# Statistics of RNA-seq differential expression testing

2012 CGRL workshop

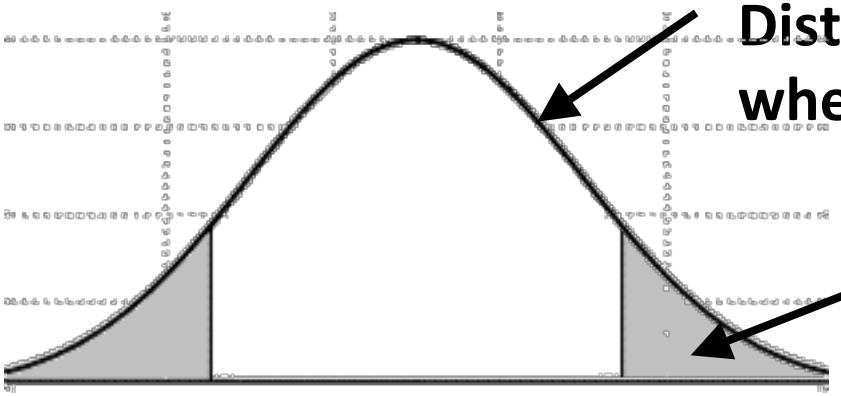
Oh Kyu Yoon

# You have a list of counts. Now what?

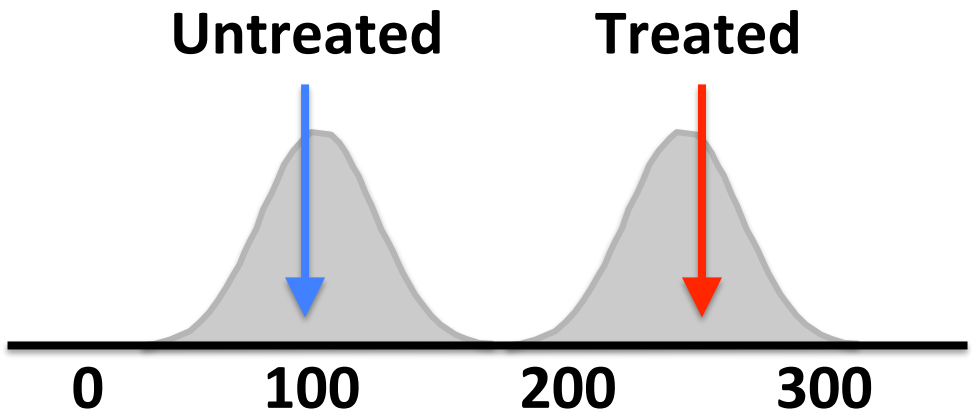
| Gene | Untreated | Treated |
|---|---|---|
| Gene 1 | 100 | 250 |
| Gene 2 | 10 | 20 |
| Gene 3 | 5000 | 4500 |
| : | : | : |
| : | : | : |

# Differential expression asks: Is difference in expression bigger than dispersion?

$$\textbf{DE statistic T} = \frac{\textbf{Expression1} - \textbf{Expression2}}{\textbf{Standard Error}}$$



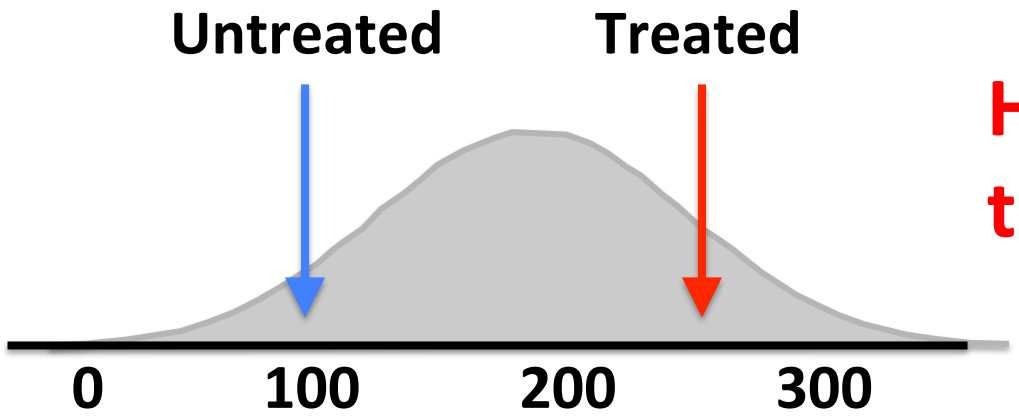
# Differential expression asks: Is difference in expression bigger than dispersion?




$$\text{DE statistic } T = \frac{250 - 100}{50} = 3$$

Look up in table → P-value 0.0013

# Differential expression asks: Is difference in expression bigger than dispersion?

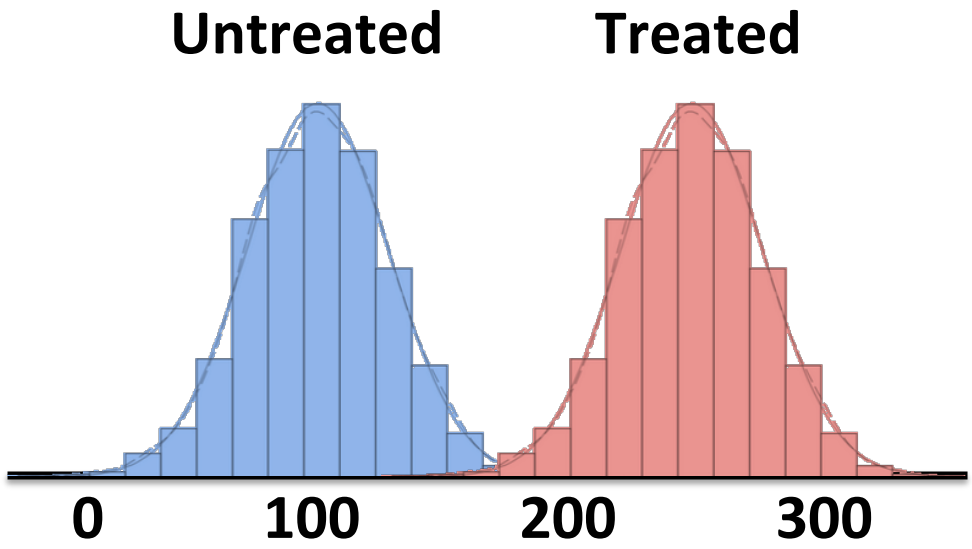



**How do you know the dispersion?**

$$\text{DE statistic } T = \frac{250 - 100}{180} = 0.83$$

Look up in table → P-value 0.2

# You need replicates to accurately estimate dispersion

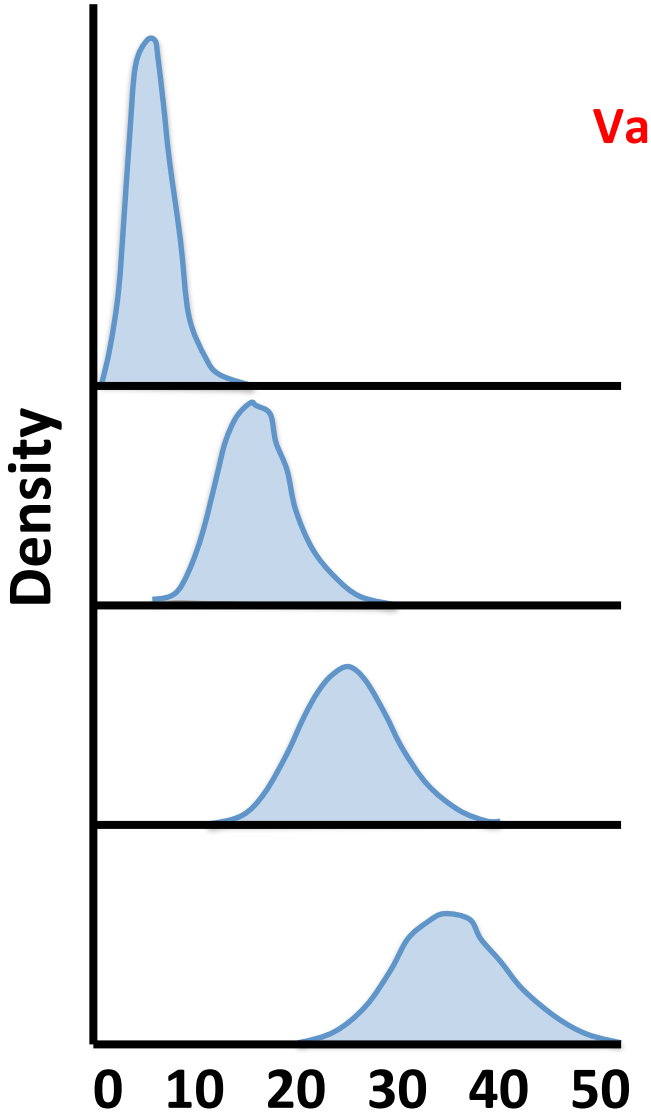



**It is unrealistic to have more than a few RNA-seq replicates**

**You need to make some assumptions about dispersion**

# RNA-seq data are counts

## ➡ Counts have Poisson distribution




**Variance = Mean**

**Mean=5**
**Variance=5**

**Mean=15**
**Variance=15**

**Mean=25**
**Variance=25**

**Mean=35**
**Variance=35**

**Need to use**
**a statistical test**
**designed for count data**

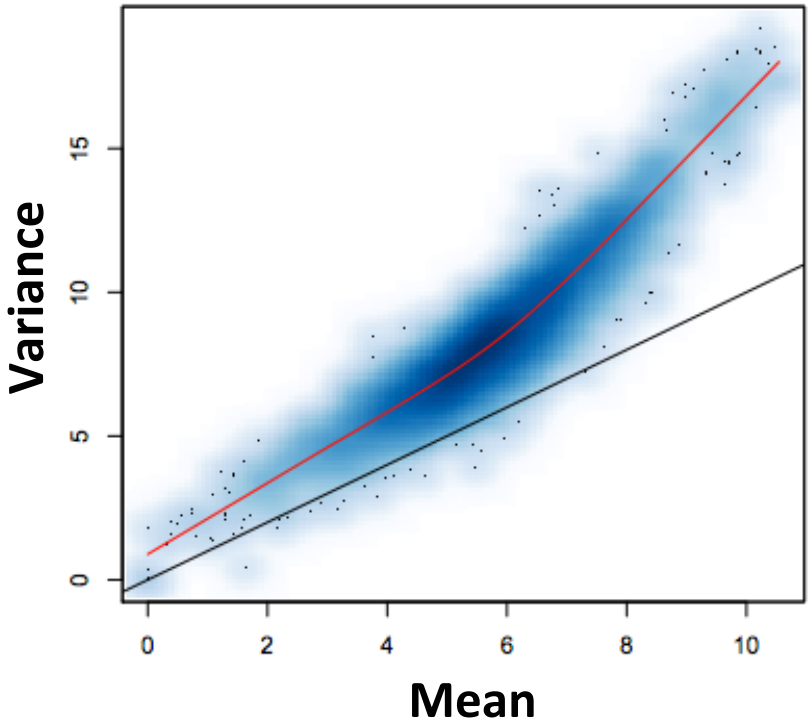
# Fisher's exact test

- Very easy to use
- Used with 2x2 contingency table
- Based on hypergeometric distribution

| | Untreated | Treated | Total |
|---|---|---|---|
| Gene 1 | 100 | 250 | 350 |
| Other genes | 9,999,900 | 12,999,750 | 24,999,650 |
| Total | 10M | 13M | 25M |

```
> cont.table <- matrix(c(100,250,9999900,12999750),nrow=2,byrow=T)
> output <- fisher.test(cont.table)
> pvalue <- output$p.value
```

# There are additional variations that make variance larger than the mean




Source of variation:
- Biological variation
- Technical variation from library prep
- GC bias, transcript length bias
- Flowcell effect

etc

**Generalized linear model (GLM)**

**allows incorporation of known additional variations**

**Negative binomial**

**models unexplained variance as**

**Variance = Mean + $\phi$ Mean$^2$**

# Generalized Linear Model (GLM)

- Linear regression that allows distributions such as Poisson
- Can incorporate replicates and other variables

| Gene | Untreated | | | | Treated | | | |
|---|---|---|---|---|---|---|---|---|
| | Lib Prep 1 | | Lib Prep 2 | | Lib Prep 1 | | Lib Prep 2 | |
| | FC1 | FC2 | FC1 | FC2 | FC1 | FC2 | FC1 | FC2 |
| Gene 1 | 95 | 105 | 110 | 83 | 313 | 301 | 325 | 295 |
| Gene 2 | 10 | 7 | 12 | 5 | 19 | 18 | 24 | 20 |
| Gene 3 | 4930 | 4990 | 5050 | 4850 | 4549 | 4529 | 4869 | 4497 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Total | 10M | 11M | 11M | 8M | 10M | 9M | 12M | 10M |

**log(Counts) ~ Treatment + Lib_Prep + Flowcell**

# log(Counts) ~ log(Total) + Treatment + Lib_Prep + Flowcell

**Design matrix**

| Treatment | Lib_Prep | Flowcell | Count | Total reads |
|---|---|---|---|---|
| 1 | 1 | 1 | 95 | 10 |
| 1 | 1 | 2 | 105 | 11 |
| 1 | 2 | 1 | 110 | 11 |
| 1 | 2 | 2 | 83 | 8 |
| 2 | 1 | 1 | 313 | 10 |
| 2 | 1 | 2 | 301 | 9 |
| 2 | 2 | 1 | 325 | 12 |
| 2 | 2 | 2 | 295 | 10 |

```
> counts <- c(95,105,110,83,313,301,325,295)
> treatment <- c(1,1,1,1,2,2,2,2)
> lib_prep <- c(1,1,2,2,1,1,2,2)
> flowcell <- c(1,2,1,2,1,2,1,2)
> norm.factor <- c(10,11,11,8,10,9,12,10)
> glm.gene1 <- glm(counts ~ treatment + lib_prep + flowcell,
                   family=poisson(),offset=log(norm.factor))
> summary(glm.gene1)
```

# Generalized Linear Model (GLM)

```
> summary(glm.gene1)
> pvalues <- summary(glm.gene1)$coefficients[,4]
```

```
Deviance Residuals:
      1         2         3         4         5         6         7         8
-0.4398   -1.0361    0.9558    0.6298    0.3866    0.4969   -0.6663   -0.1838

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.19709    0.14748   8.117 4.77e-16 ***
treatment    1.12561    0.05801  19.404  < 2e-16 ***
lib_prep    -0.08603    0.04969  -1.731   0.0834 .
flowcell     0.05942    0.04966   1.197   0.2315
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 444.0285  on 7  degrees of freedom
Residual deviance:   3.4512  on 4  degrees of freedom
AIC: 67.415
```

# Negative Binomial

- **NB is similar to Poisson, but Variance = Mean + $\phi$ Mean$^2$**
- **This can capture most of unexplained variability**
- **The difficult step is calculating the dispersion parameter $\phi$**

**EdgeR and DESeq will calculate the dispersion parameter**

# Don’t forget multiple testing correction

```
> library(multtest)
> mt.rawp2adjp(rawp,c(“Bonferroni”,”BH”))
```