STAT588/BIOL588: Genomic Data Science
Lecture 18: Next Sequencing Data Anlaysis:
(Chapter 5 in Gondro's book)

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

Next Generation Sequencing (NGS) Data Analysis: Preprocess and Differential Expression Analysis

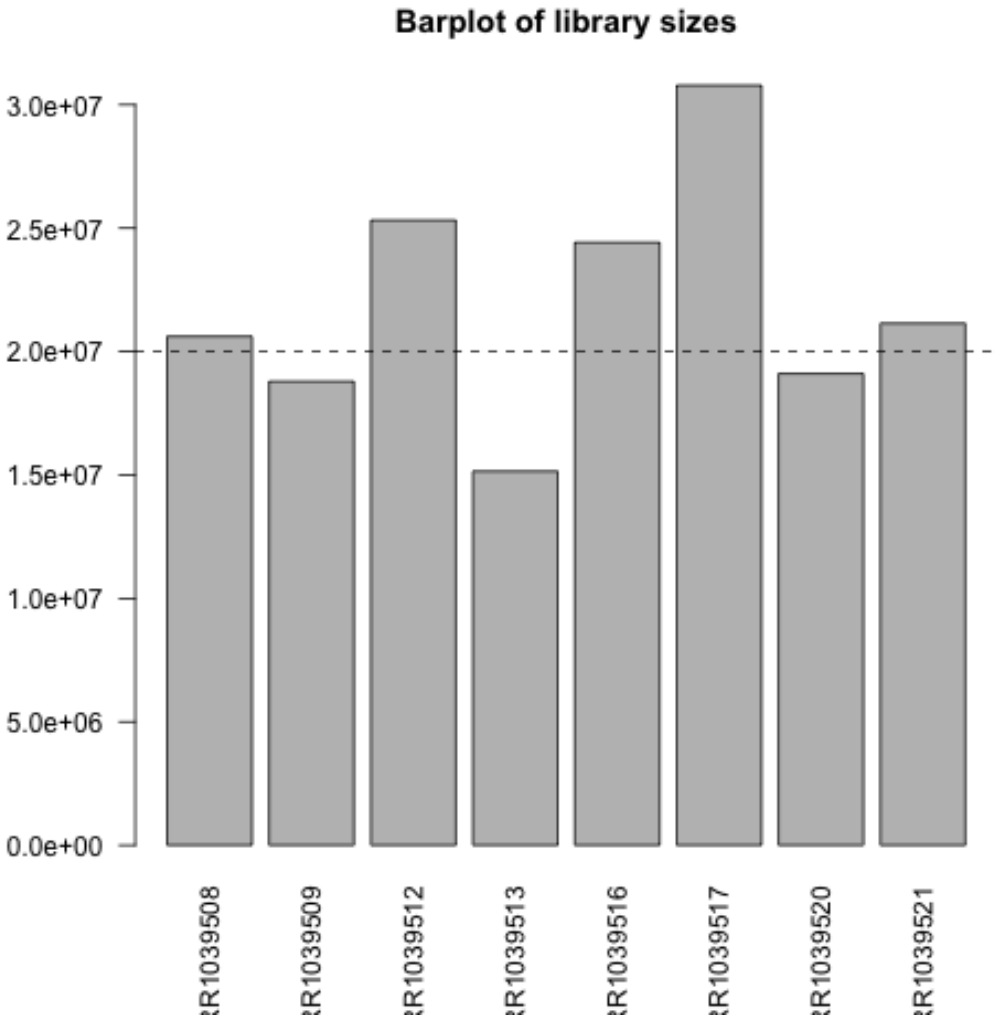
- Read Counts Matrix
- Filtering
- Poisson Regression for Count Data
- Negative Binomial Model
- Experimental Design

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	
...						

Quality Control: Library Sizes

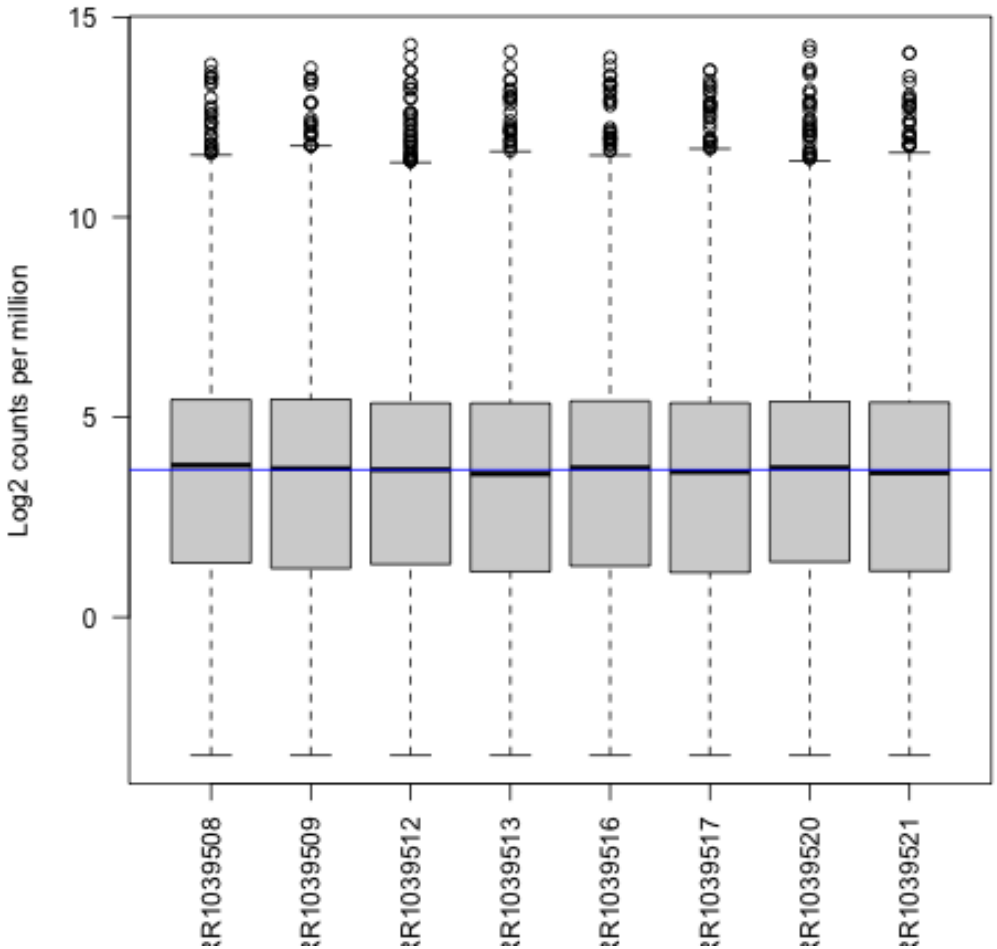


Distribution plots

Count data is not normally distributed, so if we want to examine the distributions of the raw counts we need to log the counts. The "cpm" function to get log2 counts per million, which are corrected for the different library sizes. The cpm function also adds a small offset to avoid taking log of zero.

```
> logcounts <- cpm(dgeObj,log=TRUE)
> boxplot(logcounts, xlab="", ylab="Log2 counts per million",las=2)
> # Let's add a blue horizontal line that corresponds to the median logCPM
> abline(h=median(logcounts), col="blue", main="Boxplots of logCPMs")
```

Quality Control: Distribution plots

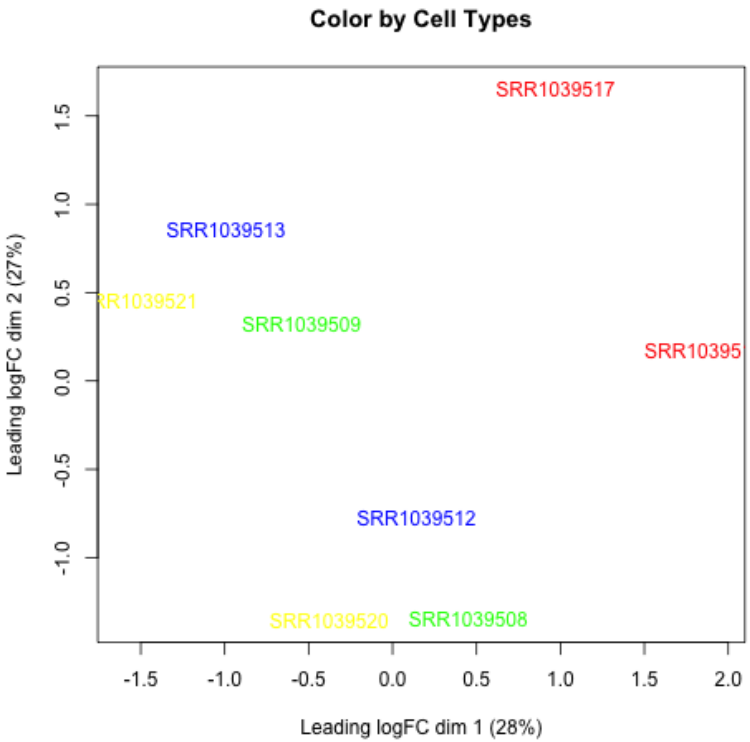


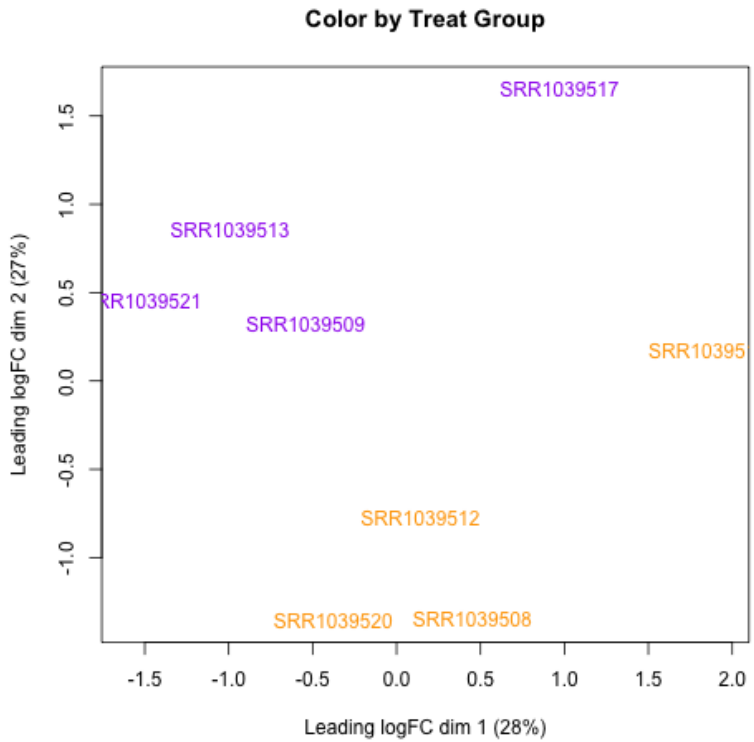
Data Exploration: Multidimensional Scaling

An MDSplot is a visualisation of 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





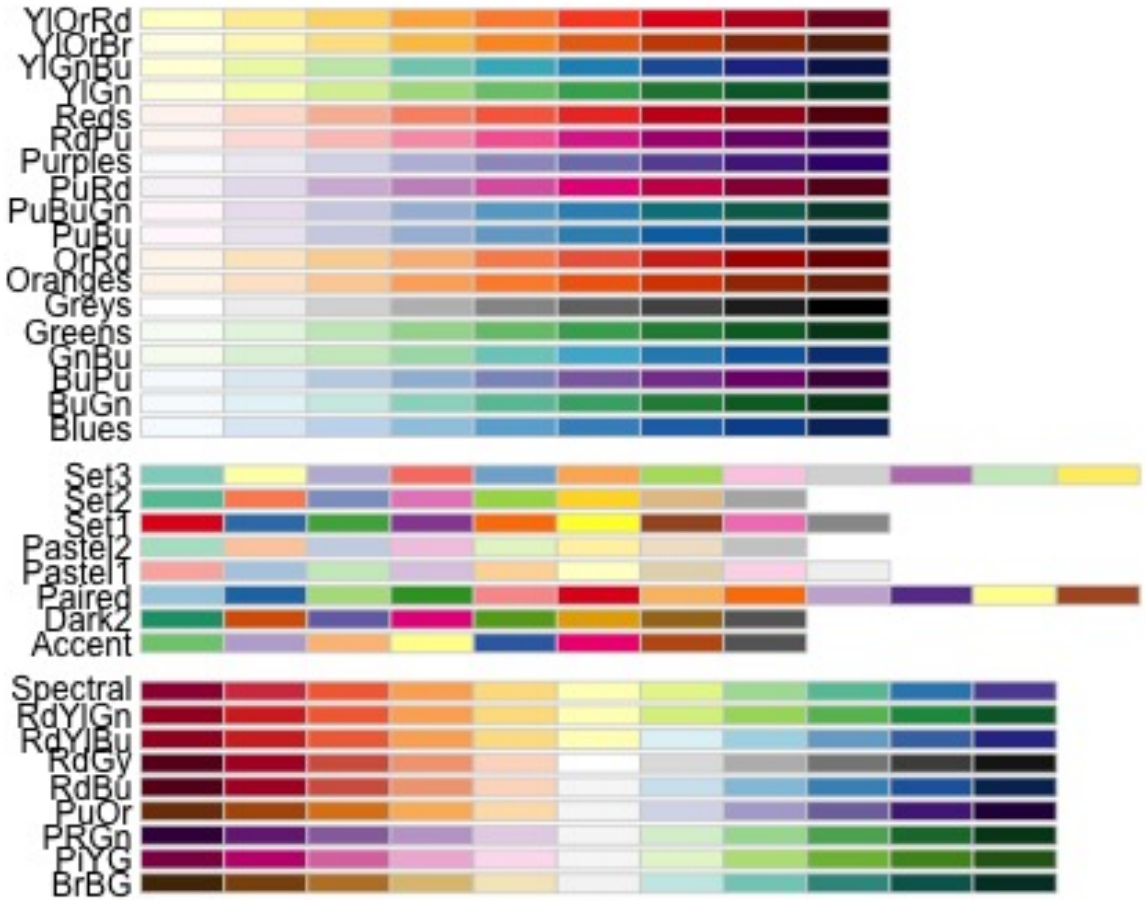
Data Exploration: Hierarchical Clustering with Heatmaps

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.

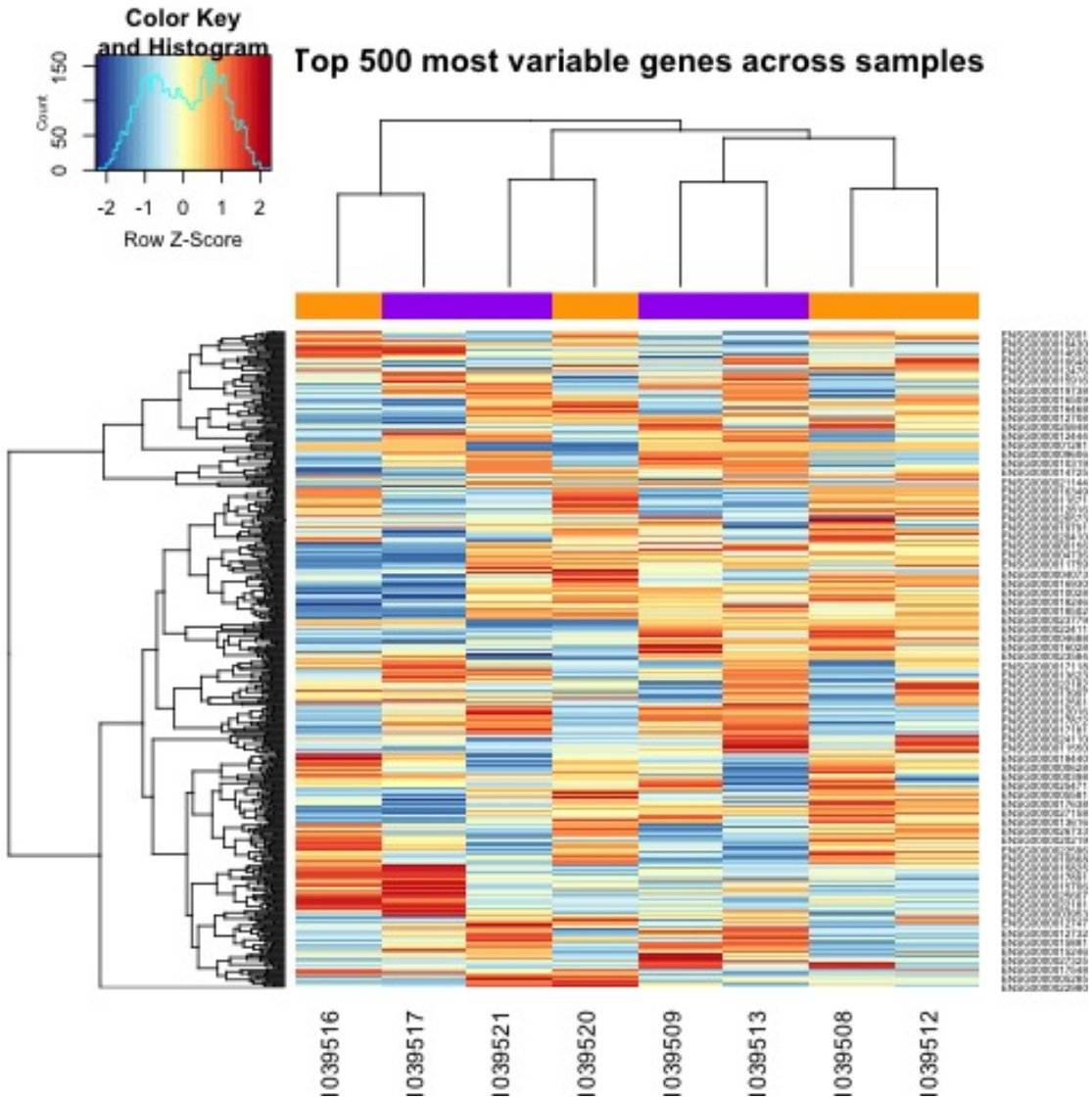
```
> var_genes <- apply(logcounts, 1, var)
> # Get the gene names for the top 500 most variable genes
> select_var <- names(sort(var_genes, decreasing=TRUE))[1:500]
```

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

Differential Expression

- Microarray method are not directly applicable: continuous intensity versus count data, but ideas can be borrowed.
- Sample size: the number of samples in each experimental conditions.
- Test is implemented one-gene-at-a-time.

Poisson Regression for Count Data

The Poisson distribution is commonly used for "count" data (ie: the number of occurrences of some event).
In a Poisson distribution model, the mean is equal to the variance.
However, in RNAseq data the observed variance usually exceeds the mean (overdispersion).
To account for overdispersion, we use the negative binomial distribution (Possion distribution with overdispersion) to model RNAseq data.

- Poisson: $var = \mu$
- NB: $var = \mu(1 + \mu\phi)$
- quasi-likelihood NB: $\sigma_g^2(\mu_{gi} + \phi\mu_{gi}^2)$

Overdispersion in edgeR

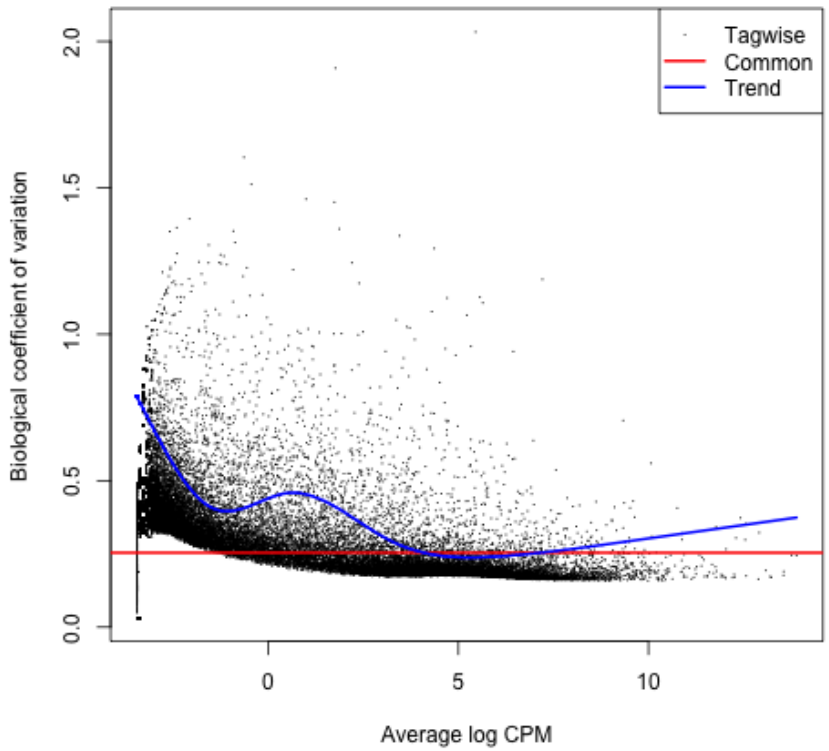
$$var(y_{gi}) = \sigma_g^2(\mu_{gi} + \phi\mu_{gi}^2),$$

where ϕ is the global NB dispersion parameter and σ_g^2 is the gene-specific quasi-likelihood (QL) dispersion parameter.

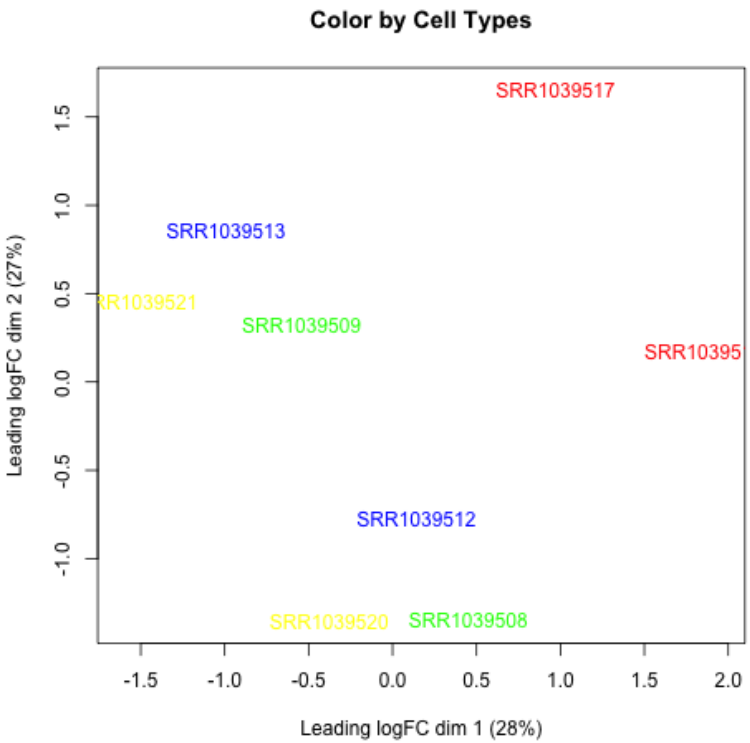
The NB dispersion, ϕ, describes the overall biological variability across all genes. It represents the observed variation that is attributable to inherent variability in the biological system, in contrast to the Poisson variation from sequencing. The QL dispersion picks up any gene-specific variability above and below the overall level.

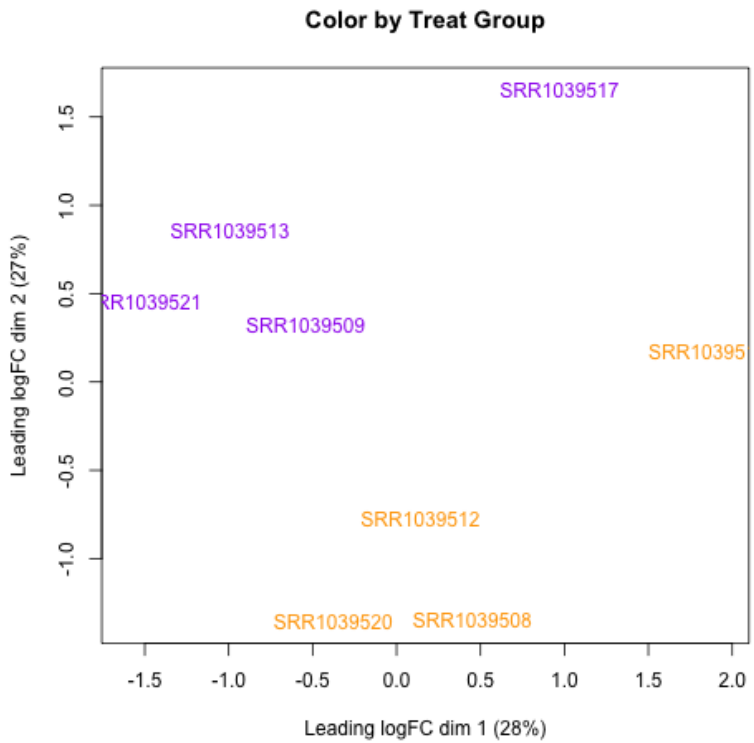
Estimated Dispersion

```
>y<-DGEList(counts=countMat, group=group)
>y<-calcNormFactors(y)
>y
#########stimate dispersion parameters
>y <- estimateCommonDisp(y)
>y <- estimateGLMTrendedDisp(y)
>y<- estimateTagwiseDisp(y)
> plotBCV(y)
```



Data Exploration: Multidimensional Scaling





What is the most important sources of variations in the data? Do you think if the interaction term will be important?

DE & pathway analysis

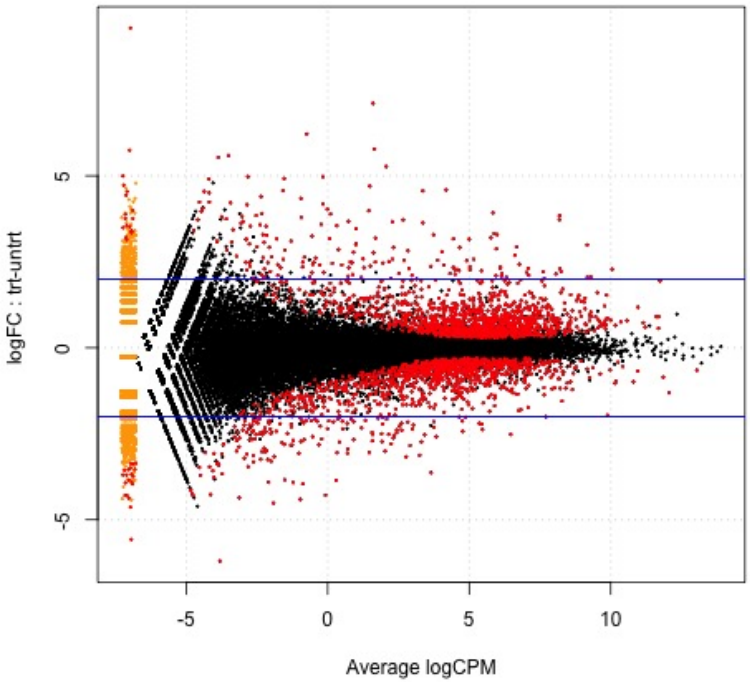
There are three popular methods for DE analysis using RNAseq data: edgeR, DESeq2, and limma. Some example R codes are in Lab18.R

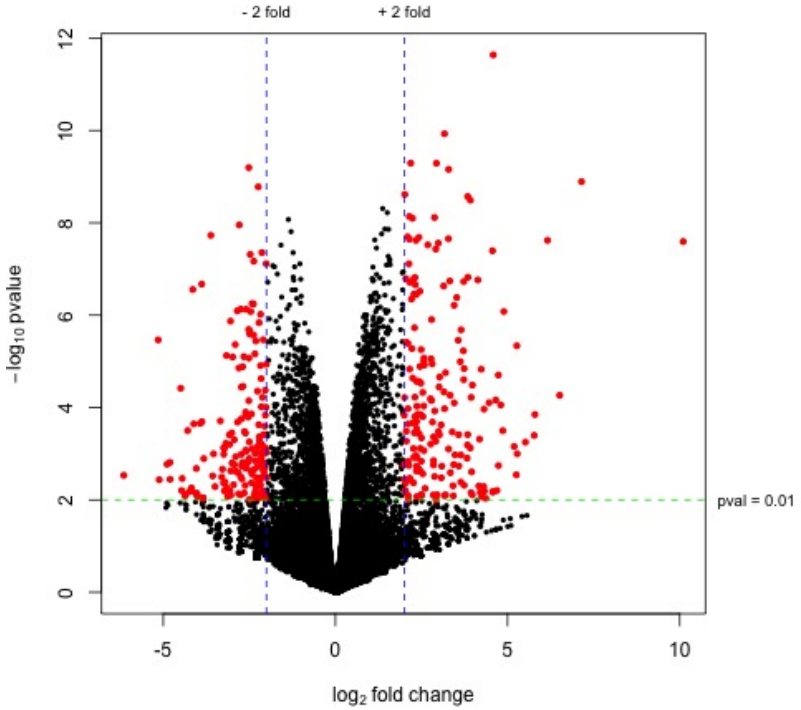
Testing for Differential Expression in edgeR

```
>design<-model.matrix(~group)
>y <- estimateDisp(y, design)
> design <- model.matrix(~group)
> fit <- glmQLFit(y, design)
> qlf.2vs1 <- glmQLFTest(fit, coef=2)
> top<-topTags(qlf.2vs1)
```

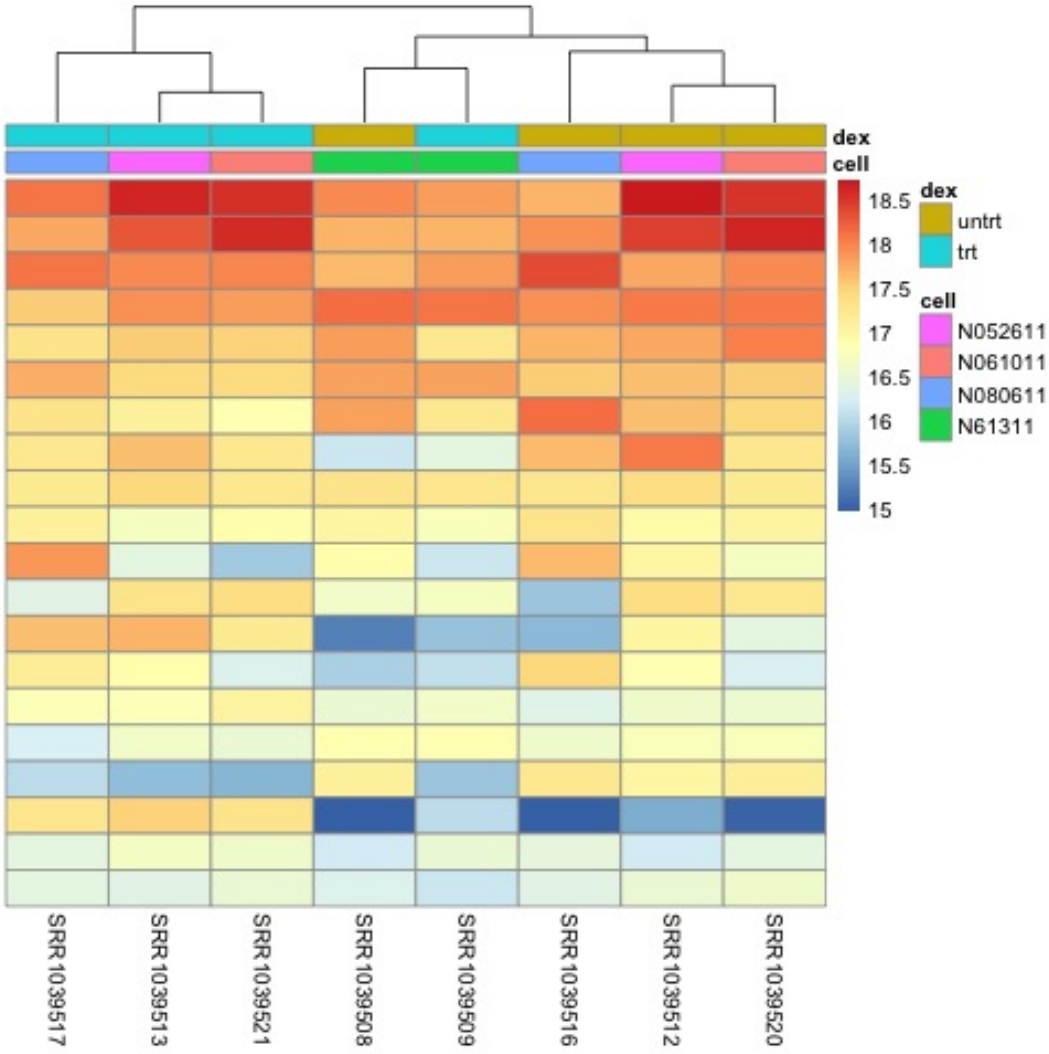
	logFC	logCPM	F	PValue	FDR
ENSG00000152583	4.59	5.54	586.44	0.00	0.00
ENSG00000179094	3.17	5.18	318.34	0.00	0.00
ENSG00000125148	2.19	7.41	252.90	0.00	0.00
ENSG00000120129	2.93	7.31	252.60	0.00	0.00
ENSG00000178695	-2.52	6.96	244.08	0.00	0.00
ENSG00000189221	3.29	6.77	240.77	0.00	0.00
ENSG00000109906	7.15	4.16	218.39	0.00	0.00
ENSG00000196517	-2.24	3.45	209.54	0.00	0.00
ENSG00000162614	2.01	7.97	197.28	0.00	0.00
ENSG00000101347	3.84	9.21	193.98	0.00	0.00

MA and Volcano plot

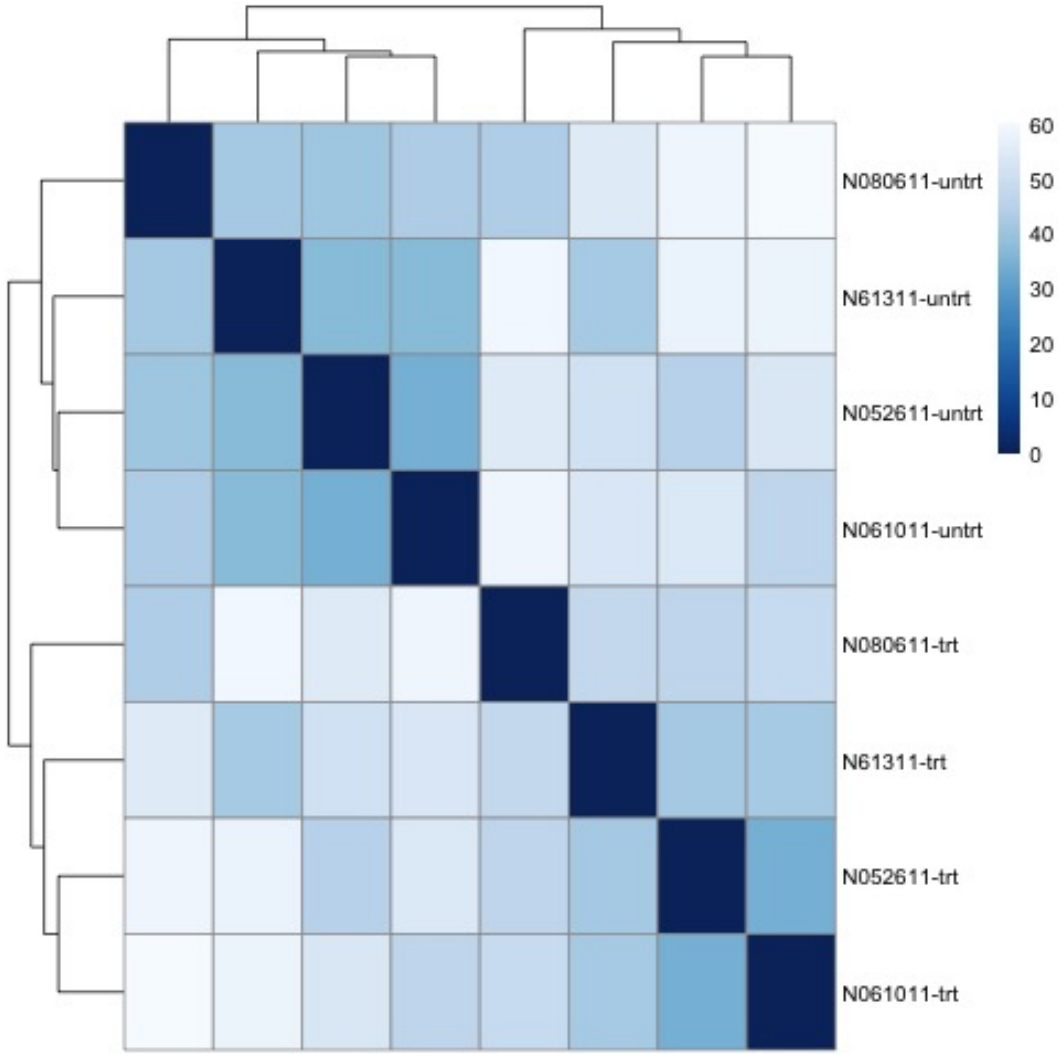




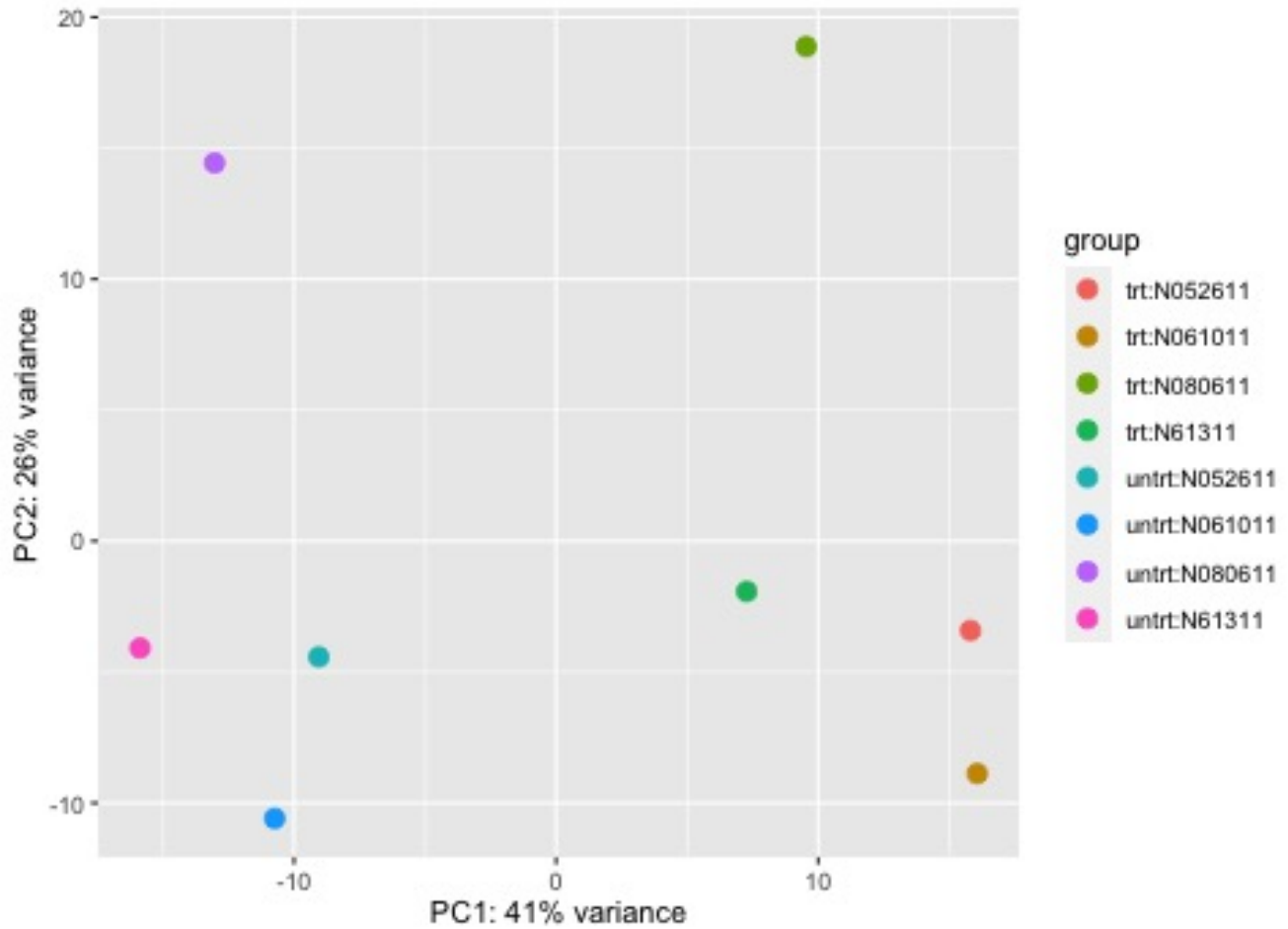
Heatmap for DE genes



Sample Distance



PCA plot


20
10
0
-10
PC2: 26% variance
-10
0
10
PC1: 41% variance
group
trt:N052611
trt:N061011
trt:N080611
trt:N61311
untrt:N052611
untrt:N061011
untrt:N080611
untrt:N61311