Class 17

Neha (PID: A17567541)

With each sample having its own directory containing the Kallisto output, we can import the transcript count estimates into R using the Bioconductor package `tximport`

```
library(tximport)
library(rhdf5)
```

```
# setup the folder and filenames to read
folders <- dir(pattern="SRR21568*", all.files=TRUE)
folders
```

```
[1] "SRR2156848_quant" "SRR2156849_quant" "SRR2156850_quant" "SRR2156851_quant"
```

```
samples <- sub("_quant", "", folders)
files <- file.path( folders, "abundance.h5" )
files
```

```
[1] "SRR2156848_quant/abundance.h5" "SRR2156849_quant/abundance.h5"
[3] "SRR2156850_quant/abundance.h5" "SRR2156851_quant/abundance.h5"
```

```
names(files) <- samples
```

```
txi.kallisto <- tximport(files, type = "kallisto", txOut = TRUE)
```

```
1 2 3 4
```

```
head(txi.kallisto$counts)
```

```
                SRR2156848 SRR2156849 SRR2156850 SRR2156851
ENST00000539570          0          0    0.00000          0
ENST00000576455          0          0    2.62037          0
ENST00000510508          0          0    0.00000          0
ENST00000474471          0          1    1.00000          0
ENST00000381700          0          0    0.00000          0
ENST00000445946          0          0    0.00000          0
```

```
colSums(txi.kallisto$counts)
```

```
SRR2156848 SRR2156849 SRR2156850 SRR2156851
   2563611    2600800    2372309    2111474
```

```
sum(rowSums(txi.kallisto$counts)>0)
```

```
[1] 94561
```

We will then filter out the annotated transcripts with no reads.

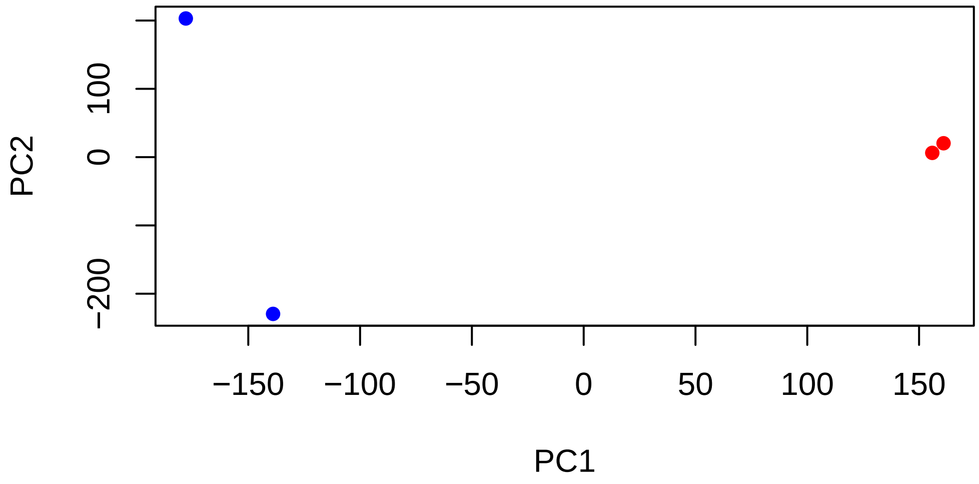
```
to.keep<-rowSums(txi.kallisto$counts)>0
kset.nonzero<-txi.kallisto$counts[to.keep,]


keep2 <-apply(kset.nonzero,1,sd)>0
x<-kset.nonzero[keep2,]
```

Principal component analysis:

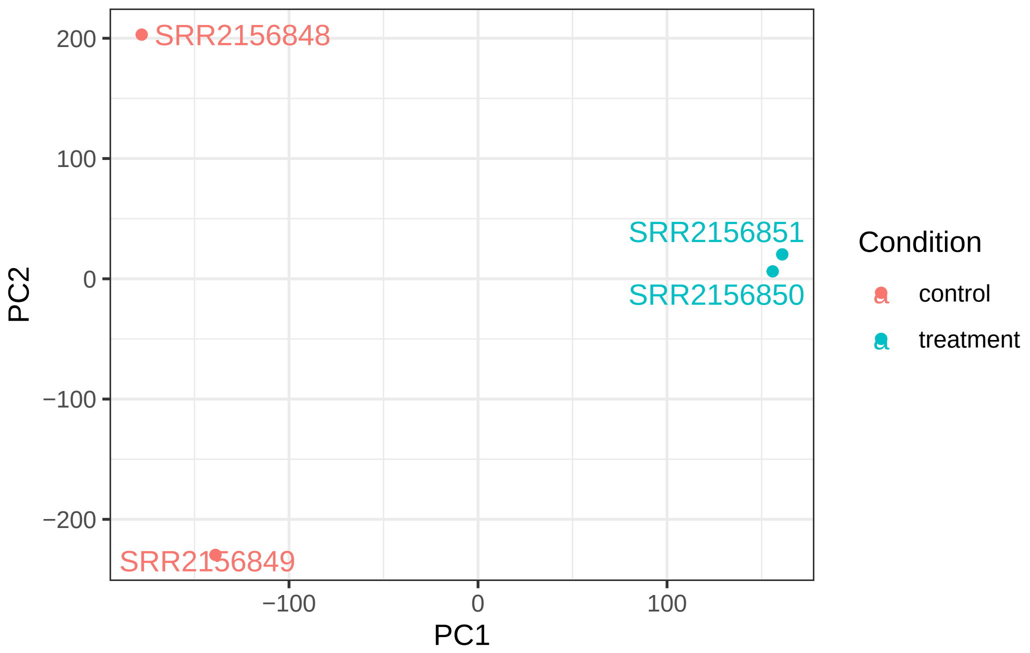
```
pca<-prcomp(t(x),scale=TRUE)
summary(pca)
```

```
Importance of components:
                            PC1      PC2      PC3   PC4
Standard deviation     183.6379 177.3605 171.3020 1e+00
Proportion of Variance   0.3568   0.3328   0.3104 1e-05
Cumulative Proportion    0.3568   0.6895   1.0000 1e+00
```

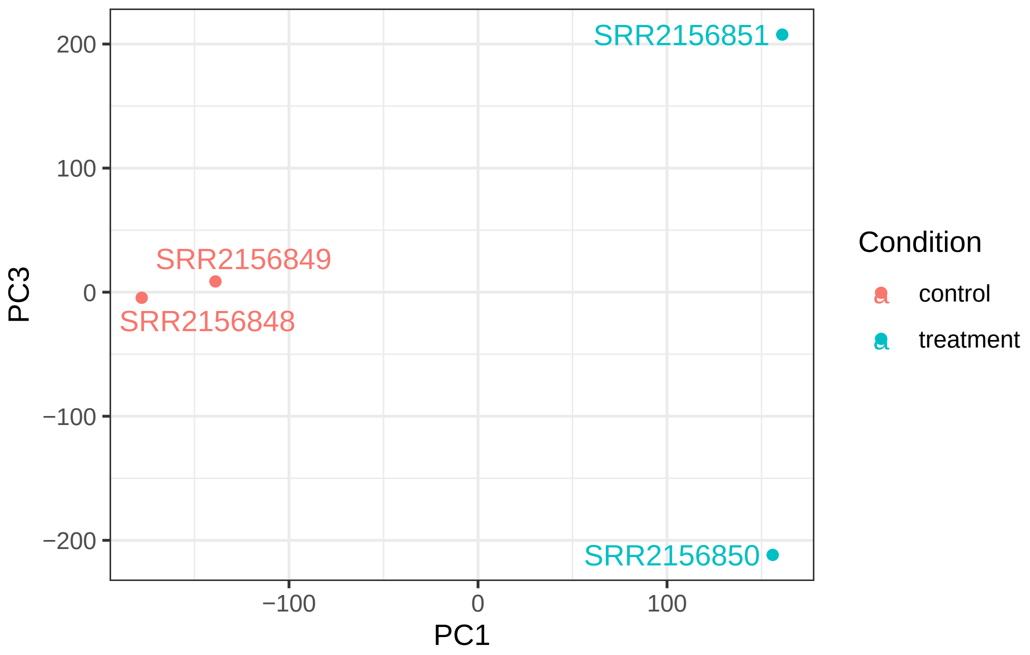
```r
plot(pca$x[,1], pca$x[,2],col=c("blue","blue","red","red"), xlab="PC1", ylab="PC2", pch=16
```



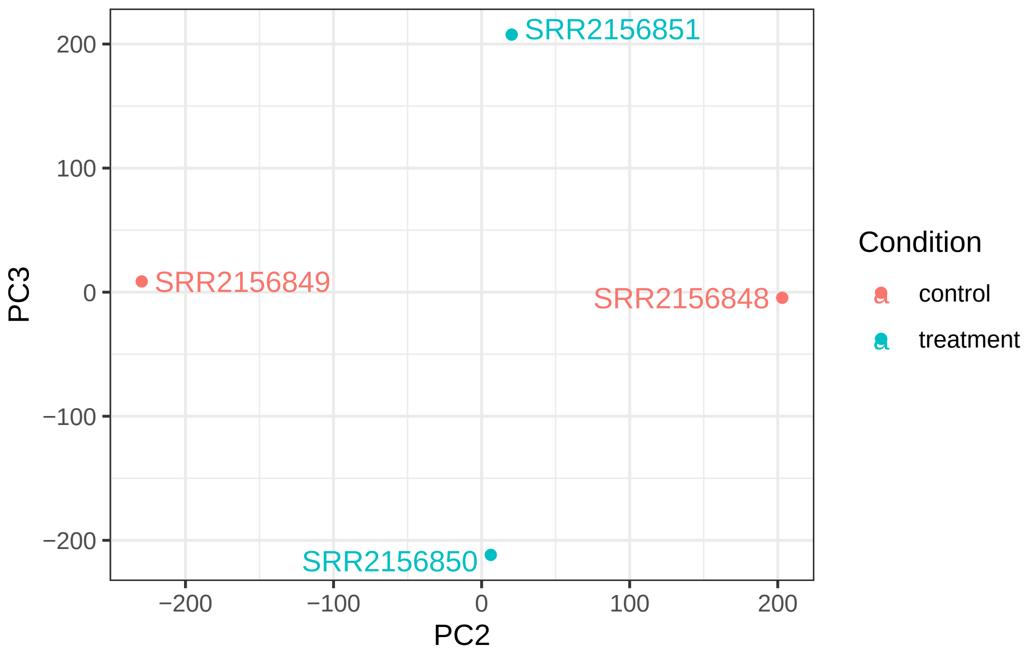
```r
library(ggplot2)
library(ggrepel)

# Make metadata object for the samples
colData <- data.frame(condition = factor(rep(c("control", "treatment"), each = 2)))
rownames(colData) <- colnames(txi.kallisto$counts)

# Make the data.frame for ggplot
y <- as.data.frame(pca$x)
y$Condition <- as.factor(colData$condition)

ggplot(y) +
  aes(PC1, PC2, col=Condition) +
  geom_point() +
  geom_text_repel(label=rownames(y)) +
  theme_bw()
```



```
ggplot(y) +
  aes(PC1, PC3, col=Condition) +
  geom_point() +
  geom_text_repel(label=rownames(y)) +
  theme_bw()
```



```
ggplot(y) +
  aes(PC2, PC3, col=Condition) +
  geom_point() +
  geom_text_repel(label=rownames(y)) +
  theme_bw()
```


SRR2156851
SRR2156849
SRR2156848
SRR2156850
200
100
0
−100
−200
PC3
−200
−100
0
100
200
PC2
Condition
control
treatment