Final_Project

Title

Study the gene expression pattern in TCGA of different age men with prostate cancer using DeSEQ2

Author

Cheng-Hsiang Lu

Overview of project

I will study differentially expressed genes between two groups. One group of people is younger than 65 years old, while the other group is older than 65 years old. This analysis will utilize the package DESeq2 and follow the specific vignette: link

For this analysis, I will use the TCGA cohort and have identidfied 331 RNA-seq counts files for tumors that fit within my cohort.Saperated by the age of 65 years old, 128 samples are from the group beyond 65 years old and 203 samples are from the group under 65 years old. Within the analysis, I will control for race and primary gleason grade.

Data

I will use the data from GDC Examining clinical data, there are total 331 cases from 55 to 75 years old, and each group has 128 and 203 samples .

Milestone_1

Data filtering

First, go to GDC and click on "**Repository**".

On the left side "**Files**" filters:

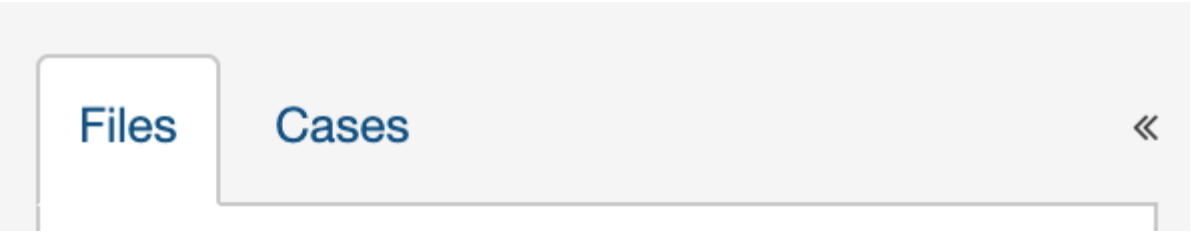
Data Category - "**transcriptome profiling**".

Data Type - "**Gene Expression Quantification**".

Experimental Strategy - "**RNA-Seq**".

Workflow Type - "**HTSeq - Counts**".

Access - "**open**".



Add a File Filter

Search Files

e.g. 142682.bam, 4f6e2e7a-b...

Data Category

	# Files
[x] transcriptome profiling	17,753

Data Type

	# Files
[x] Gene Expression Quantification	17,753

Experimental Strategy

	# Files
[x] RNA-Seq	17,753

Workflow Type

	# Files
[x] HTSeq - Counts	17,753
[ ] HTSeq - FPKM	17,753
[ ] HTSeq - FPKM-UQ	17,753
[ ] STAR - Counts	6,630

Data Format

	# Files
[ ] txt	17,753

Platform

On the left side "**Cases**" filters:

First, click "**Add a Case/Biospecimen Filter**"

Then, type "**Gleason Grade**" and select "**primary_gleason_grade**".

Diagnoses Primary Gleason Grade - "**pattern 3**" and "**pattern 4**".

Primary Site - "**prostate gland**".

Program - "**TCGA**".

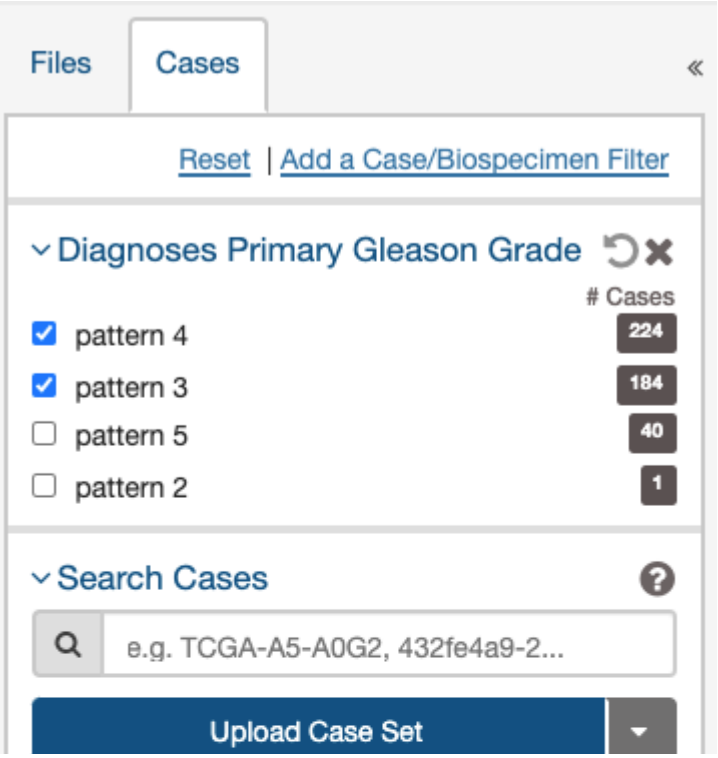
Disease Type - "**adenomas and adenocarcinomas**".

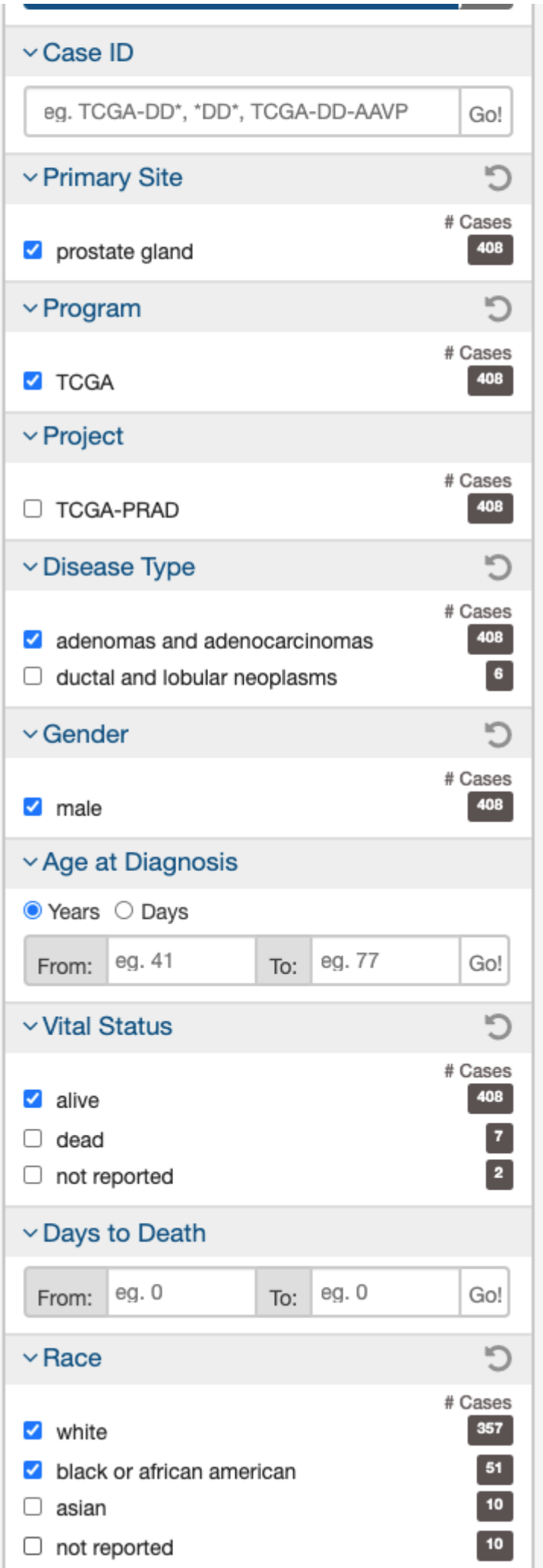
Gender - "**male**".

Age at Diagnosis - From "**55**" to "**64**".

Vital Status - "**alive**".

Race - "**white**" and "**black or african american**".



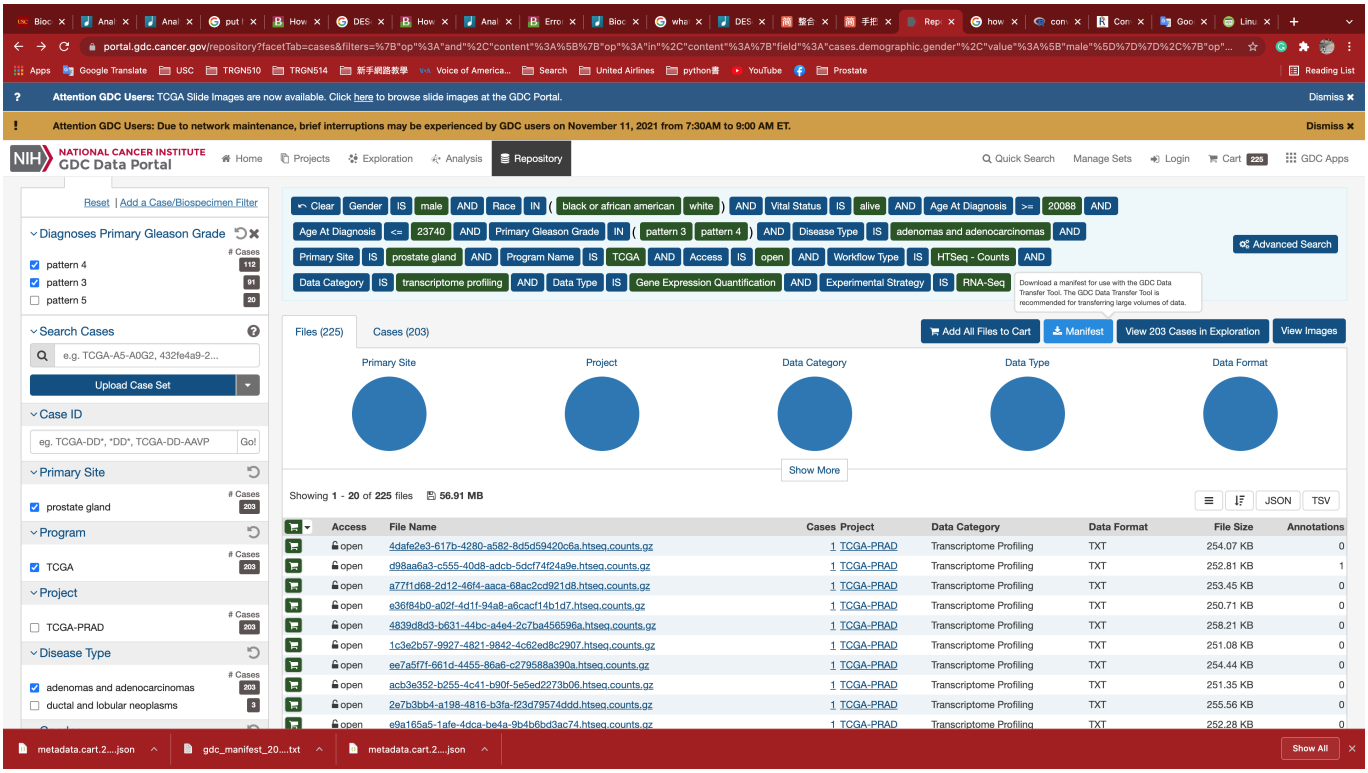

Case ID
eg. TCGA-DD*, *DD*, TCGA-DD-AAVP
Go!
Primary Site
Cases
prostate gland
408
Program
Cases
TCGA
408
Project
Cases
TCGA-PRAD
408
Disease Type
Cases
adenomas and adenocarcinomas
408
ductal and lobular neoplasms
6
Gender
Cases
male
408
Age at Diagnosis
Years
Days
From:
eg. 41
To:
eg. 77
Go!
Vital Status
Cases
alive
408
dead
7
not reported
2
Days to Death
From:
eg. 0
To:
eg. 0
Go!
Race
Cases
white
357
black or african american
51
asian
10
not reported
10

In this group, I got 225 files but only 203 cases.

Later, open a new webpage of GDC. I will select another group with all the same filters except Age at Diagnosis (From "**65**" to "**75**"). I got 146 files but only 128 cases.

Data downloading

After selecting all files to Cart in GDC, I have downloaded TCGA data by clicking **Manifest**.



You have to download "gdc-client" form GDC Data Transfer Tool by choosing **gdc-client_v1.6.1_OSX_x64.zip** and put it in your work directory and copy your work directory path into the ".zshrc" file like this:

```
vi ~/.zshrc

export PATH="/path to your gdc-client/:${PATH}"
```

Then, after reopening the terminal, please use the command:

```
nohup gdc-client download -m ~/path_to_your_file/your_manifest.txt &
```

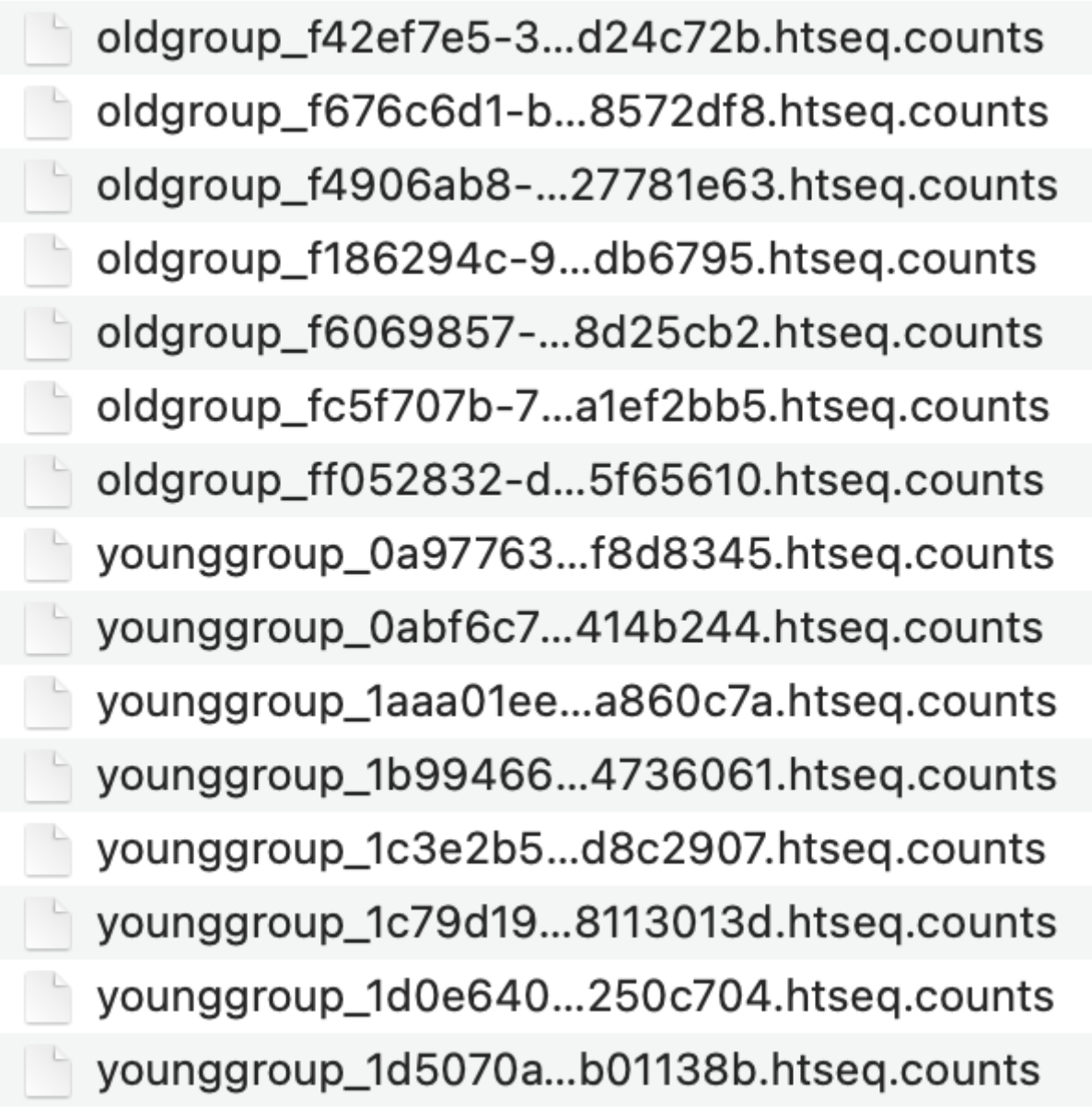
I will put all files in new directory.

unzip all files in by using the command:

```
gunzip *htseq.counts
```

The first group which age between 55-64, I will put them in a folder called "young" and change all their names with the prefix "younggroup".

The second group which age between 65-75, I will put them in a folder called "old" and change all their names with the prefix "oldgroup".



Then, merge all files into a new folder called "all".

Next Steps

I will run throught the SOP I presented above and try to ruduce errors within my contexts. Maybe run more data to test my script. Then, I will start to create plots from the vignette.

Data

I have uploaded" Sample_young.csv", "result.txt", and all my "htseq.counts" files.

Known Issues

I have met issue with the content in DESeq2 guildlines. However, after discussing with Dr. Craig, problems solved but still need to retest my whole testing scripts.

It is hard to put all files into the scripts that I run, but I will put more data and samples into my scripts eventually.

Milestone2

Modified my Milestone_1(optional)

I have modified my Milestone_1 with more details about how to download the data step by step. Then, I reloaded the data and put more screenshots to follow through.

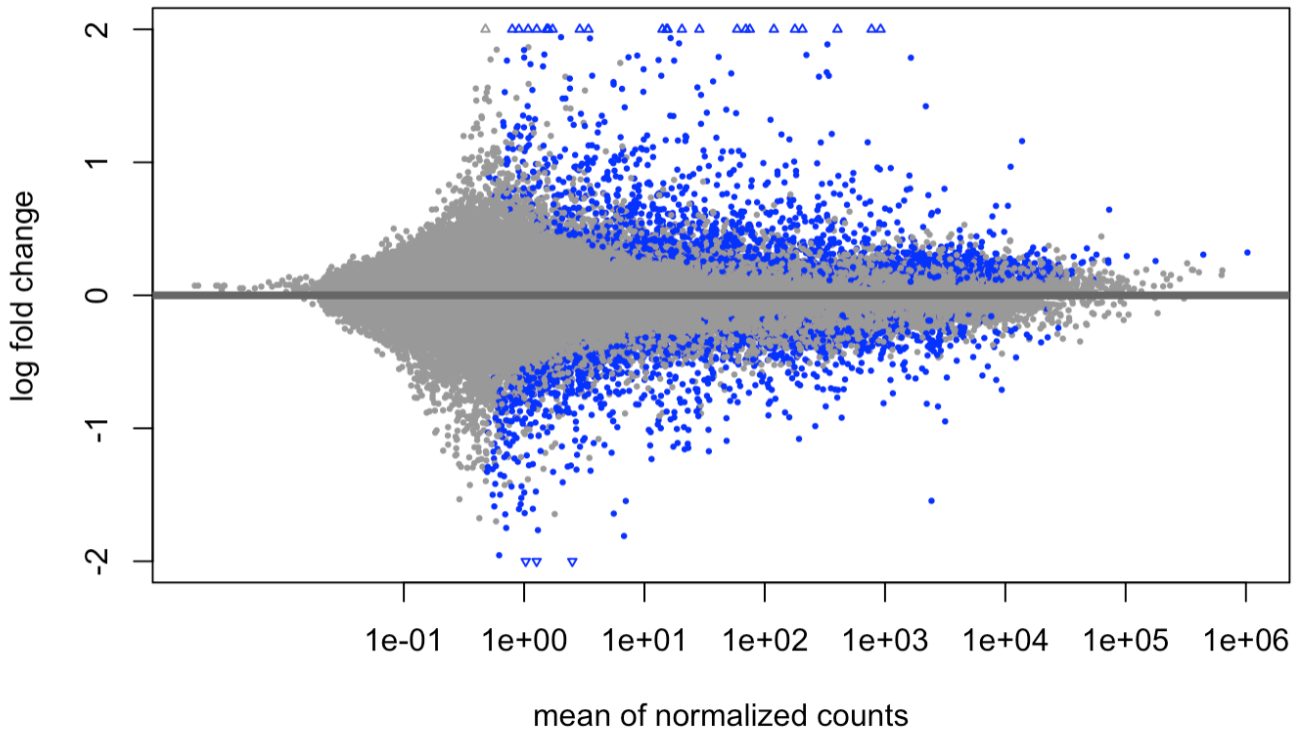
Input all samples

After testing with more files, now I will start putting all my samples in my "HTseq_counts_files" folder and going through the script that I previously made.

Differential expression analysis

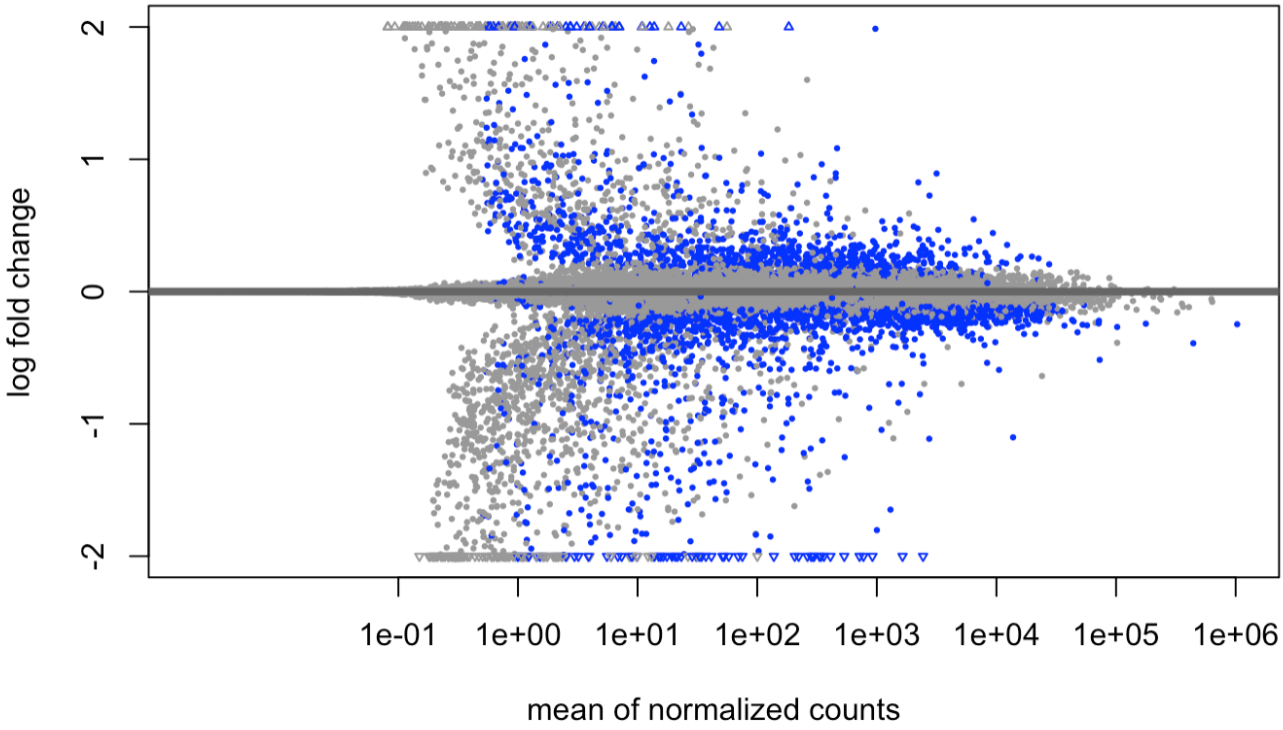
MA-plot

```{r}
plotMA(res, ylim=c(-2,2))
```



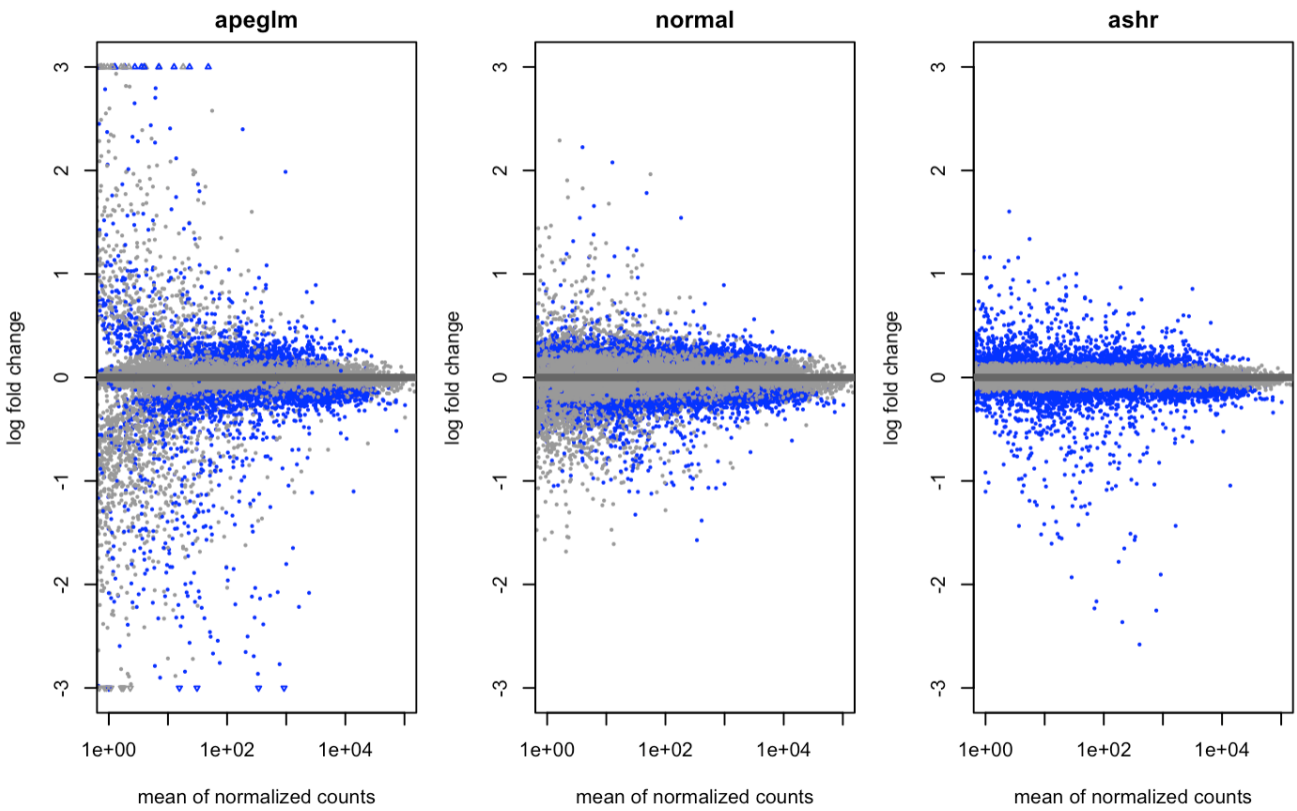
With this plot, I remove the noise associated with log2 fold changes from low count genes without requiring arbitrary filtering thresholds.

```{r}
plotMA(resLFC, ylim=c(-2,2))
```



Setting with 'coef=2'.

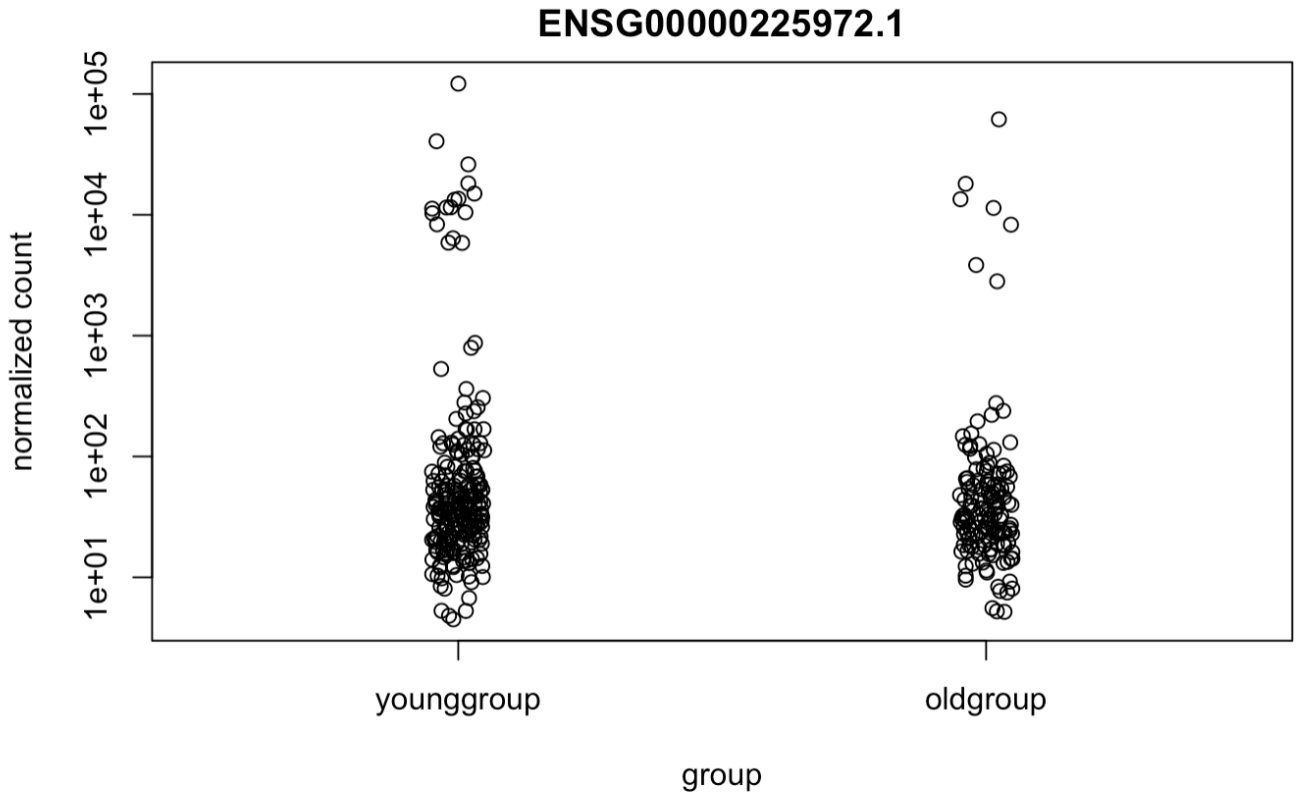
```{r}
par(mfrow=c(1,3), mar=c(4,4,2,1))
xlim <- c(1,1e5); ylim <- c(-3,3)
plotMA(resLFC, xlim=xlim, ylim=ylim, main="apeglm")
plotMA(resNorm, xlim=xlim, ylim=ylim, main="normal")
plotMA(resAsh, xlim=xlim, ylim=ylim, main="ashr")
```



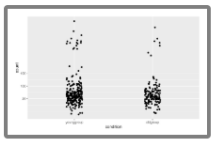
Plot counts

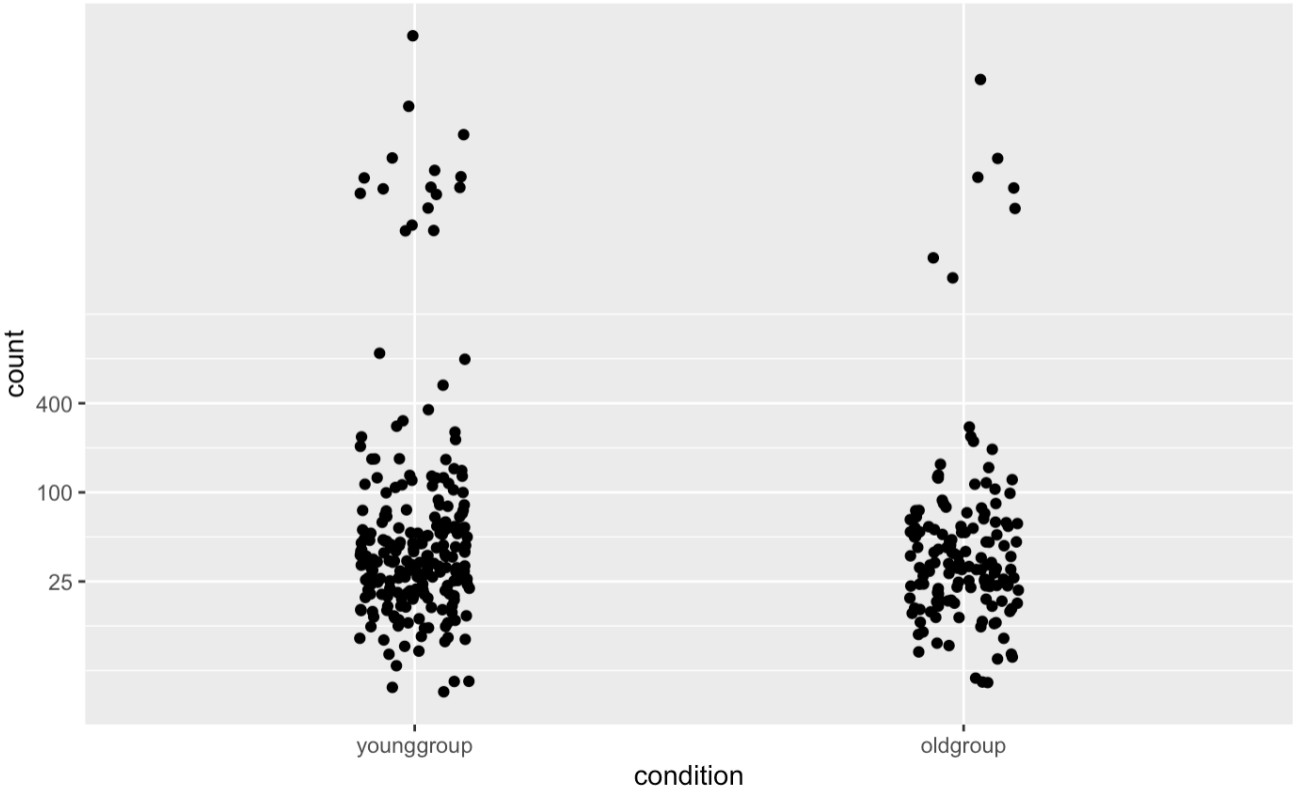
Examine the counts of reads for a single gene across the groups.

```{r}
plotCounts(dds, gene=which.min(res$padj), intgroup="condition")
```



```{r}
d <- plotCounts(dds, gene=which.min(res$padj), intgroup="condition",
                returnData=TRUE)
library("ggplot2")
ggplot(d, aes(x=condition, y=count)) +
  geom_point(position=position_jitter(w=0.1,h=0)) +
  scale_y_log10(breaks=c(25,100,400))
```



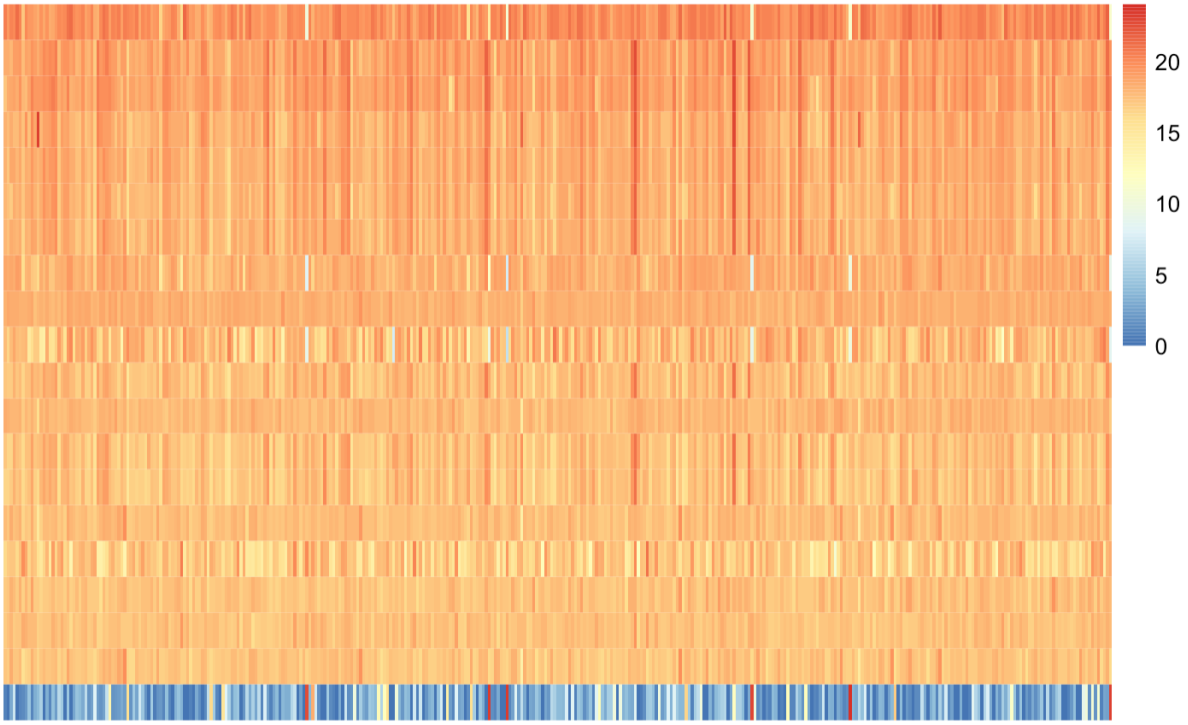


Heatmap

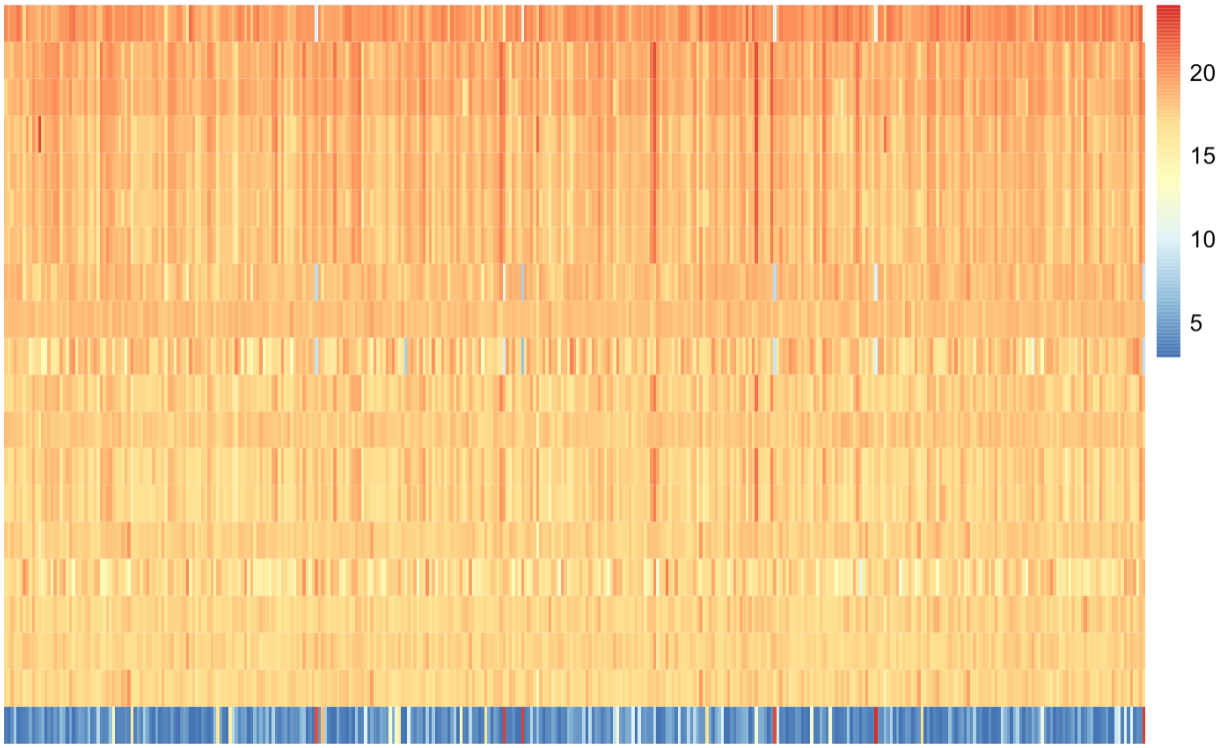
I will show how to produce a heatmap for various transformations of the data.

```{r}
library("pheatmap")
select <- order(rowMeans(counts(dds,normalized=TRUE)),
                decreasing=TRUE)[1:20]
df <- as.data.frame(colData(dds)[,c("condition")])

pheatmap(assay(ntd)[select,], cluster_rows=FALSE, cluster_cols=FALSE, show_rownames = F, show_colnames = F) #,
annotation_col=df)
```

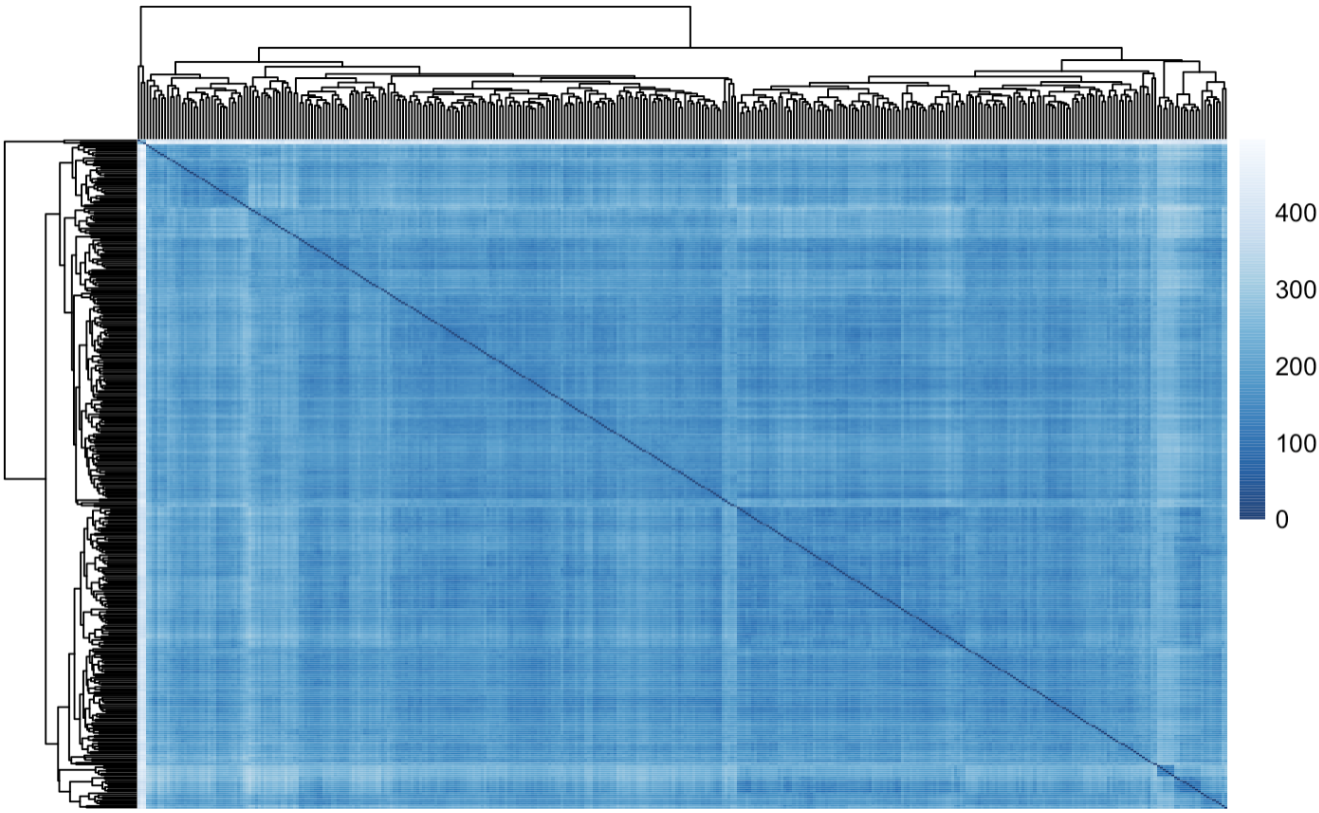



```{r}
pheatmap(assay(vsd)[select,], cluster_rows=FALSE, cluster_cols=FALSE, show_rownames=FALSE, show_colnames =
F)#,cluster_cols=FALSE, annotation_col=df)
```



Heatmap of the sample-to-sample distances.

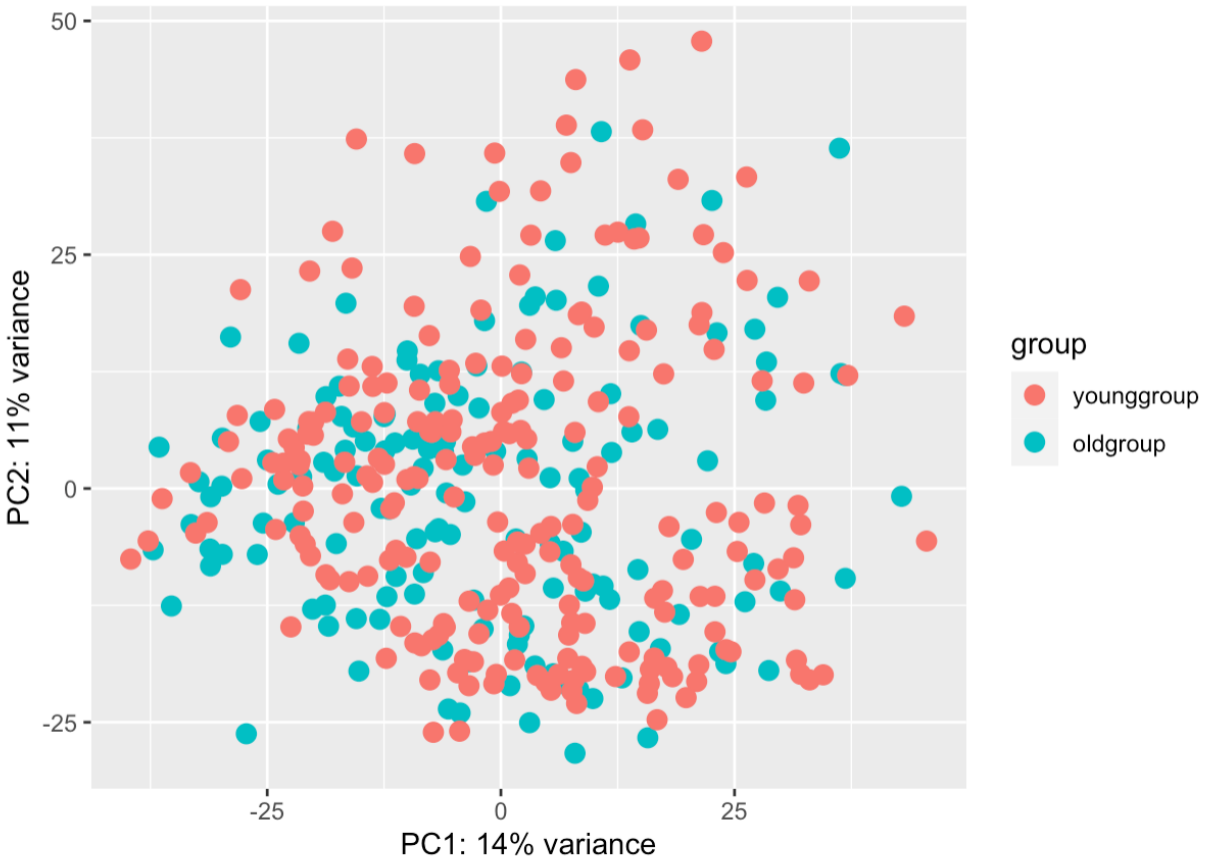
```{r}
library("RColorBrewer")
sampleDistMatrix <- as.matrix(sampleDists)
rownames(sampleDistMatrix) <- paste(vsd$condition, vsd$type, sep="-")
colnames(sampleDistMatrix) <- NULL
colors <- colorRampPalette( rev(brewer.pal(9, "Blues")) )(255)
pheatmap(sampleDistMatrix,
         clustering_distance_rows=sampleDists,
         clustering_distance_cols=sampleDists,
         col=colors, show_rownames=FALSE)
```



Principal component plot

It shows the samples in the 2D plane spanned by their first two principal components.

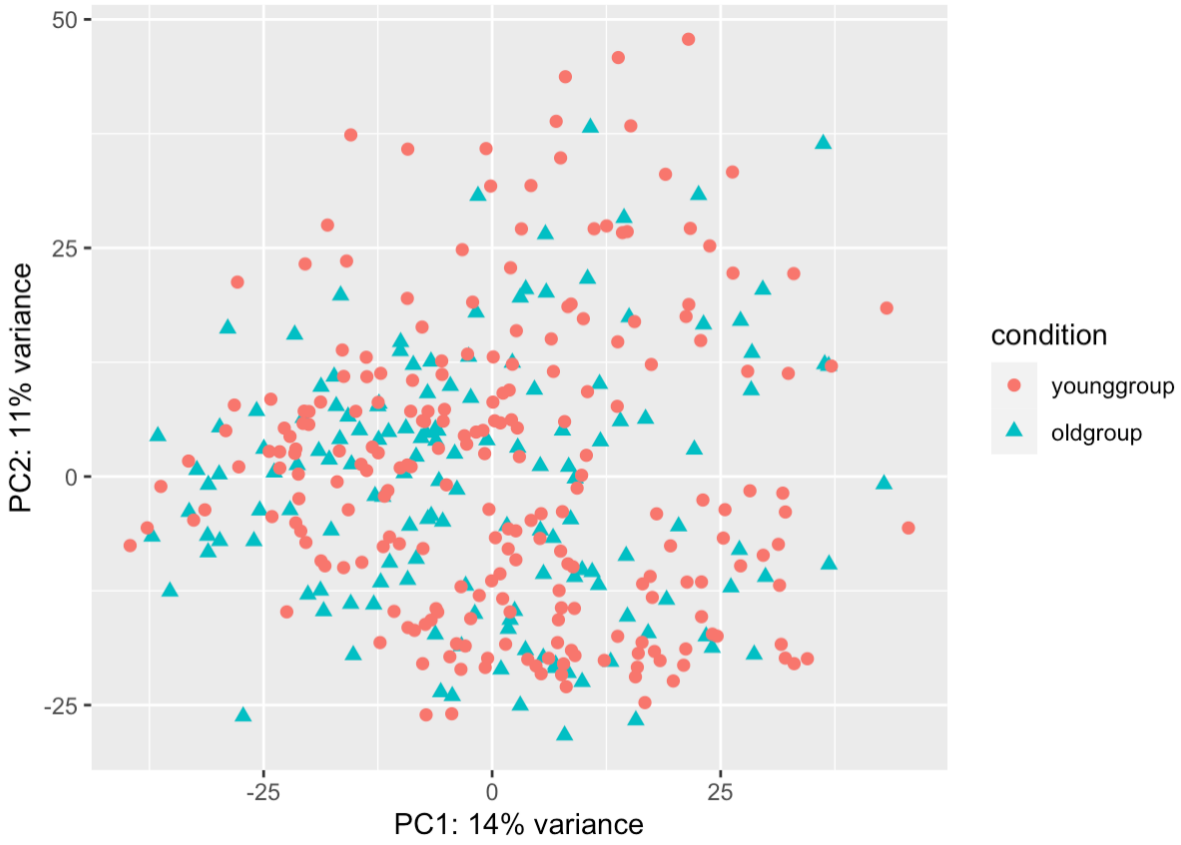
```{r}
plotPCA(vsd, intgroup=c("condition"))
```



I Acustomize the PCA plot using the ggplot function.

```{r}
pcaData <- plotPCA(vsd, intgroup=c("condition"), returnData=TRUE)
percentVar <- round(100 * attr(pcaData, "percentVar"))
ggplot(pcaData, aes(PC1, PC2, color=condition, shape=condition)) +
  geom_point(size=2) +
  xlab(paste0("PC1: ",percentVar[1],"% variance")) +
  ylab(paste0("PC2: ",percentVar[2],"% variance")) +
  coord_fixed()

```



Data

All of my data will be uploaded to my GitHub acount.

Feedback

Change all my ensembl id to real gene names.

See how many genes are significantly different (up-regulated or down-regulated).

Try to get no more than 500 genes.

Look into the link Dr. Craig gave us to look into the biology part(put in the gene list I have).

Heatmaps need to be fixed.

Known issues

I will change my p-value to 0.05 and 2foldchange to 2 and see what will happen with my plots.

I will try to change all my ensemble id to hugo id.

I have faced a problem with the heatmap error. Everytime I try to put the reference of "annotation_col=df" into my code, it will not work.

Deliverable

A complete repository with clear documentation and description of my analysis and results.