5987X
## ENSG00000001167.14 8568 5088 6487 1665  2384  2894  1465  2876
## ENSG00000002549.12 2552 4769 6106 7290  3913  3966  5270  4787
## ENSG00000002822.15   14    6   31   18  1012  1145  1320  1401
## ENSG00000003096.14   66  344   92  142   299   479   729   633
## ENSG00000003137.8   113  101 9183  476  6696   558  1952  1609
## ENSG00000004777.18 3548  885  717 1006  1869  4623  1940  4095
```

Deseq Normalization

```
#removing the rows having sum of counts 0
new_df <- new_df[which(rowSums(new_df)>0),]

#meta data
col_data  <- data.frame(condition = c("Disease","Disease","Disease","Disease","Control","Control","Cont

row.names(col_data) <- selected_columns


dds <- DESeqDataSetFromMatrix(countData = new_df, colData = col_data , design = ~ condition)

filter_counts <- rowSums(counts(dds)) >= 50
dds <- dds[filter_counts,]

dds$condition <- relevel(dds$condition, ref = "Control")
dds <- DESeq(dds)
res <- results(dds)

#summary(res)
#colnames(dds)


normalized_counts <- as.data.frame(counts(dds, normalized=TRUE))

significat_gene <- data.frame(res)

gene_df <- merged_df1[c("gene_name")]
```

```
## [1] "Meta data :"
```

```
##       condition   sample_type
## A034    Disease Primary_tumor
## A109    Disease Primary_tumor
## A33J    Disease Primary_tumor
## A266    Disease Primary_tumor
## 5GZXL   Control        Normal
## 5GICC   Control        Normal
```

Data Analysis

```
#setting the filter for significant gene
padj.cutoff <- 0.05
lfc.cutoff <- 0.58

threshold <- significat_gene$padj < padj.cutoff & abs(significat_gene$log2FoldChange) > lfc.cutoff
length(which(threshold))
significat_gene$threshold <- threshold


significat_gene <-  merge(significat_gene,gene_df, by = 0, all = FALSE)
rownames(significat_gene) <- significat_gene$Row.names

sigOE <- data.frame(subset(significat_gene, threshold==TRUE))



## Order significant results by padj values

sigOE_ordered <- significat_gene[order(significat_gene$padj), ]
top20_sigOE_genes <- rownames(sigOE_ordered[1:20, ])



## normalized counts for top 20 significant genes
normalized_counts <- counts(dds, normalized=T)
top20_sigOE_norm <- normalized_counts[top20_sigOE_genes, ]
top20_sigOE_norm <- merge(top20_sigOE_norm,gene_df, by = 0, all = FALSE)
#colnames(top20_sigOE_norm )

select_col <- c("gene_name","A034", "A109","A33J","A266","5GZXL","5GICC","5H113","5987X")
top20_sigOE_norm <- top20_sigOE_norm[select_col]


## use melt to modify the format of the data frame
melted_top20_sigOE <- data.frame(melt(top20_sigOE_norm))

## check the column header in the "melted" data frame


colnames(melted_top20_sigOE) <- c("gene_name", "samplename", "normalized_counts")

melted_top20_sigOE<- merge(melted_top20_sigOE, col_data, by.x = "samplename", by.y = 0, all.x = FALSE)
```

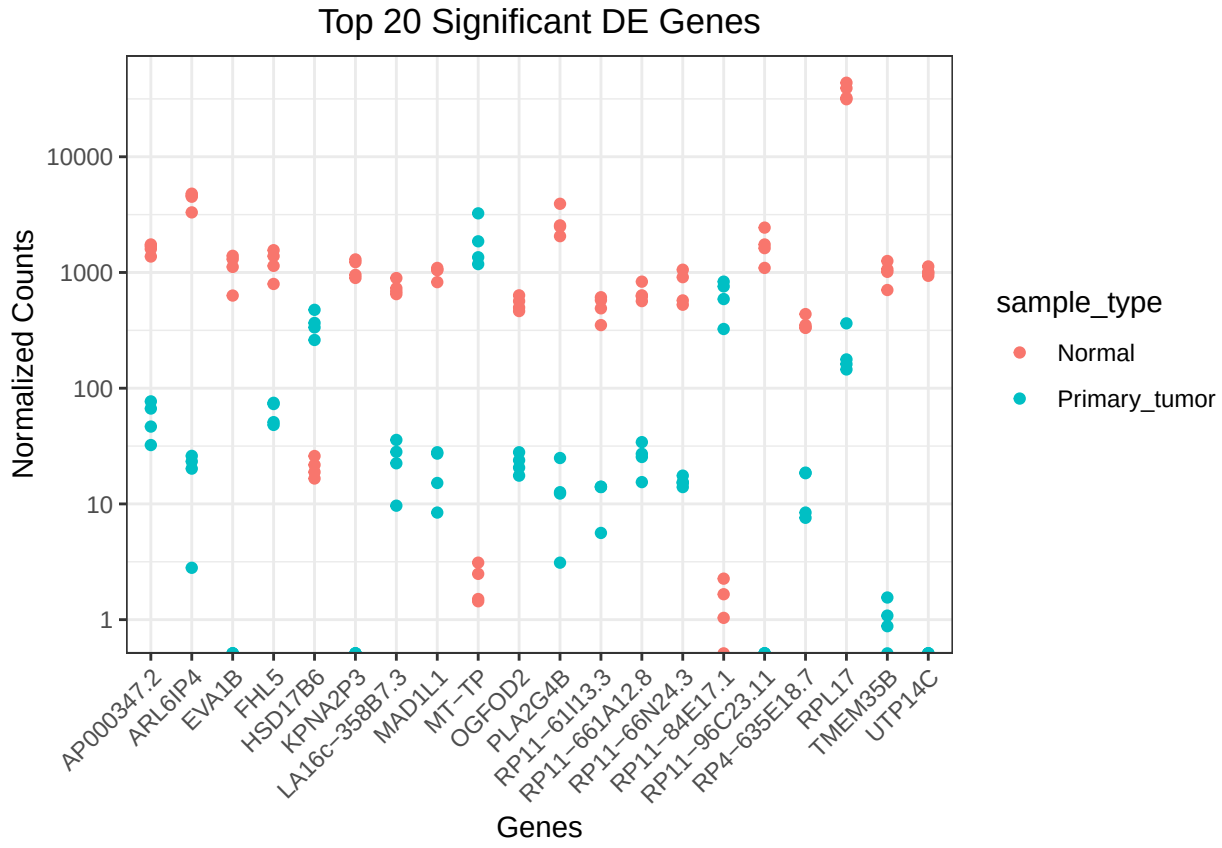
Plot for Significant DE Genes

This Plot shows the Top 20 significant genes which has been differentially expressed

```
## plot using ggplot2
sig_gene_plot<-ggplot(melted_top20_sigOE)+
  geom_point(aes(x = gene_name, y = normalized_counts, color = sample_type)) +
```

```
  scale_y_log10() +
  xlab("Genes") +
  ylab("Normalized Counts") +
  ggtitle("Top 20 Significant DE Genes") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5))

print(sig_gene_plot)
```



HeatMap

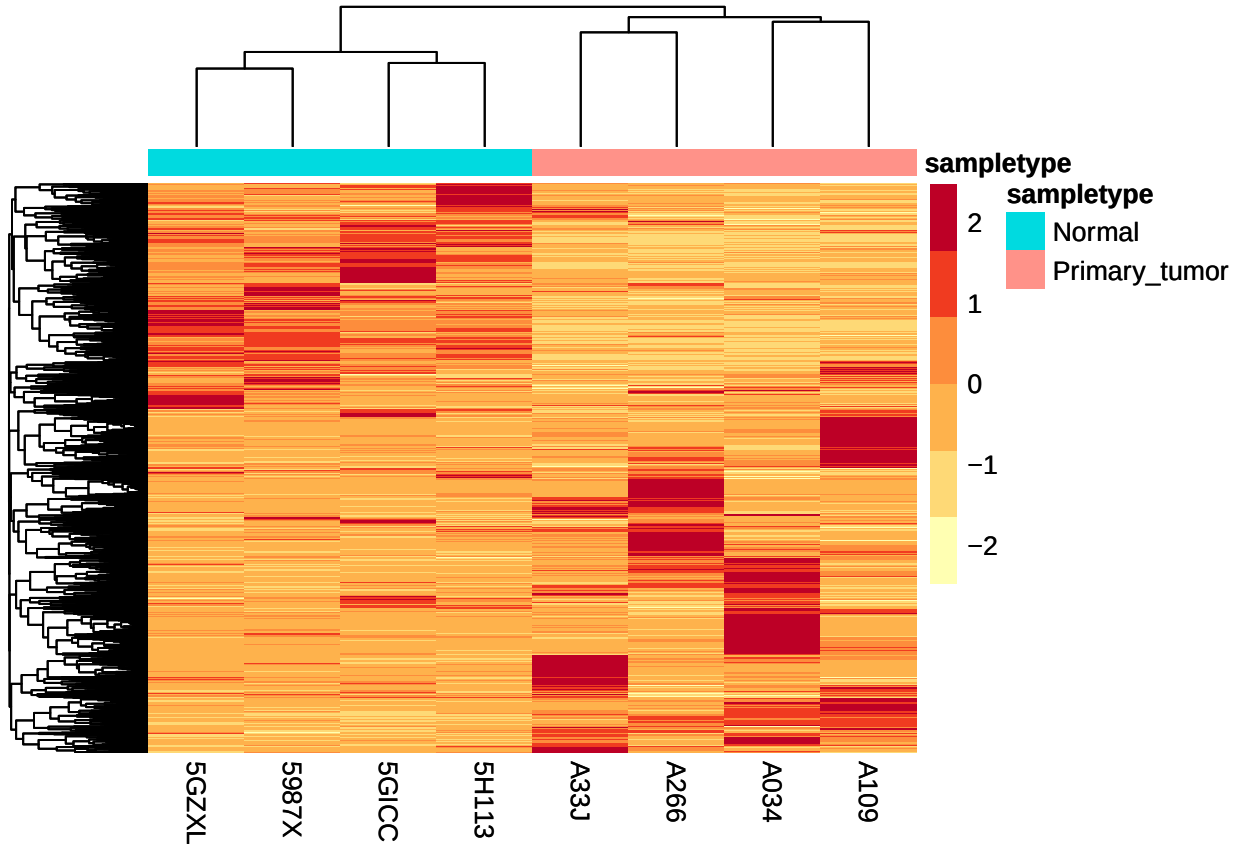
From the Heatmap, we correlate the expressing of genes in the control and Tumor condition. By the plot, under Tumor condition more number of gene has high correlated value. The cluster nodes gives much information how they correlated

```
# Heatmap

norm_OEsig <- normalized_counts[rownames(normalized_counts),]
### Set a color palette
heat.colors <- brewer.pal(6, "YlOrRd")
annotation <- data.frame(sampletype=col_data[,'sample_type'],
                         row.names=rownames(col_data))
```

```
heat_map <-pheatmap(norm_OEsig, color = heat.colors, cluster_rows = T, show_rownames=F,annotation= anno
         fontsize_row = 10, height=20)

print(heat_map )
```



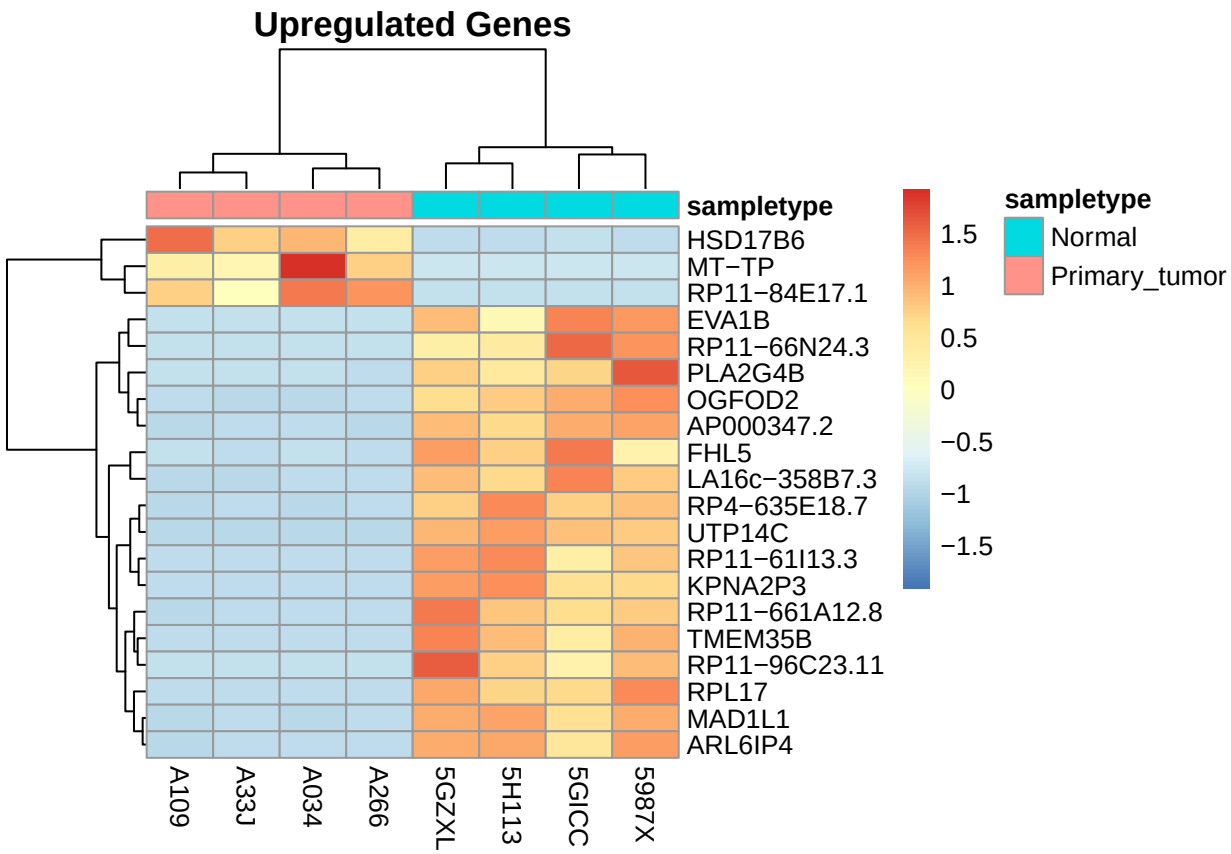
From the below graph, the genes which can be expressed in specific condition and how the genes can be correlated under each condition for each sample has been identified.

```
up <-  top20_sigOE_norm
row.names(up) <- up$gene_name
colnames(up)
```

```
## [1] "gene_name" "A034"      "A109"      "A33J"      "A266"      "5GZXL"
## [7] "5GICC"     "5H113"     "5987X"
```

```
up_df = up[c("A034","A109","A33J","A266","5GZXL","5GICC","5H113","5987X")]


pheatmap(up_df,scale = 'row',main = "Upregulated Genes",cluster_rows = T,cluster_cols = T ,annotation= a
```



PCA plot

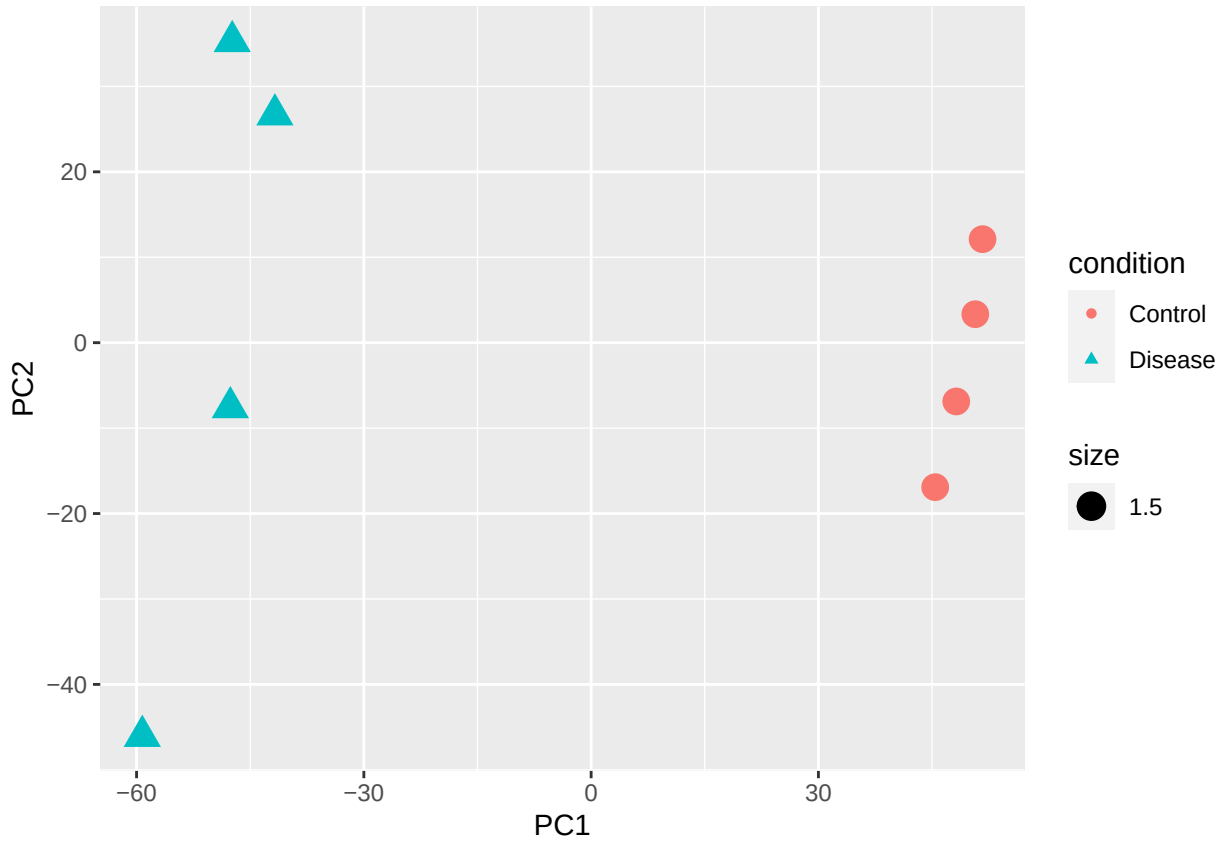
There are two functions within DEseq2 to transform the data in such a manner, the first is to use a regularized logarithm **rlog()** and the second is the variance stablizing transform **vst()**. There are pros and cons to each method, we will use vst() h. It help to cluster the samples.

```
dds_norm <- vst(dds,blind = FALSE)
dds_norm


pca_results <-  plotPCA(dds_norm,intgroup = c("condition"),returnData = TRUE) # This argument tells R t

annotated_pca_plot<-ggplot( pca_results,
  aes(x = PC1,y = PC2,shape=condition,color= condition,   size = 1.5)) +geom_point()

print(annotated_pca_plot)
```

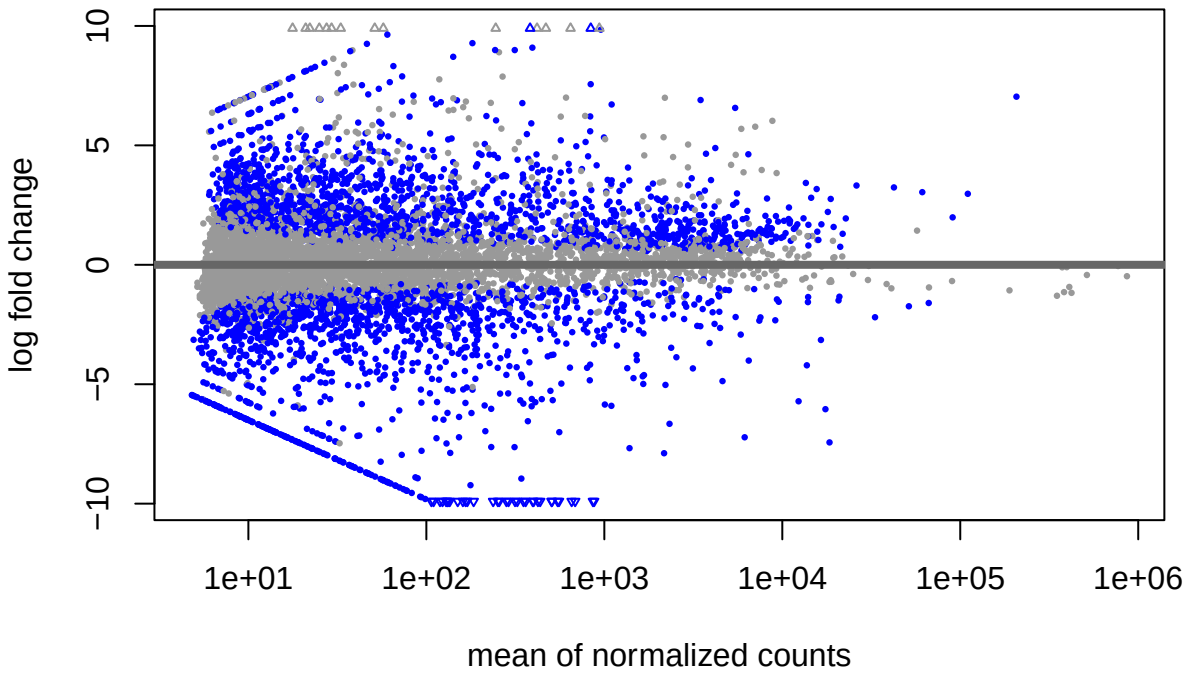


MA plot

MA plots display a log ratio (M) vs an average (A) in order to visualize the differences between two groups.The expression of genes to remain consistent between conditions and so the MA plot should be similar to the shape of a trumpet with most points residing on a y intercept of 0. The blue color dots indicate the gene that are differentialy expressed and the Triangle sign indicate the genes has higher fold changes.

```
 # MA plot

MA <-plotMA(res)
```




```
print(MA)
```

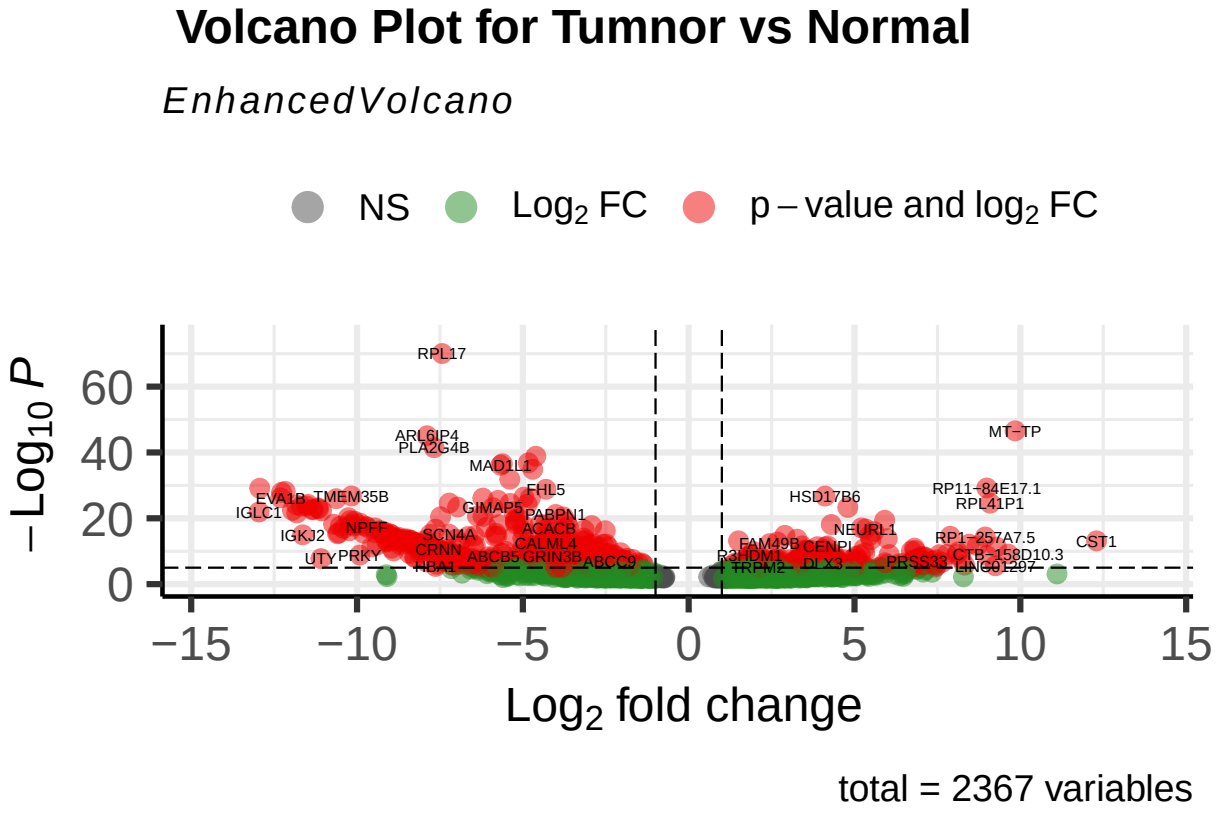
```
## NULL
```

Volcono plot

Volcano plots provide an effective means for visualizing the direction, magnitude, and significance of changes in gene expression.The log2-fold differences between the groups are plotted on the x-axis and the -log10 p-value differences are plotted on the y-axis. The horizontal dashed line represents the significance threshold specified in the analysis, usually derived using a multiple testing correction.

Genes whose expression is decreased versus the comparison group are located to the left of zero on the x-axis while genes whose expression is increased are illustrated to the right of zero. Genes with statistically significant differential expression lie above a horizontal threshold. Closer to zero indicates less change while moving away from zero in either direction indicates more change

```
significat_gene <- data.frame(res)
significat_gene <-  merge(significat_gene,gene_df, by = 0, all = FALSE)
sigOE <- data.frame(subset(significat_gene, threshold==TRUE))
sigOE <- data.frame(subset(significat_gene, threshold==TRUE))
EnhancedVolcano(sigOE,
                lab = sigOE$gene_name,
                x = 'log2FoldChange',
                y = 'pvalue',title = ' Volcano Plot for Tumnor vs Normal',labSize = 2, pointSize = 3.0)
```



Gene ENrichemnt Analysis

The enrichment analysis helps to get information of the genes which involves in biological,molecular, and cellualar process. This helps to see the which are the pathway are getting affected

```
EA <- data.frame(log2FoldChange = sigOE$log2FoldChange,gene_name =  sigOE$gene_name,ensemble_version= si


a_gene <- c("log2FoldChange" = EA[, 1])
names(a_gene) <- EA[,2]
a_gene <-sort(a_gene,decreasing = TRUE)
library(org.Hs.eg.db)
```

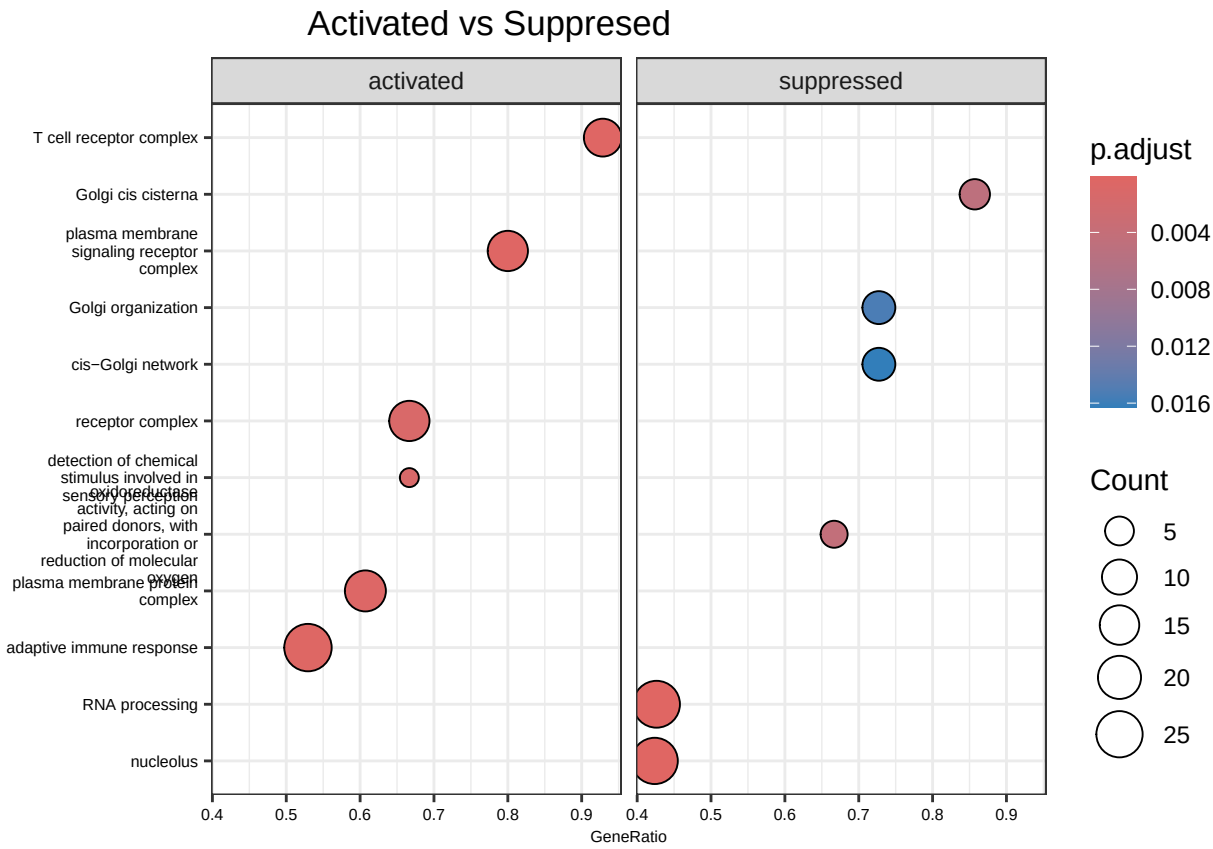
```
gse <- gseGO(geneList=a_gene,
             ont = "ALL",
             keyType =  "SYMBOL",
             nPerm = 10000,
             minGSSize = 3,
             maxGSSize = 1000,
             pvalueCutoff = 0.05,
             verbose = TRUE,
             OrgDb = org.Hs.eg.db,
             pAdjustMethod = "none")
```

```
##            ONTOLOGY         ID                                Description
## GO:0006396       BP GO:0006396                               RNA processing
## GO:0005730       CC GO:0005730                                    nucleolus
## GO:0042101       CC GO:0042101                      T cell receptor complex
## GO:0098802       CC GO:0098802 plasma membrane signaling receptor complex
## GO:0002250       BP GO:0002250                     adaptive immune response
## GO:0098797       CC GO:0098797            plasma membrane protein complex
##            setSize enrichmentScore       NES       pvalue     p.adjust
## GO:0006396      61      -0.5653102 -2.486991 0.0001294498 0.0001294498
## GO:0005730      59      -0.5541270 -2.418838 0.0001301744 0.0001301744
## GO:0042101      14       0.7531966  2.633604 0.0002655337 0.0002655337
## GO:0098802      20       0.6001848  2.344501 0.0002951594 0.0002951594
## GO:0002250      51       0.3980755  2.061087 0.0004040404 0.0004040404
## GO:0098797      28       0.4804696  2.106487 0.0006451613 0.0006451613
##                qvalue rank                    leading_edge
## GO:0006396 0.1690966  227 tags=41%, list=10%, signal=38%
## GO:0005730 0.1690966  227 tags=41%, list=10%, signal=38%
## GO:0042101 0.1917060  491 tags=93%, list=21%, signal=74%
## GO:0098802 0.1917060  663 tags=80%, list=28%, signal=58%
## GO:0002250 0.2099394  491 tags=53%, list=21%, signal=43%
## GO:0098797 0.2793548  531 tags=61%, list=22%, signal=48%
##
## GO:0006396        SNORA1/UMOD/SNORD21/SCARNA13/SNORD35B/SNORA27/SNORA21/SNORA44/SNORA52/SNORA62/SNORA
## GO:0005730             SNORA1/SNORD21/SCARNA13/SNORD35B/SNORA27/SNORA21/SNORA44/SNORA52/SNORA62/SNORA
## GO:0042101
## GO:0098802
## GO:0002250 IGHG1/IGHV1-69/CXCL13/TRBV5-6/TRAV17/TRAV19/IGHV5-51/IGKV1D-13/TRAV4/IGHV1-69D/FCGR1A/TRBV
## GO:0098797                                                                                          TR
```

Dot Plot

Dot-plot representation of the gene expression marker genes for the identified cell types. The size of dots represents the relative gene expression in percent for each cluster

```
dotplot(gse,showCategory=6, title = "          Activated vs Suppresed ",
        font.size = 6,
        label_format = 25,
        split=".sign")+facet_grid(.~.sign)
```


Activated vs Suppresed
activated
suppressed
T cell receptor complex
Golgi cis cisterna
plasma membrane signaling receptor complex
Golgi organization
cis−Golgi network
receptor complex
detection of chemical stimulus involved in sensory perception
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
plasma membrane protein complex
adaptive immune response
RNA processing
nucleolus
0.4
0.5
0.6
0.7
0.8
0.9
GeneRatio
p.adjust
0.004
0.008
0.012
0.016
Count
5
10
15
20
25