Read counts to DGE: Exploratory plots

```
library(DESeq2)
library(magrittr)
library(ggplot2)
```

```
load(file = "RNAseqGierlinski.RData")
```

Similarity assessments and clustering

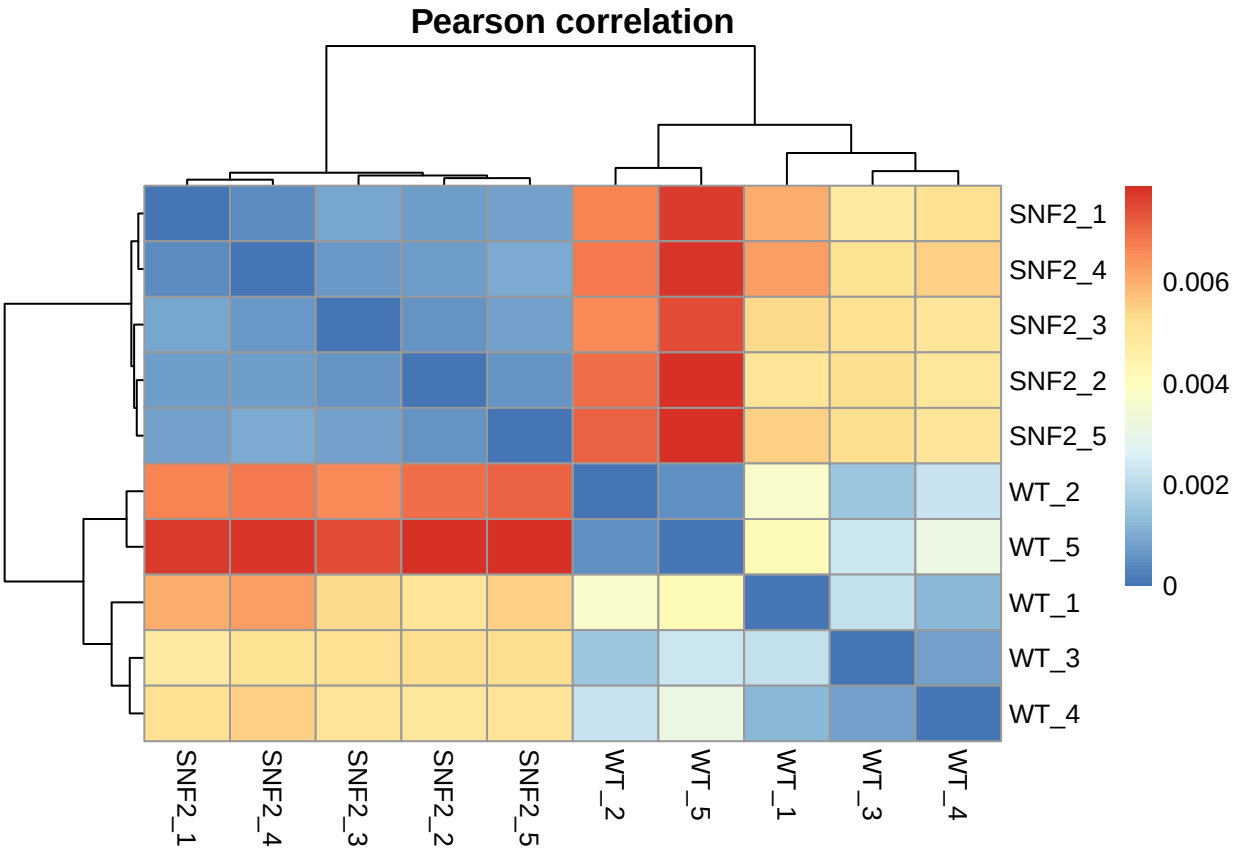
Sample clustering using Pearson correlation

The ENCODE consortium recommends that *"for messenger RNA, (. . .) biological replicates [should] display greater than 0.9 correlation for transcripts/features"*.

The Pearson correlation coefficient is a measure of the strength of the linear relationship between two variables and is often used to assess the similarity of RNA-seq samples in a pair-wise fashion. It is defined as the **covariance of two variables divided by the product of their standard deviation**.

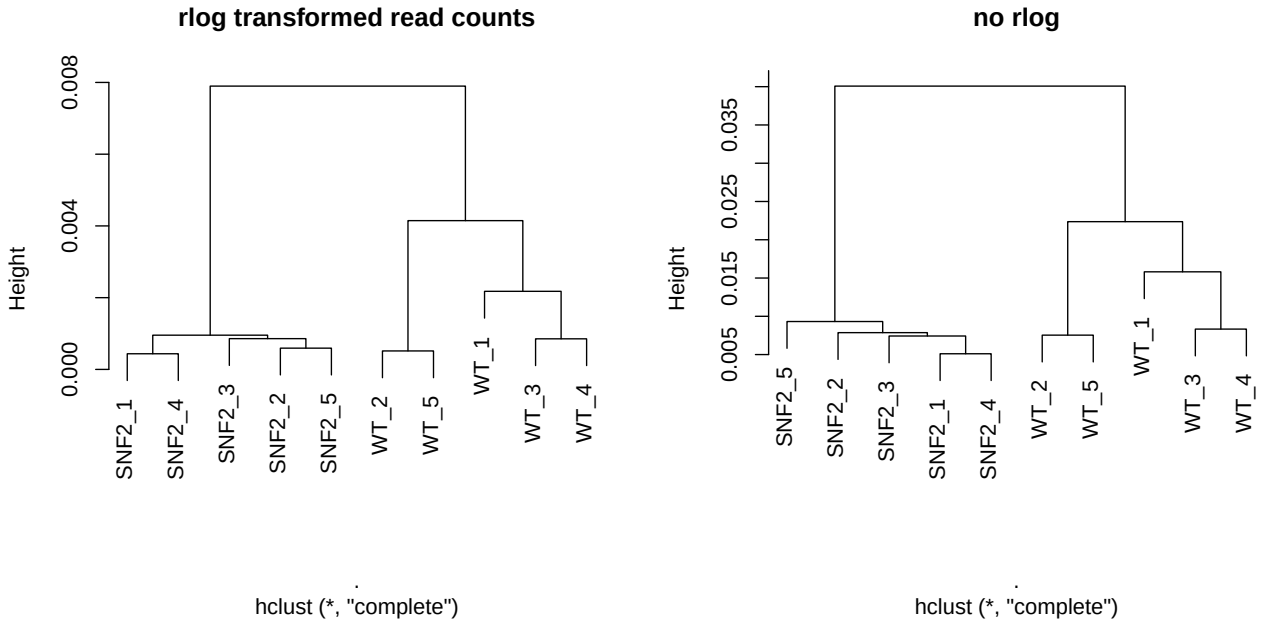
Clustered heatmap of correlation coefficients:

```
corr_coeff <- cor(rlog.norm.counts, method = "pearson")
as.dist(1-corr_coeff, upper = TRUE) %>%as.matrix %>%
  pheatmap::pheatmap(., main = "Pearson correlation")
```



Just plot the dendrogram, comparing the effects of the `rlog` transformation.

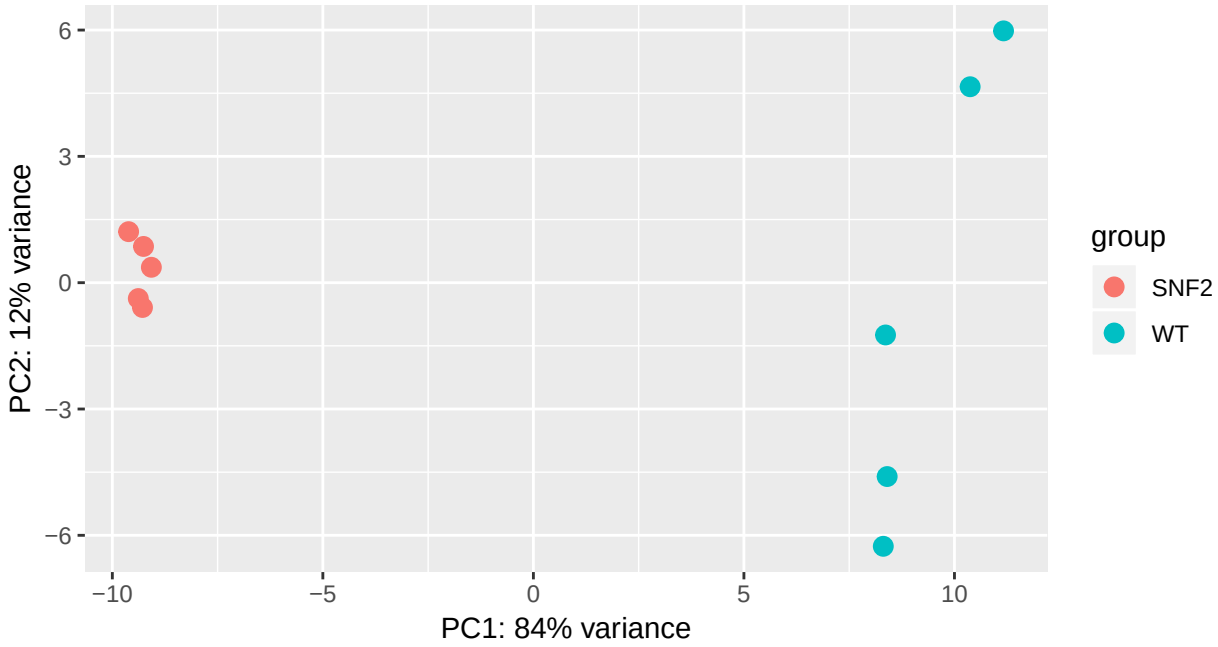
```
par(mfrow=c(1,2))
# Pearson corr. for rlog.norm values
as.dist(1 - corr_coeff) %>% hclust %>%
  plot( ., labels = colnames(rlog.norm.counts),
        main = "rlog transformed read counts")

# Pearson corr. for log.norm.values
as.dist( 1 - cor(log.norm.counts, method = "pearson")) %>%
  hclust %>% plot( ., labels = colnames(log.norm.counts),
                   main = "no rlog")
```



PCA

```
rv <- rowVars(assay(DESeq.rlog)) # equivalent to rowVars(rlog.norm.counts)
top_variable <- order(rv, decreasing = TRUE)[seq_len(500)]
pca <- prcomp(t(assay(DESeq.rlog)[top_variable, ]))
head(pca$x)
```

```
 plotPCA(DESeq.rlog)
```



pcaExplorer

pcaExplorer lets you interact with the DESeq2-based plots and analyses. It has included hierarchical clustering of samples and PCA.

```
#source("https://bioconductor.org/biocLite.R")
#biocLite("pcaExplorer")
pcaExplorer::pcaExplorer(dds = DESeq.ds, rlt = DESeq.rlog)
```