~~~
---
title: "Class 17: Analyzing Sequencing Data in the Cloud"
author: "Claire Lua A16922295"
format: pdf
---

```{r}
library(tximport)

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)
```

```{r}
head(txi.kallisto$counts)
```

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

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

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

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


## PCA

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

```{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()
```
~~~