STAT718/BIOL703: Genomic Data Science
Gene-level DE Analysis with DESeq2 I

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

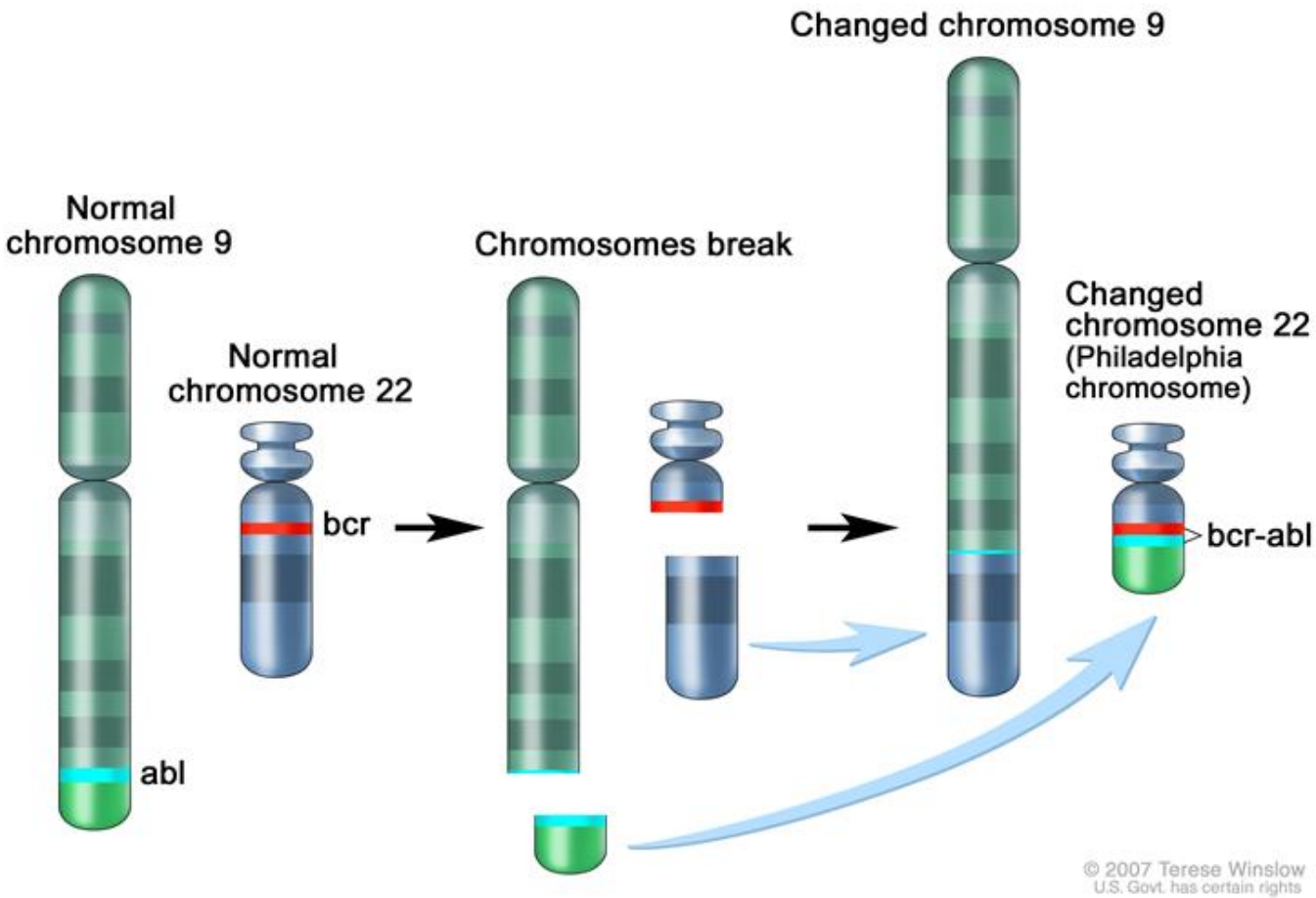
Learning Objectives

- Understanding the different steps in a differential expression analysis in the context of DESeq2
- Executing the differential expression analysis workflow with DESeq2
- Constructing design formulas appropriate for a given experimental design
- Exploring the importance of dispersion during differential expression analysis, and using the plots of the dispersion values to explore assumptions of the NB model

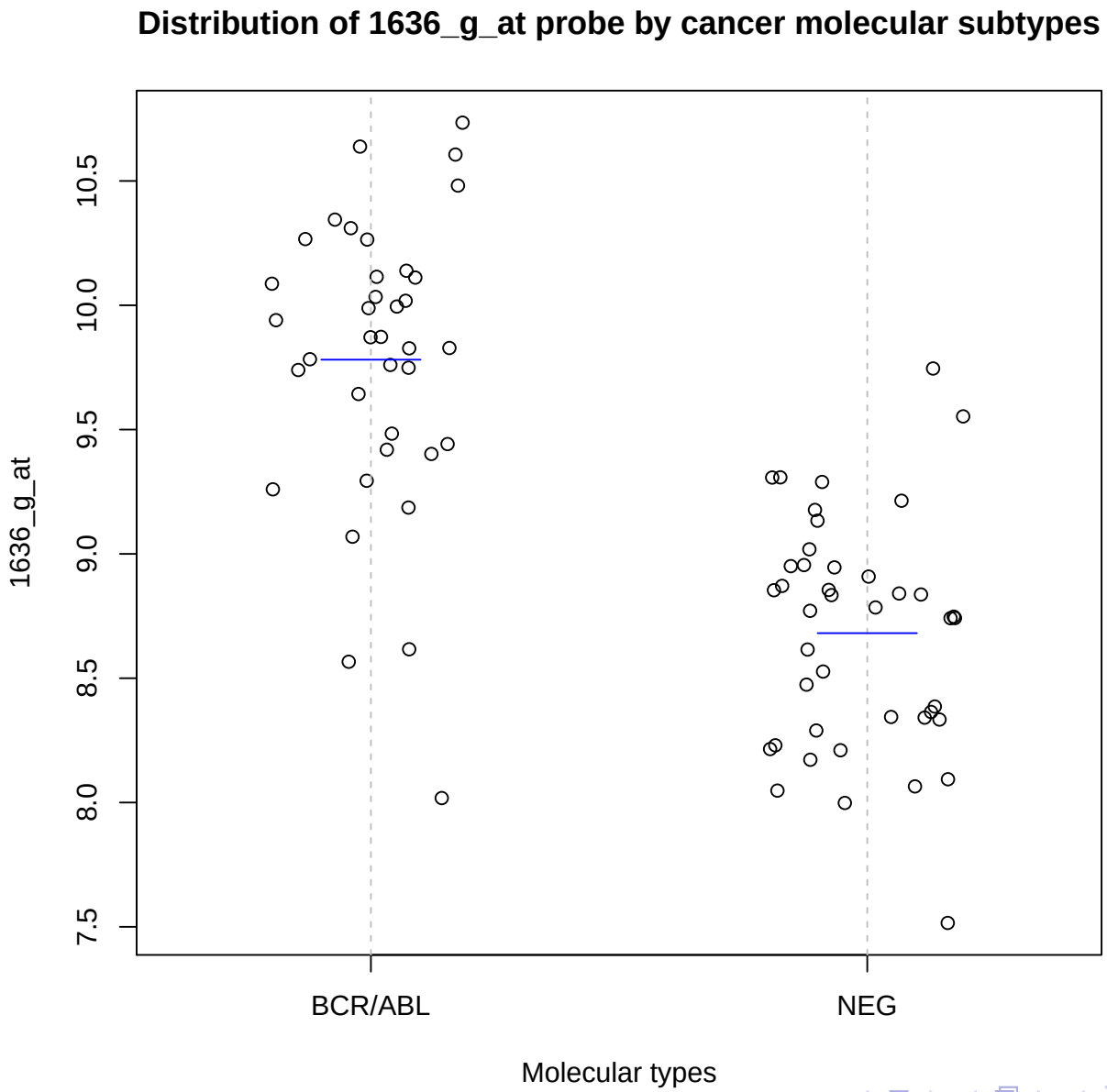
Differential expression analysis with DESeq2

The final step in the differential expression analysis workflow is fitting the raw counts to the Negative Binomial model and performing the statistical test for differentially expressed genes. In this step we essentially want to determine whether the mean expression levels of different sample groups are significantly different.

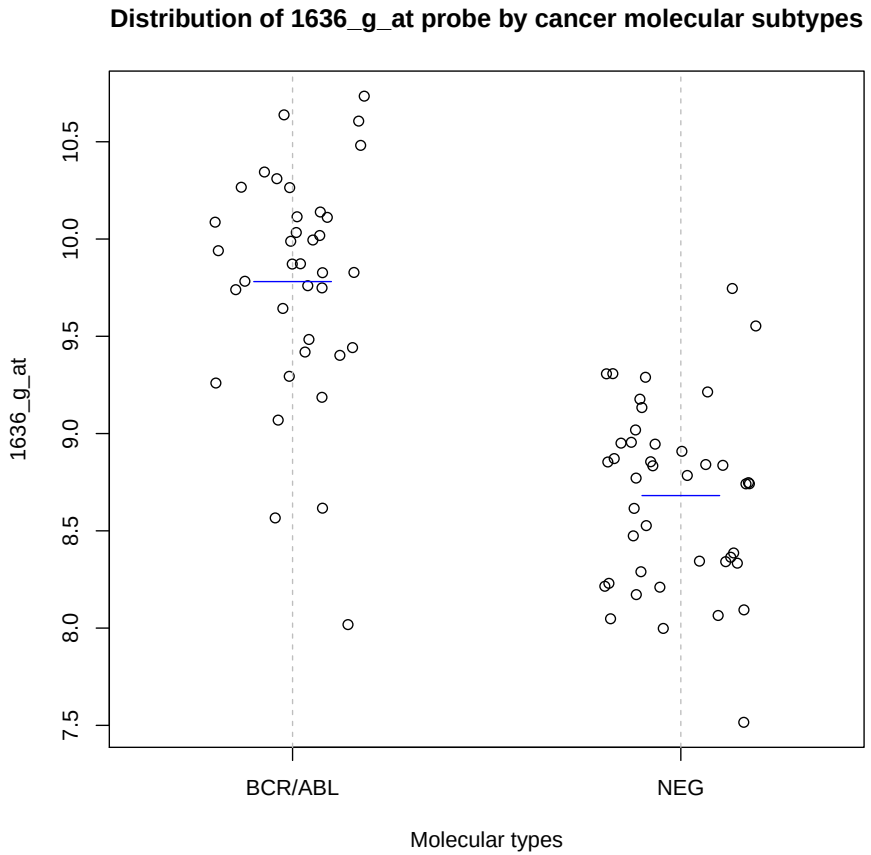
Example: Philadelphia Chromosome



Gene Expression Example (ABL1 gene)

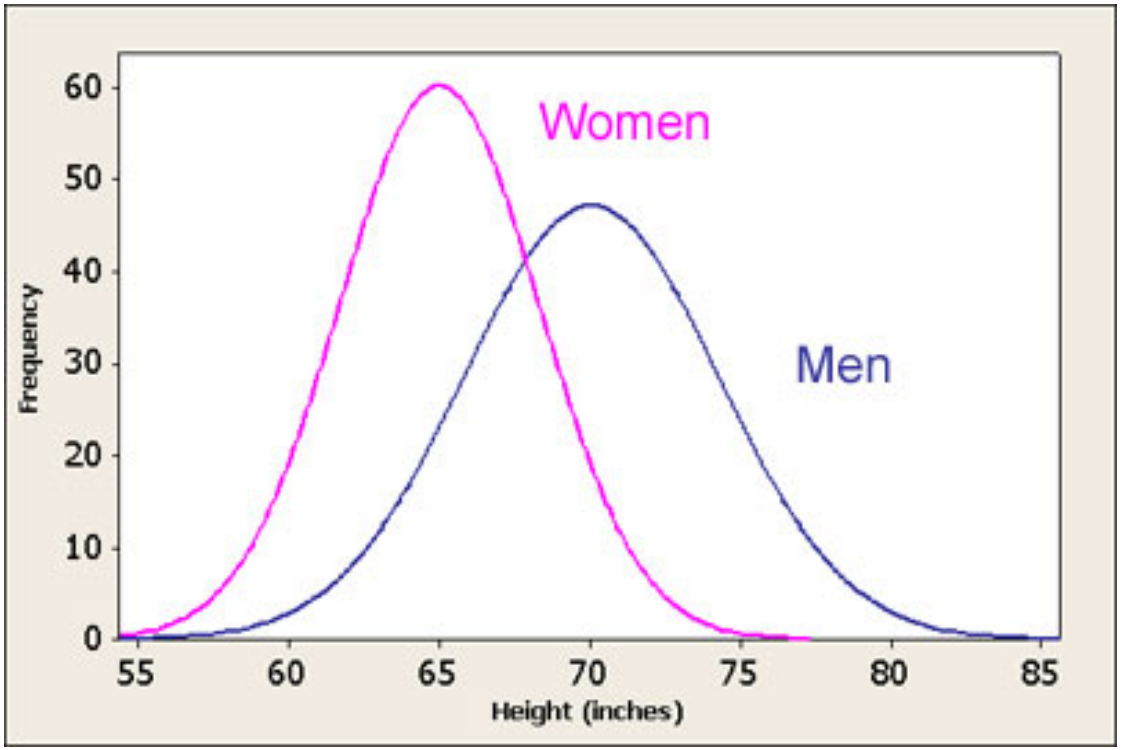


Gene Expression Example (ALL Data)



- Is this difference worth reporting?
- Some journal requires statistical significance. What does it mean?

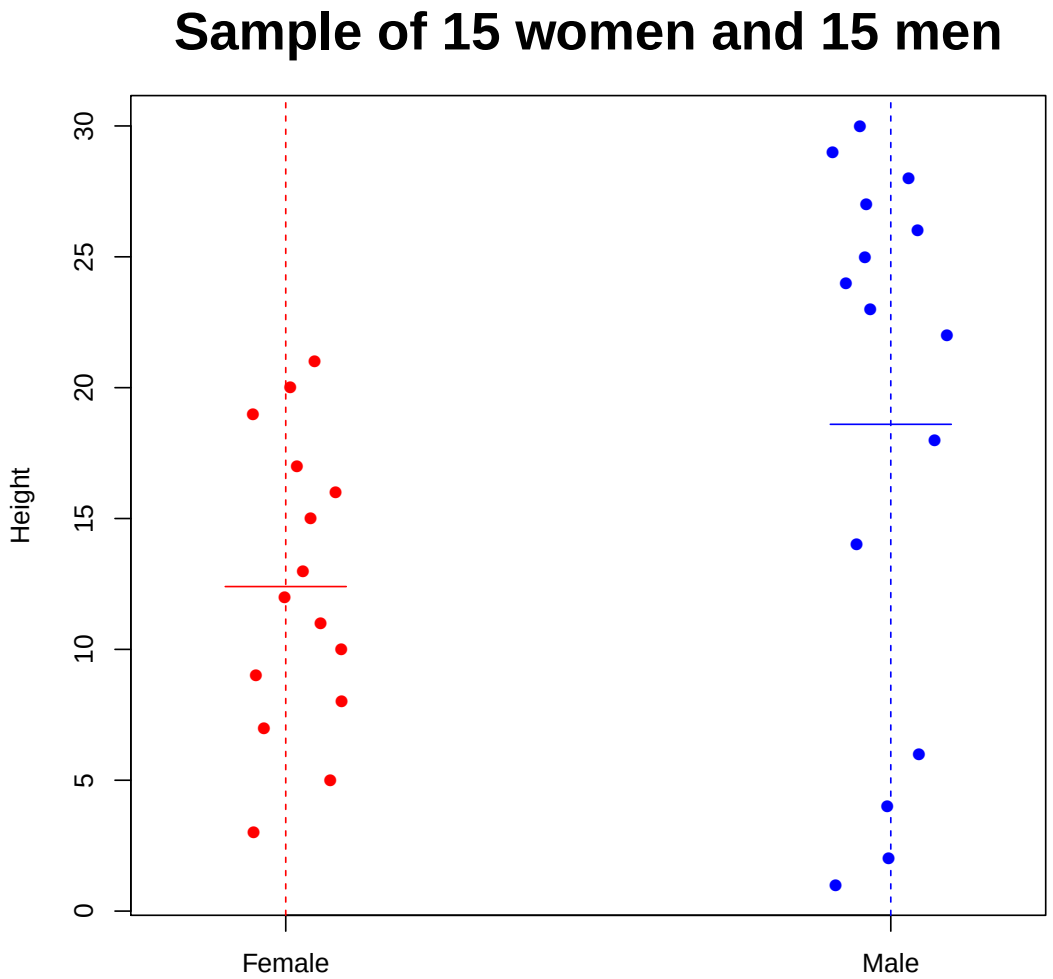
Men are taller than women



This statement refers to population averages: the population average of men's height is larger than the population average of women

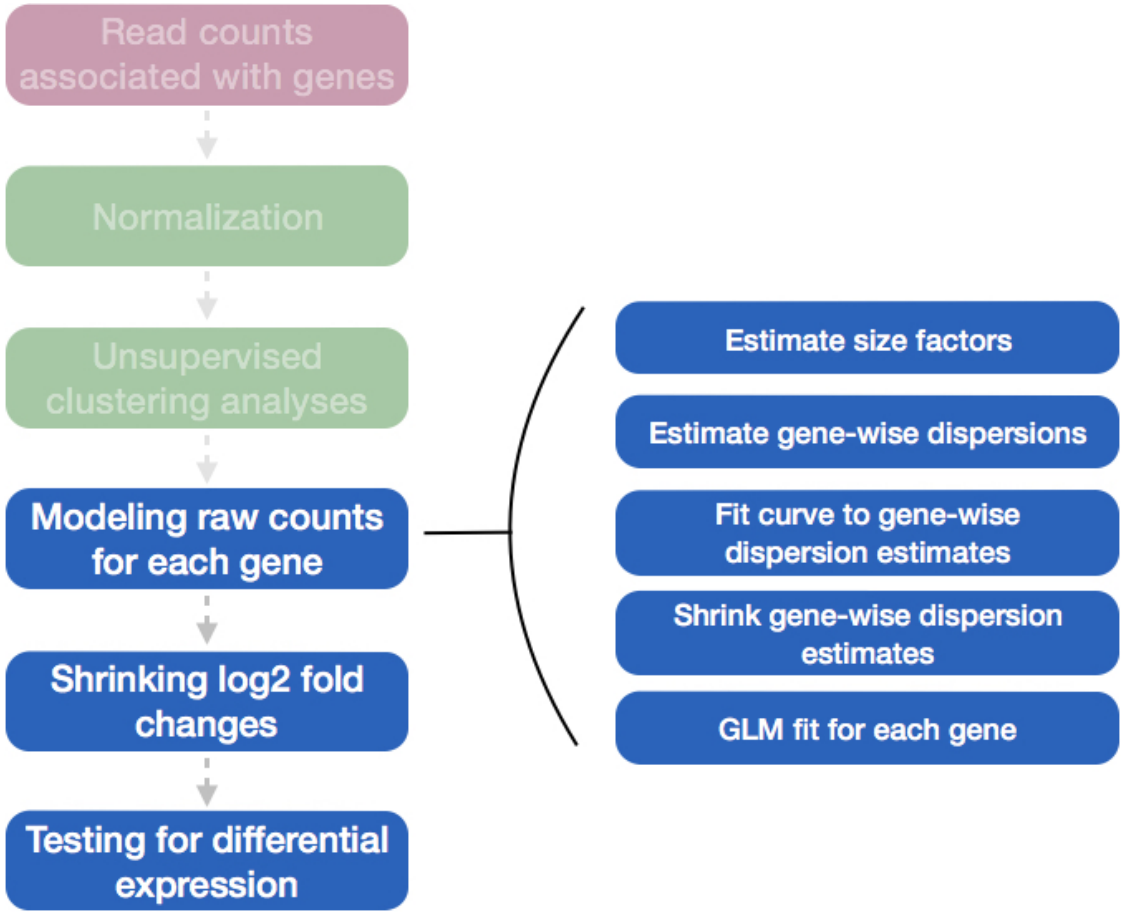
One Data Point

Law of Large Numbers



DE Analysis using DESeq2

The DESeq2 paper was published in 2014, but the package is continually updated and available for use in R through Bioconductor.



Design Formula

Prior to performing the differential expression analysis, it is a good idea to know what sources of variation are present in your data, either by exploration during the QC and/or prior knowledge. Once you know the major sources of variation, you can remove them prior to analysis or control for them in the statistical model by including them in your design formula.

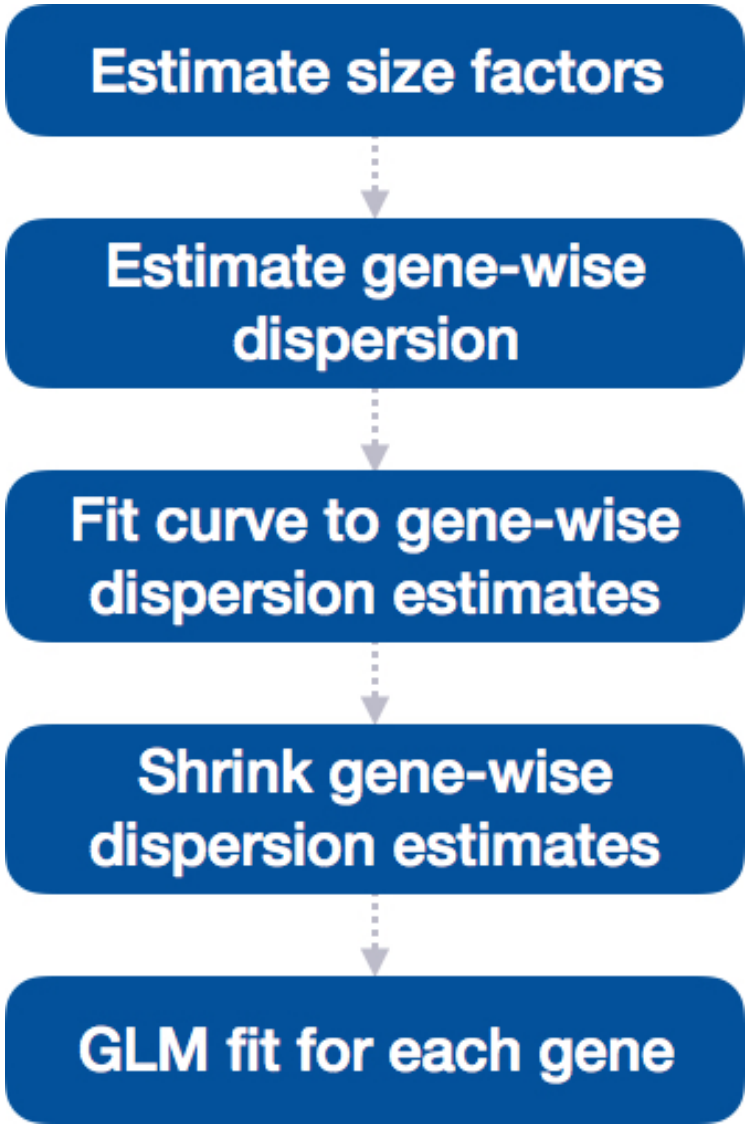
Design Formula

For example, ff you want to examine the expression differences between treatments, and you know that major sources of variation include ‘sex‘ and ‘age‘, then your design formula would be:

```
design <- ~ sex + age + treatment
```

	sex	age	litter	treatment
sample1	M	11	1	Ctrl
sample2	M	13	2	Ctrl
sample3	M	11	1	Treat
sample4	M	13	1	Treat
sample5	F	11	1	Ctrl
sample6	F	13	1	Ctrl
sample7	F	11	1	Treat
sample8	F	13	2	Treat

DESeq DE Analysis Workflow



DESeq DE Analysis

```
## Create DESeq object
dds <- DESeqDataSetFromMatrix(countData = data,
colData = meta, design = ~ sampletype)
## Run analysis
dds <- DESeq(dds)

estimating size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
```

Step 1: Estimate size factors

The first step in the differential expression analysis is to estimate the size factors. To normalize the count data, DESeq2 calculates size factors for each sample using the median of ratios method discussed previously.

```
## Check the size factors
sizeFactors(dds)
```

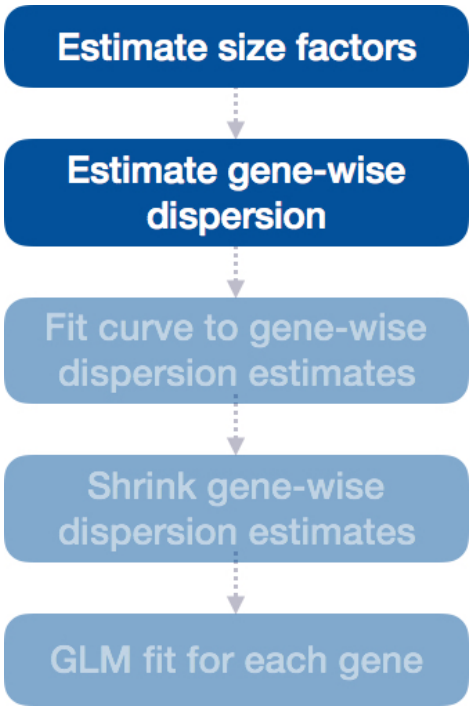
```
Mov10_kd_2 Mov10_kd_3 Mov10_oe_1 Mov10_oe_2 Mov10_oe_3 ...
 1.5646728  0.9351760  1.2016082  1.1205912  0.6534987 ...
```

How do the total number of reads for each sample correlate with the size factor?

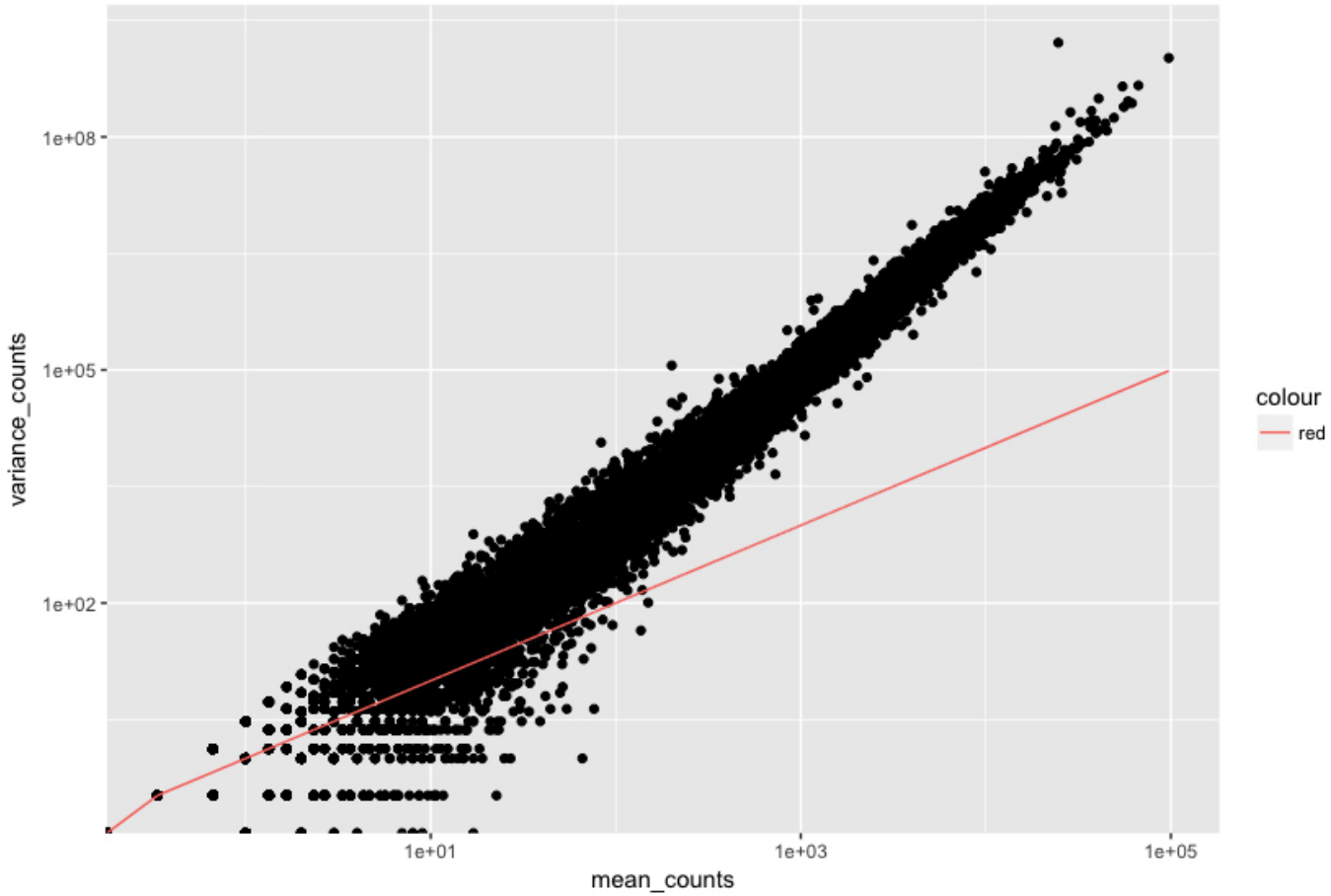
What is dispersion?

Dispersion is a measure of spread or variability in the data. Variance, standard deviation, IQR, among other measures, can all be used to measure dispersion. However, DESeq2 uses a specific measure of dispersion (α) related to the mean (μ) and variance of the data:

$$\sigma^2 = \mu(1 + \alpha\mu).$$



Disperson:Variance/Mean



How does the dispersion relate to our model?

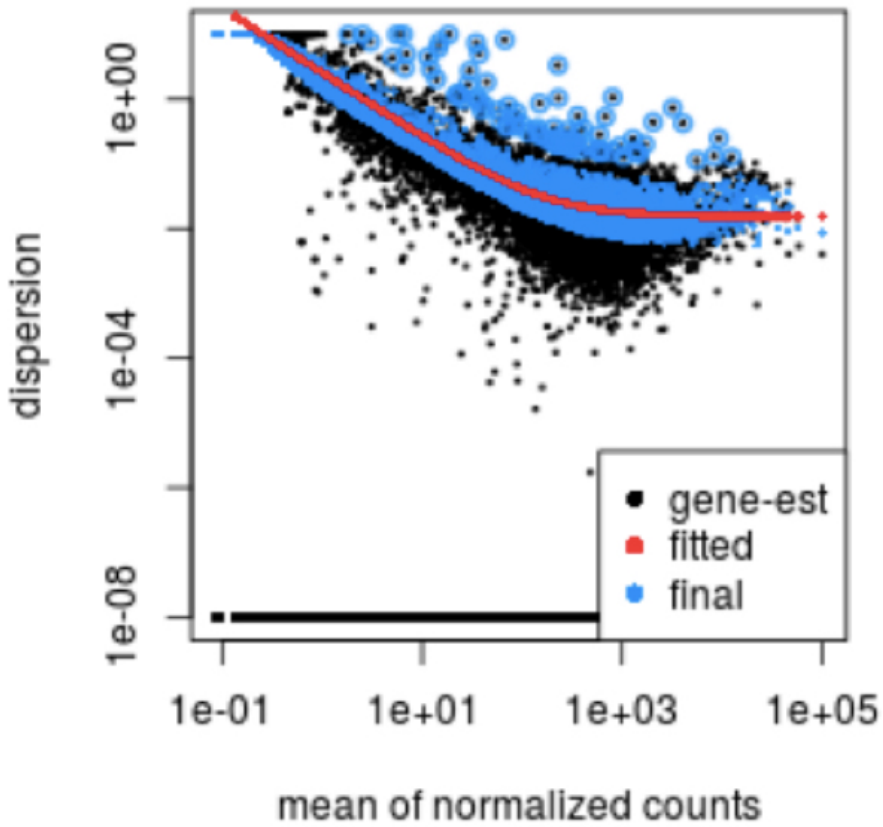
With only a few (3-6) replicates per group, the estimates of variation for each gene are often unreliable (due to the large differences in dispersion for genes with similar means).

To address this problem, DESeq2 shares information across genes to generate more accurate estimates of variation based on the mean expression level of the gene using a method called shrinkage. DESeq2 assumes that genes with similar expression levels have similar dispersion.

To achieve this, DESeq2 first estimates the dispersion for each gene separately using maximum likelihood estimate (MLE).

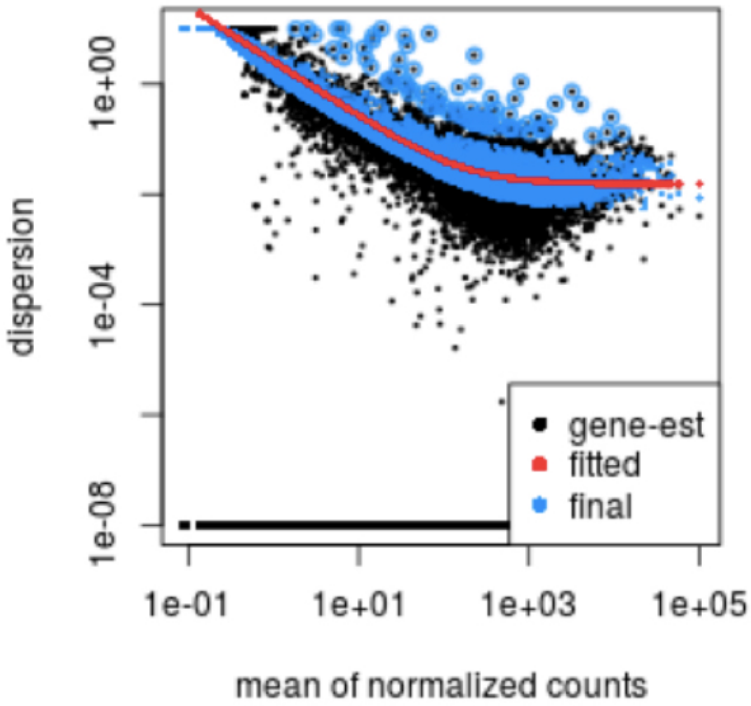
Step 3: Fit curve to gene-wise dispersion estimates

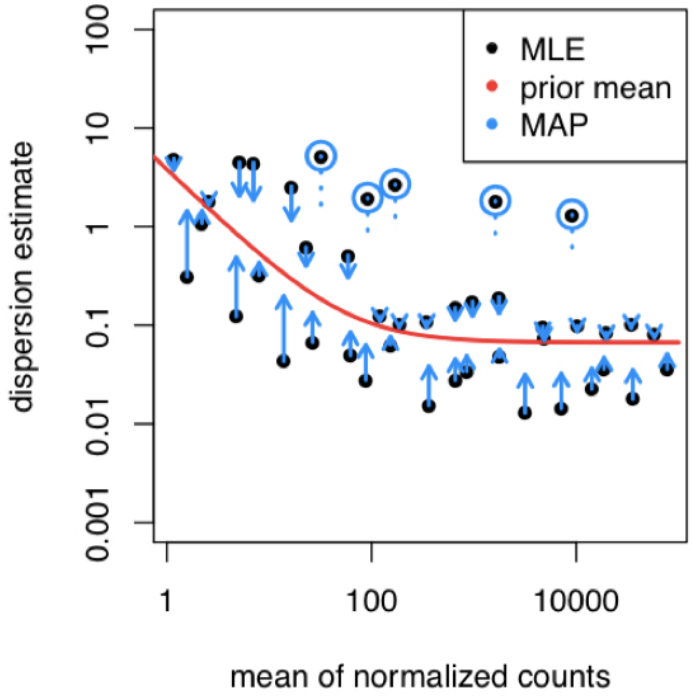
This curve is displayed as a red line in the figure below, which plots the estimate for the expected dispersion value for genes of a given expression strength. Each black dot is a gene with an associated mean expression level and MLE of the dispersion



Step 4: Shrinkage Estimates

Dispersion estimates that are slightly above the curve are also shrunk toward the curve for better dispersion estimation; however, genes with extremely high dispersion values are not. This is due to the likelihood that the gene does not follow the modeling assumptions and has higher variability than others for biological or technical reasons.





Shrinkage Estimates

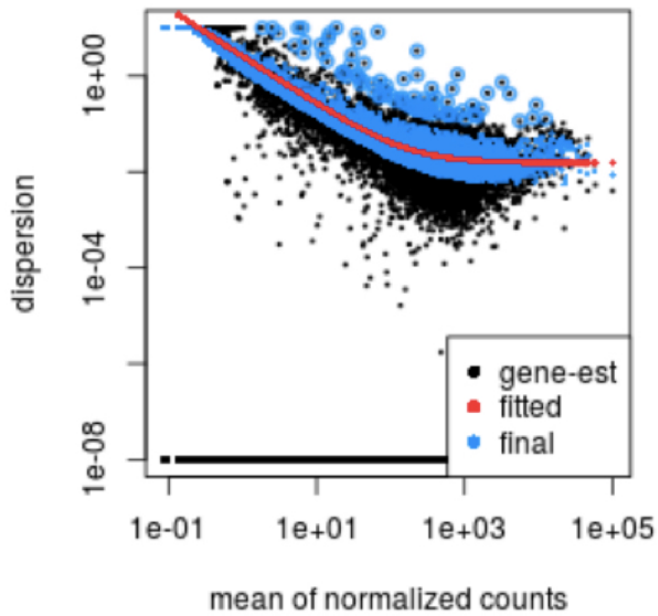
The amount of the shrinkage for each gene depends on :

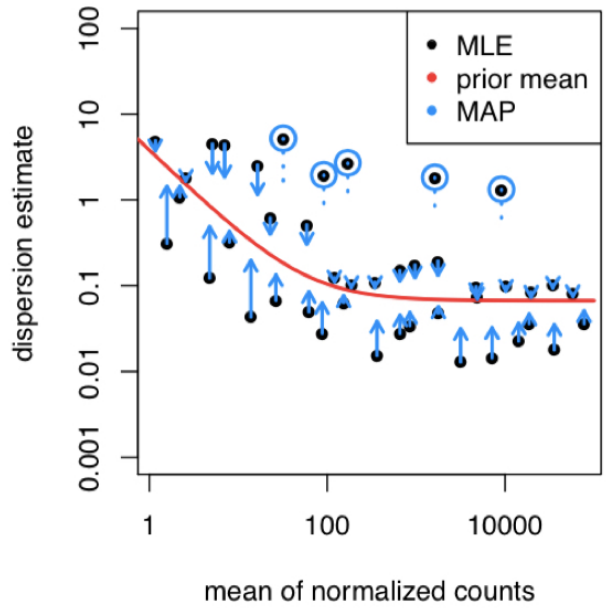
- how close gene dispersions are from the curve
- sample size (more samples = less shrinkage)

Shrinkage Estimates

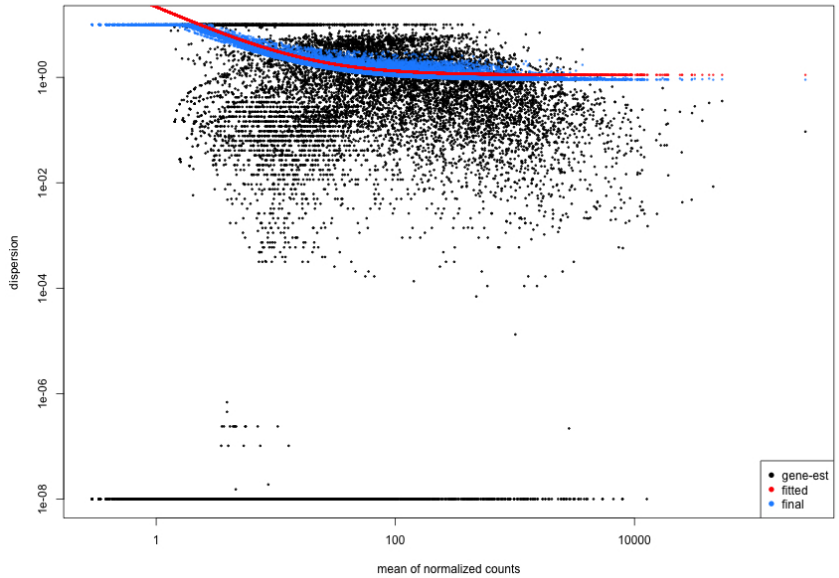
This is a good plot to examine to ensure your data is a good fit for the DESeq2 model.

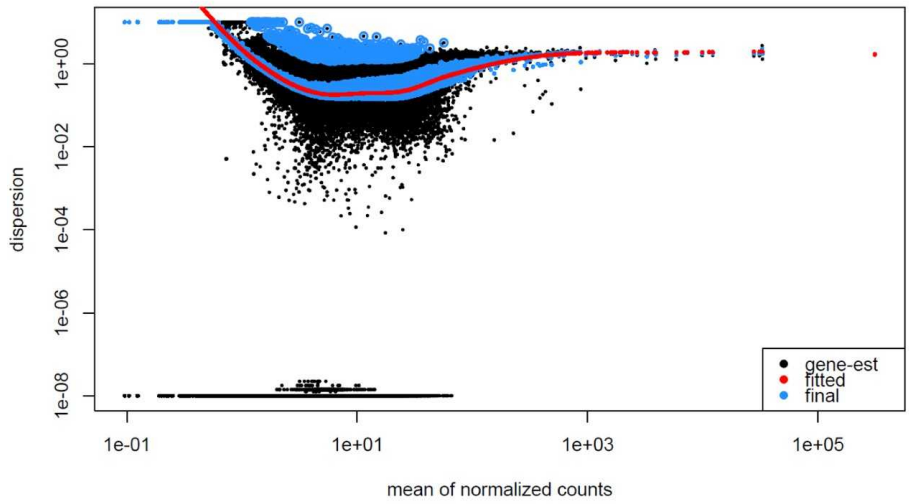
You expect your data to generally scatter around the curve, with the dispersion decreasing with increasing mean expression levels. If you see a cloud or different shapes, then you might want to explore your data more to see if you have contamination (mitochondrial, etc.) or outlier samples.





Worrisome Dispersion Plots





MOV10 Dataset

```
## Plot dispersion estimates
plotDispEsts(dds)
```