Class 17 Lab

Snehita Vallumchetla

Downstream Analysis

In order to analyze the kallisto results, I downloaded the `tximport` and `rhdf5` Bioconductor packages.

```
library(tximport)

# setup the folder and filenames to read
folders <- dir(pattern="SRR21568*")
samples <- sub("_quant", "", folders)
files <- file.path( folders, "abundance.h5" )
names(files) <- samples

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

```
1 2 3
```

Take a look at the top counts of our kallisto results.

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

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

By summing the columns, we are able to see how many transcripts we have for each sample

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

```
SRR2156848 SRR2156849 SRR2156850
   2563611    2111474    2372309
```

We can also determine how many transcripts are accounted for in at least one sample

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

```
[1] 86758
```

Let us also filter out the samples that have no reads to set us up for future analysis.

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

also removed the samples that showed no change over time.

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

##Principle Component Analysis

We can now set up a PCA of our counts matrix: centering and scaling each transcript's measured levels so that each feature contributes equally to the PCA

Setting up the pca object:

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

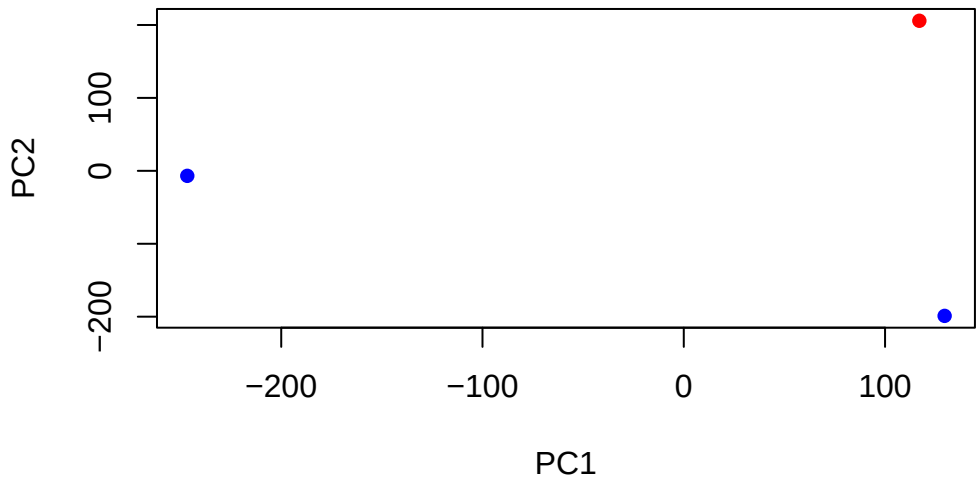
Observing summary statistics:

```
summary(pca)
```

```
Importance of components:
                            PC1      PC2     PC3
Standard deviation     213.7083 202.4346 1.41418
Proportion of Variance   0.5271   0.4729 0.00002
Cumulative Proportion    0.5271   1.0000 1.00000
```

We can make a plot in R of PC1 vs PC2

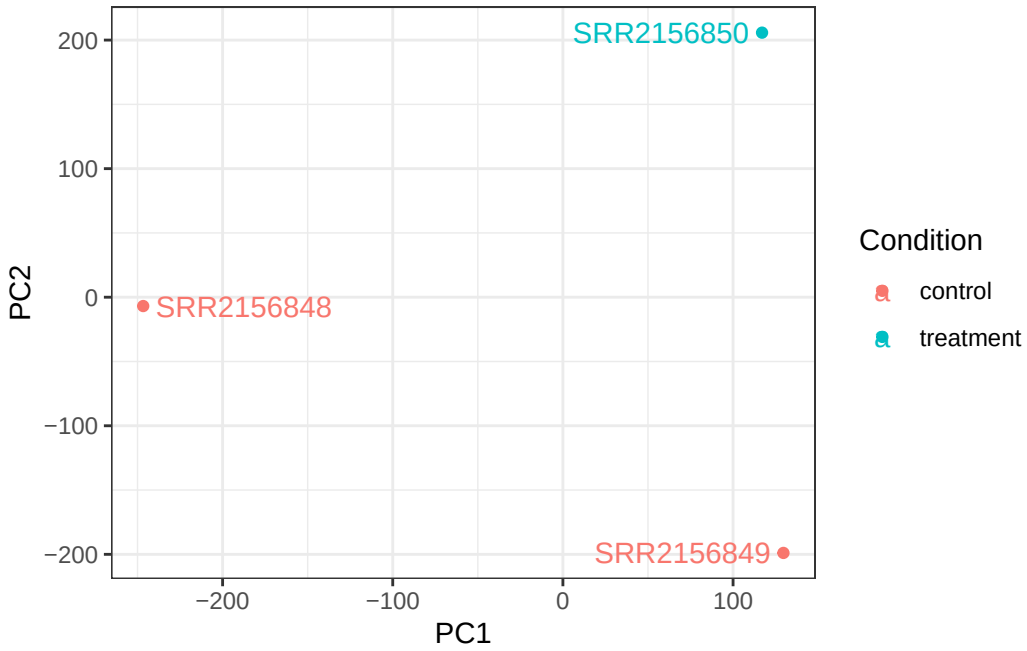
```
plot(pca$x[,1], pca$x[,2],
     col=c("blue","blue","red","red"),
     xlab="PC1", ylab="PC2", pch=16)
```



Let us make a nicer plot with more annotation using ggplot, for PC1 vs PC2, PC1 vs PC3, and PC2 vs PC3.

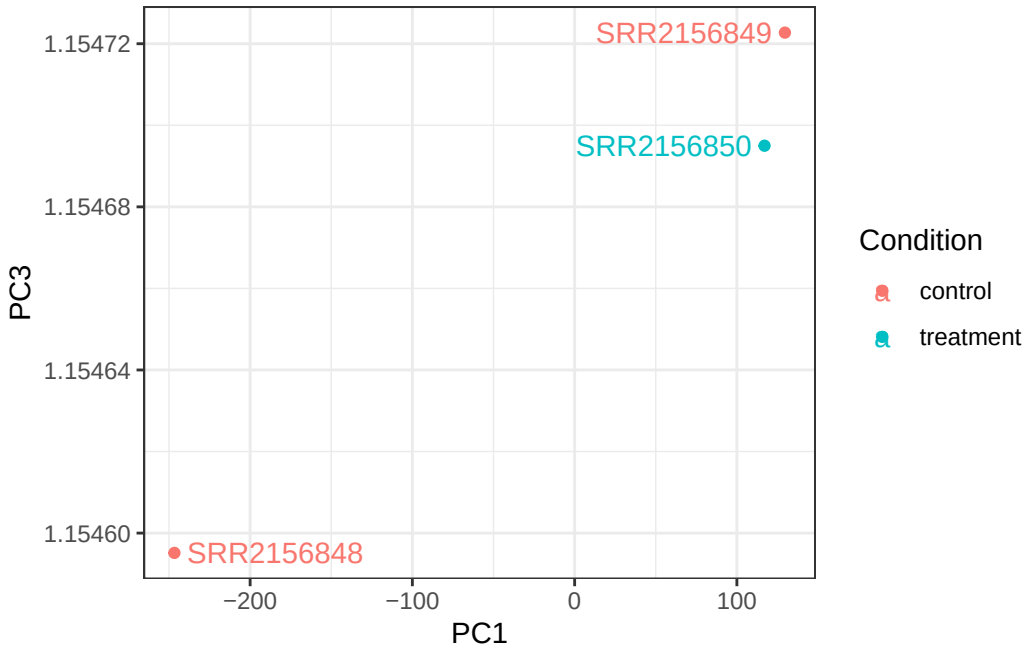
```
library(ggplot2)
library(ggrepel)

# Make metadata object for the samples
colData <- data.frame(condition = factor(rep(c("control", "control", 'treatment'))))
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()
```

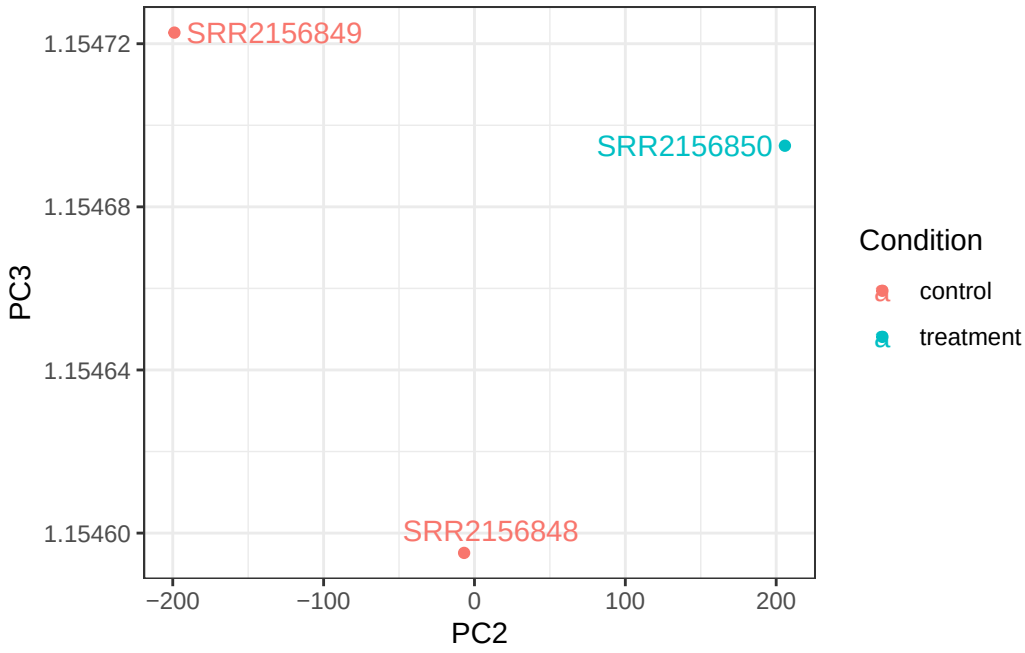


Looking at the PC1 vs PC2 we can see how the treatment group is much higher up on the y axis in comparison to the controls, showing seperation visually.

```
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()
```


SRR2156849
SRR2156850
SRR2156848
1.15472
1.15468
1.15464
1.15460
PC3
−200
−100
0
100
200
PC2
Condition
control
treatment