STAT718/BIOL703: Genomic Data Science
QC Methods for DE analysis

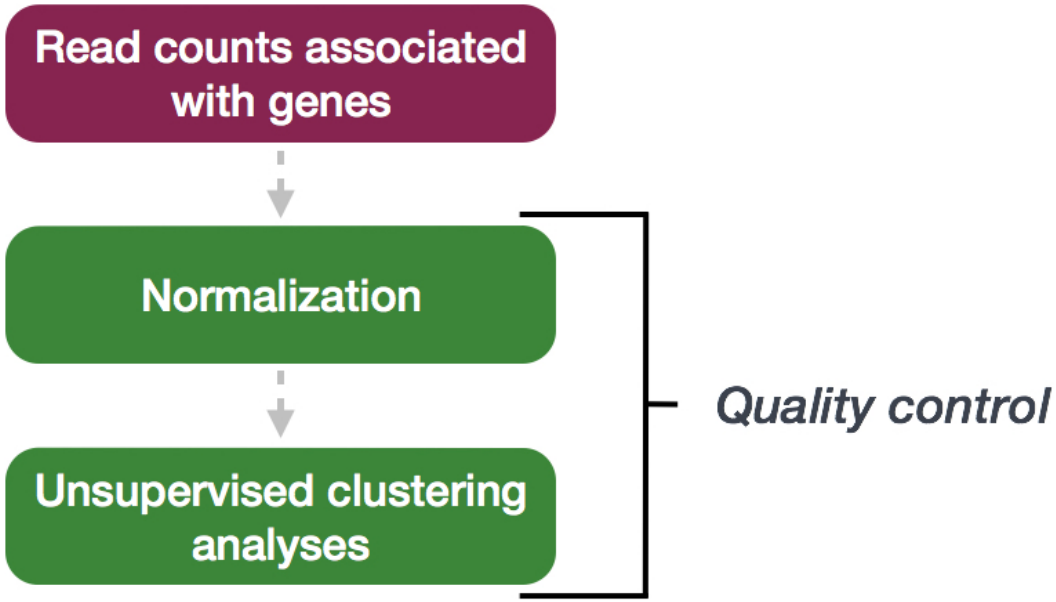
Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

Learning Objectives

- Transforming counts for unsupervised clustering methods
- Evaluating quality of samples using Principal Components Analysis
- Hierarchical clustering of samples in the dataset

Quality Control

The next step in the DESeq2 workflow is QC, which includes sample-level and gene-level steps to perform QC checks on the count data to help us ensure that the samples/replicates look good.



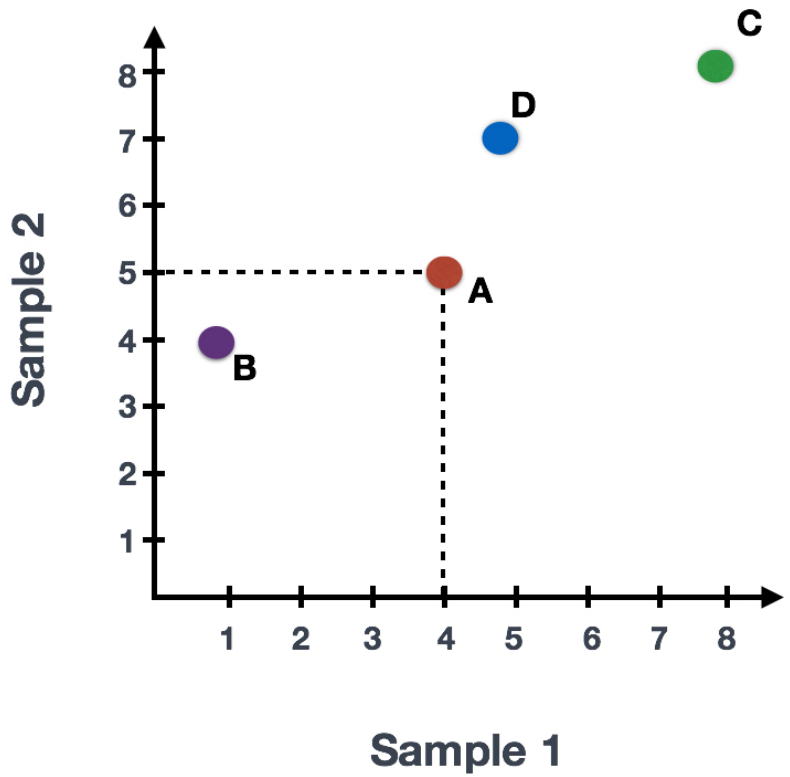
Sample-Level QC

A useful initial step in an RNA-seq analysis is often to assess overall similarity between samples:

- Which samples are similar to each other, which are different?
- Does this fit to the expectation from the experiment's design?
- What are the major sources of variation in the dataset?

Principal Component Analysis (PCA)

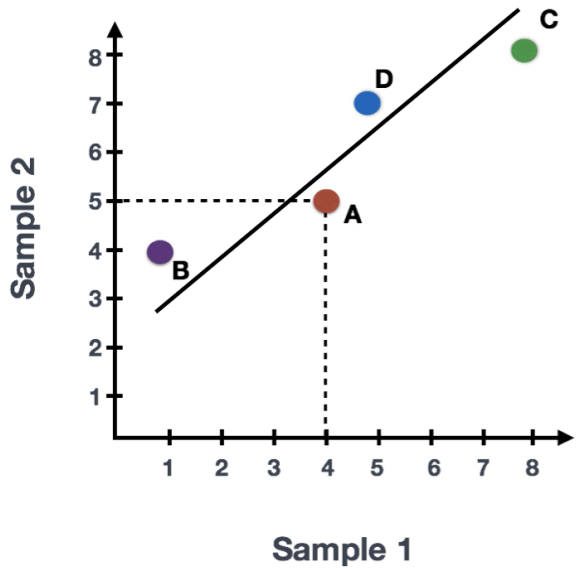
Suppose we had a dataset with two samples and four genes. Based on this expression data we want to evaluate the relationship between these samples. We could plot the counts of one sample versus another, with Sample 1 on the x-axis and Sample 2 on the y-axis as shown below:



	Sample 1	Sample 2
Gene A	4	5
Gene B	1	4
Gene C	8	8
Gene D	5	7

Principal Component Analysis (PCA)

For PCA analysis, the first step is taking this plot and drawing a line through the data in the direction representing the most variation. In this example, the most variation is along the diagonal. That is, the largest spread in the data is between the two endpoints of this line. This is called the first principal component, or PC1. The genes at the endpoints of this line (Gene B and Gene C) have the greatest influence on the direction of this line.

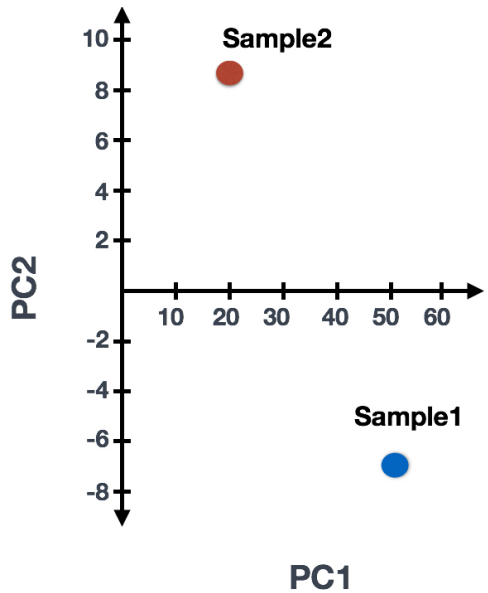


Principal Component Analysis (PCA)

After drawing this line and establishing the amount of influence per gene, PCA will compute a per sample score.

Sample1 PC1 score = (read count Gene A * influence Gene A) + (read count Gene B * influence Gene B) + .. for all genes

The end result is a 2-dimensional matrix with rows representing samples and columns reflecting scores for each of the principal components.



	PC1	PC2
Sample1	51	-7
Sample2	21	8.5

If two samples have similar levels of expression for the genes that contribute significantly to the variation represented by PC1, they will be plotted close together on the PC1 axis.

Example: Airway read counts

```
> library("airway")
> library("edgeR")
> data(airway)
> countMat<-assay(airway)
```

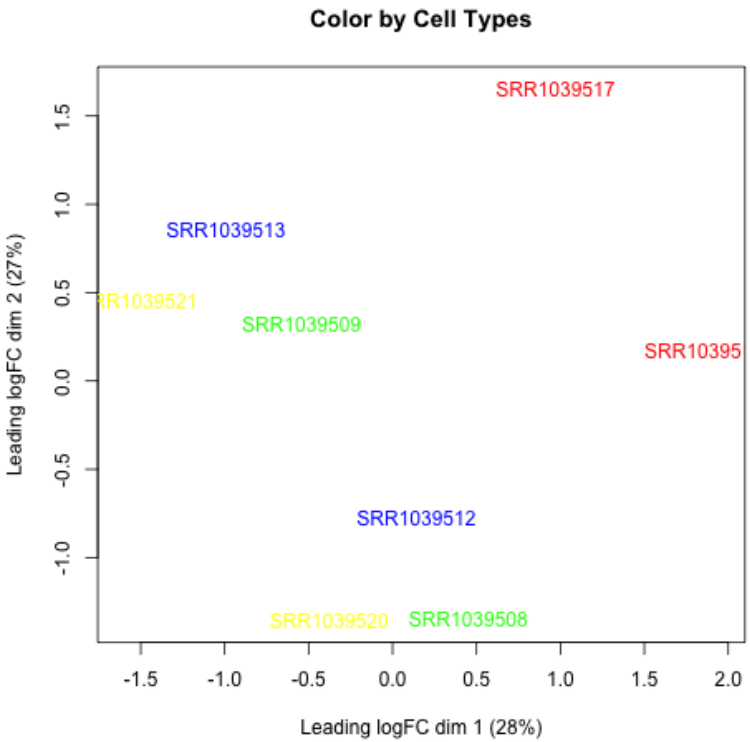
	1.bam	2.bam	3.bam	4.bam	5.bam	...
ENSG00000009724	38	28	66	24	42	
ENSG00000116649	1004	1255	1122	1313	1100	
ENSG00000120942	218	256	233	252	269	
...						

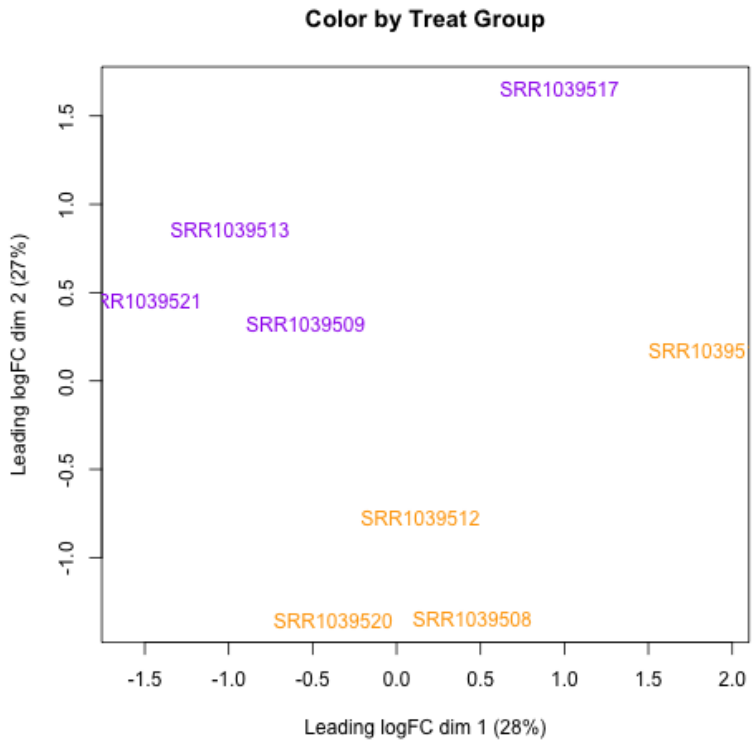
Data Exploration: Multidimensional Scaling

In R, *plotMDS* can be used to visualize the results from a principle components analysis, which determines the greatest sources of variation in the data. A principle components analysis is an example of an unsupervised analysis, where we don't need to specify the groups.

```
> col.trt <- c("purple","orange")[group]
> plotMDS(dgeObj, col=col.trt)
```

Data Exploration: Multidimensional Scaling





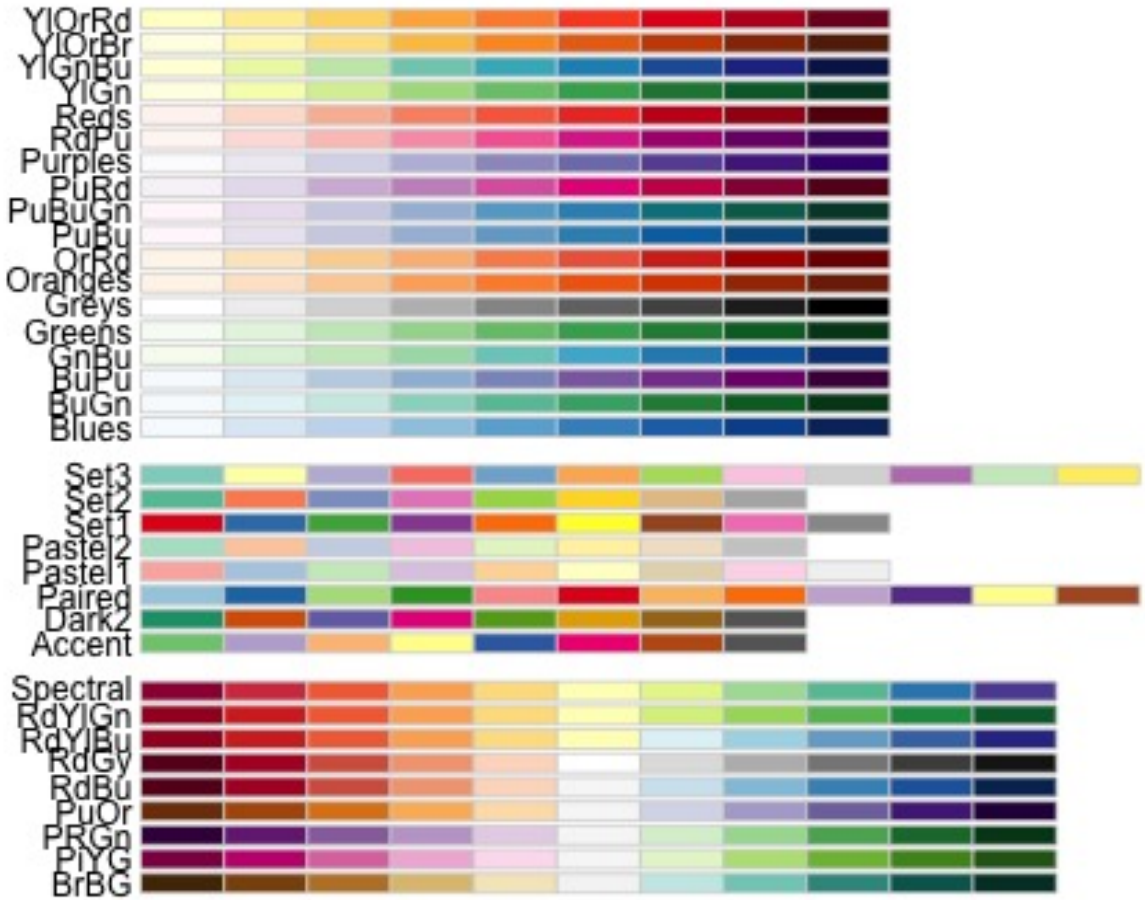
Data Exploration: Hierarchical Clustering with Heatmaps

An alternative to plotMDS for examining relationships between samples is using hierarchical clustering. Heatmaps are a nice visualization to examine hierarchical clustering of the samples. In this example, we select the the 500 most variable genes.

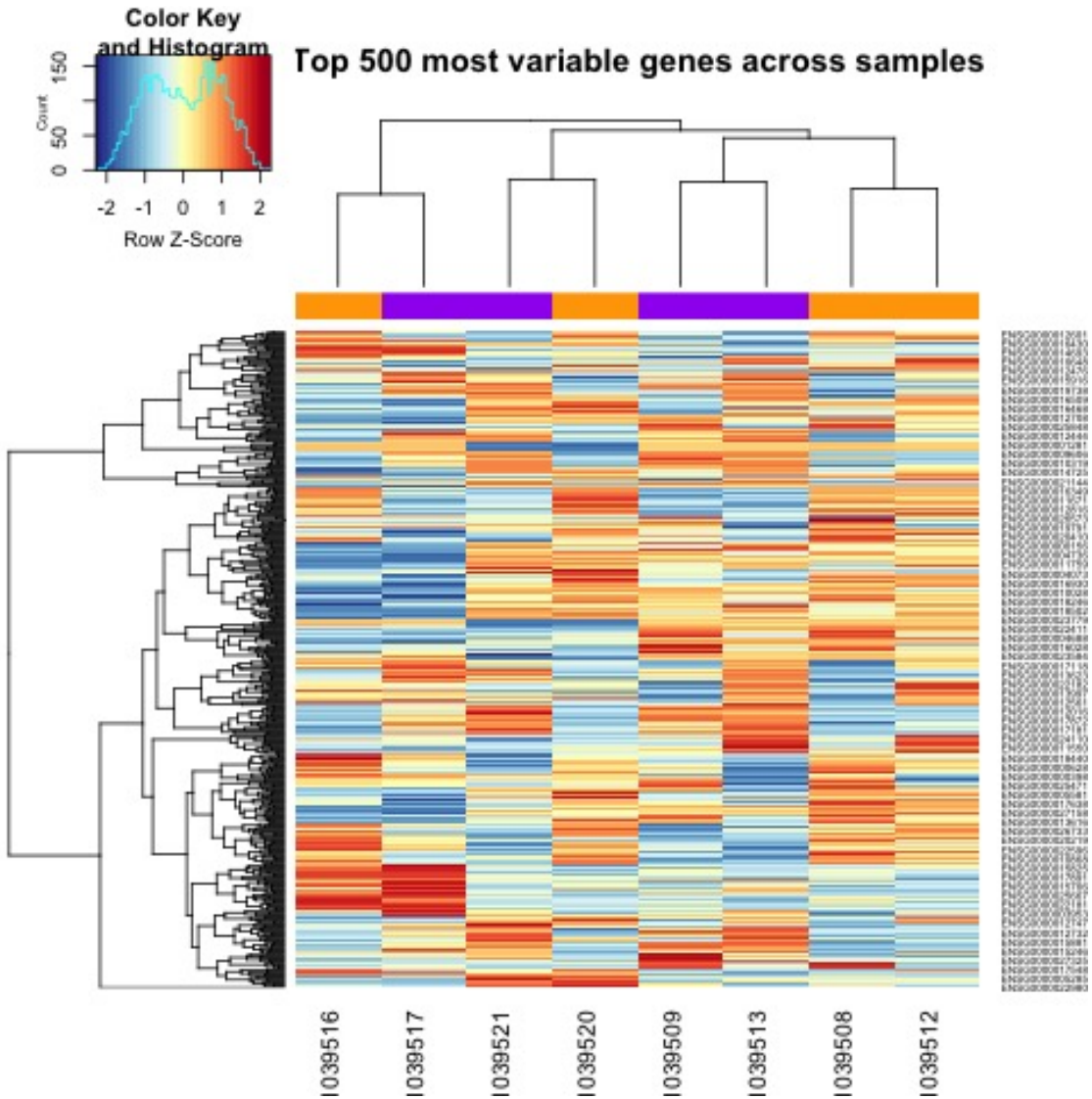
```
> var_genes <- apply(logcounts, 1, var)
> # Get the gene names for the top 500 most variable genes
> select_var <- names(sort(var_genes, decreasing=TRUE))[1:500]
```

Heatmaps: Color Scheme

```
> library(RColorBrewer)
> display.brewer.all()
```



Heatmaps


Color Key
and Histogram
Count
0
50
150
-2
-1
0
1
2
Row Z-Score
Top 500 most variable genes across samples
I039516
I039517
I039521
I039520
I039509
I039513
I039508
I039512