Class 17: Analyzing Sequence Data in the Cloud

Hyejeong Choi (PID: A16837133)

Table of contents

Downstream Analysis

Import Kallisto results using the `tximport()` function.

```
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 4
```

Look at the transcript count estimates:

```
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
```

Look at the total number of transcript counts in each sample by adding the column:

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

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

Look at how many transcripts are found by adding the total of the rows:

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

```
[1] 94561
```

Remove the transcripts that have no reads in the data:

```r
# add the rows and keep the data that are greater than zero
to.keep <- rowSums(txi.kallisto$counts) > 0

# create a new dataset
kset.nonzero <- txi.kallisto$counts[to.keep,]
```

```r
# keep the data that change between the samples and remove the data that do not
keep2 <- apply(kset.nonzero,1,sd)>0

# create a new dataset
x <- kset.nonzero[keep2,]
```

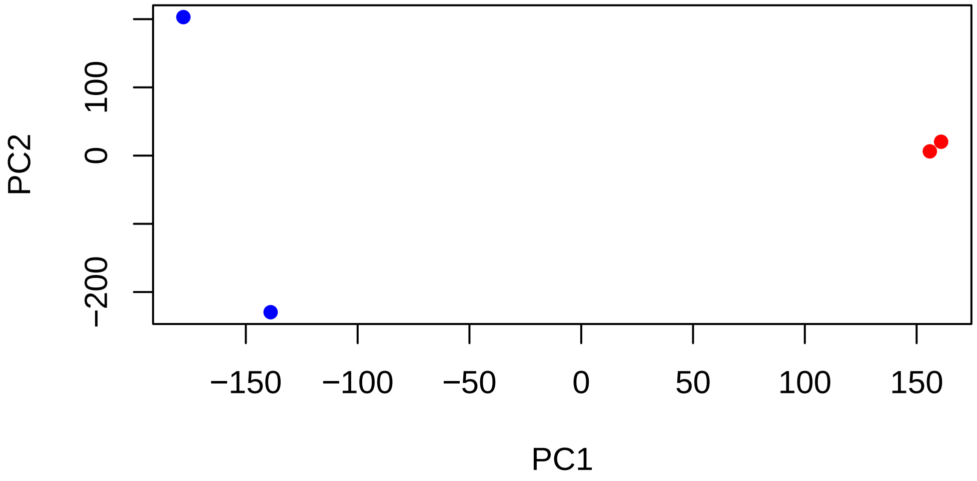
Principal Component Analysis

```
# transpose the x dataset and scale it
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
```

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



> Q. Use ggplot to make a similar figure of PC1 vs PC2 and a seperate figure PC1 vs PC3 and PC2 vs PC3.

First create a dataframe for grouping the control and treatment groups:

```
# create a dataframe and group the samples into control and treatment
# use factor() to turn the characters into a factor for easier coloring using discrete value

colors <- data.frame(group=factor(c('control','control', 'treatment','treatment')))

# make the rownames the sample names

rownames(colors) <- rownames(pca$x)

colors
```

```
               group
SRR2156848   control
SRR2156849   control
SRR2156850 treatment
SRR2156851 treatment
```

Add the group dataframe as another column into a PCA dataframe:

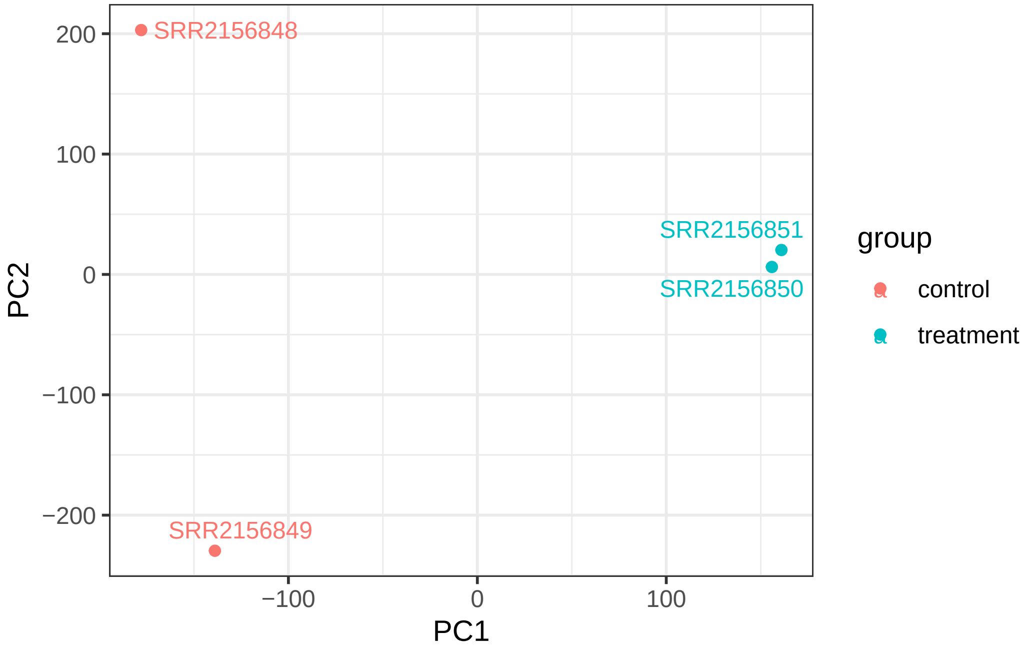
```
# convert the pca$x into a dataframe to add the group column

new_pca <- as.data.frame(pca$x)

# add the group column

new_pca$group <- colors$group

new_pca
```

```
                 PC1         PC2         PC3       PC4     group
SRR2156848 -177.9368  203.031882   -4.507483 0.8660196   control
SRR2156849 -138.9188 -229.558755    8.656814 0.8659919   control
SRR2156850  155.8981    6.206921 -211.755452 0.8660168 treatment
SRR2156851  160.9486   20.312009  207.599341 0.8660462 treatment
```

PC1 vs PC2

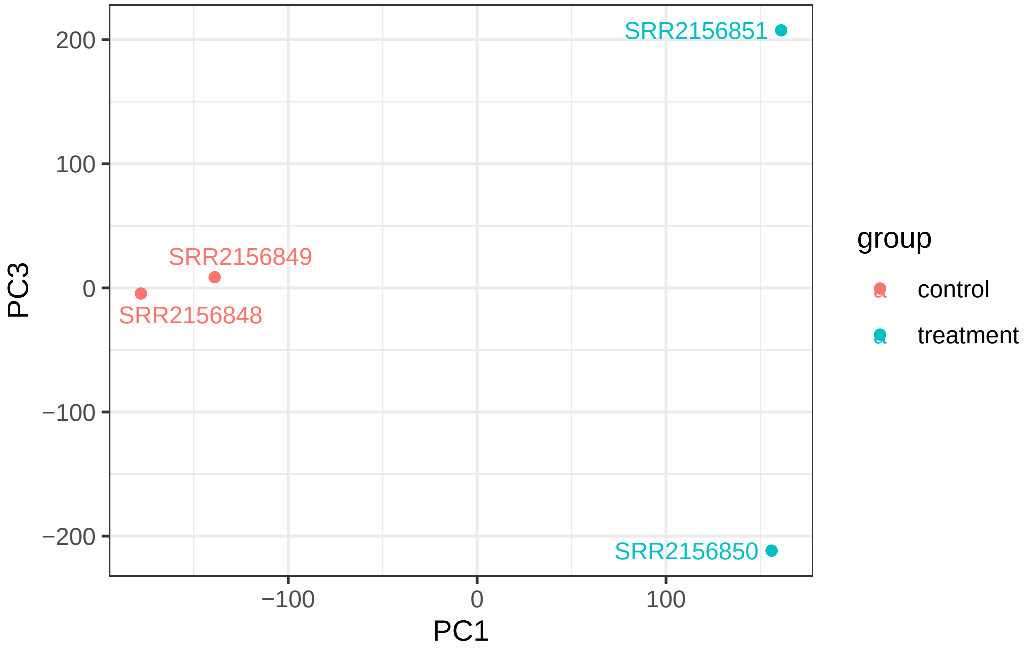
```
library(ggplot2)
library(ggrepel)
```

```
ggplot(new_pca) +
  aes(PC1, PC2, label=rownames(new_pca), col=group) +
  geom_point() +
  geom_text_repel(size=3) +
  theme_bw()
```



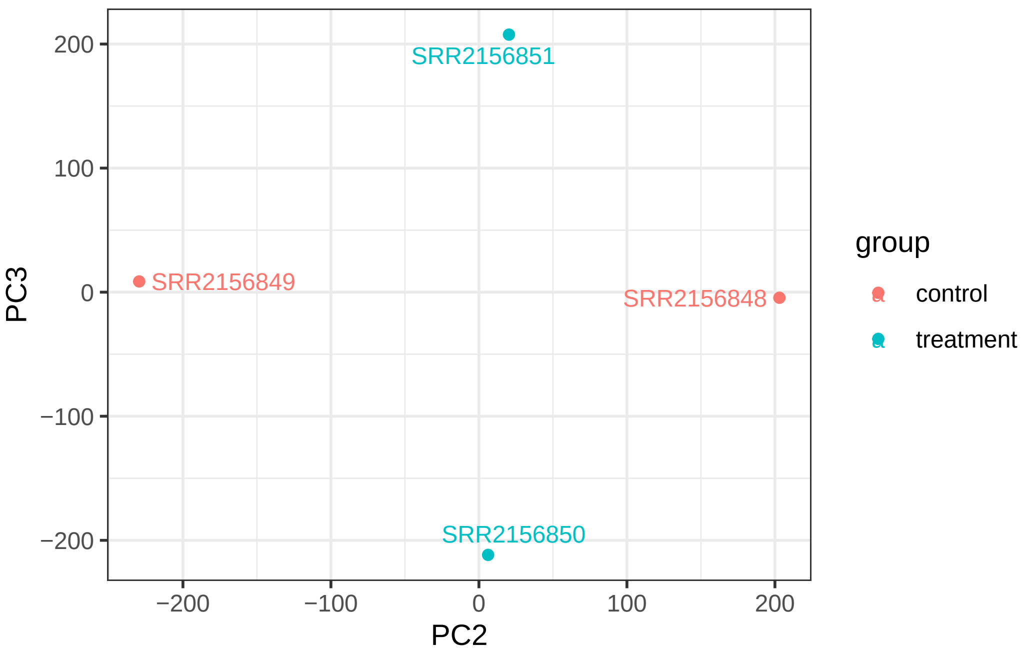
PC1 vs PC3

```
ggplot(new_pca) +
  aes(PC1, PC3, label=rownames(new_pca), col=group) +
  geom_point() +
  geom_text_repel(size=3) +
  theme_bw()
```



PC2 vs PC3

```
ggplot(new_pca) +
  aes(PC2, PC3, label=rownames(new_pca), col=group) +
  geom_point() +
  geom_text_repel(size=3) +
  theme_bw()
```


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