Differential expression statistics

Recap: transcript count bias

- Comes from:
  - Random sampling error
  - PCR bias
- Results in:
  - Highly abundant transcripts are oversampled
  - Lowly abundant transcripts are undersampled
- Dynamic range typically 10^5
- Can't have zero counts (no sequence!)

RNAseq data presents unique statistical challenges

Statistical models

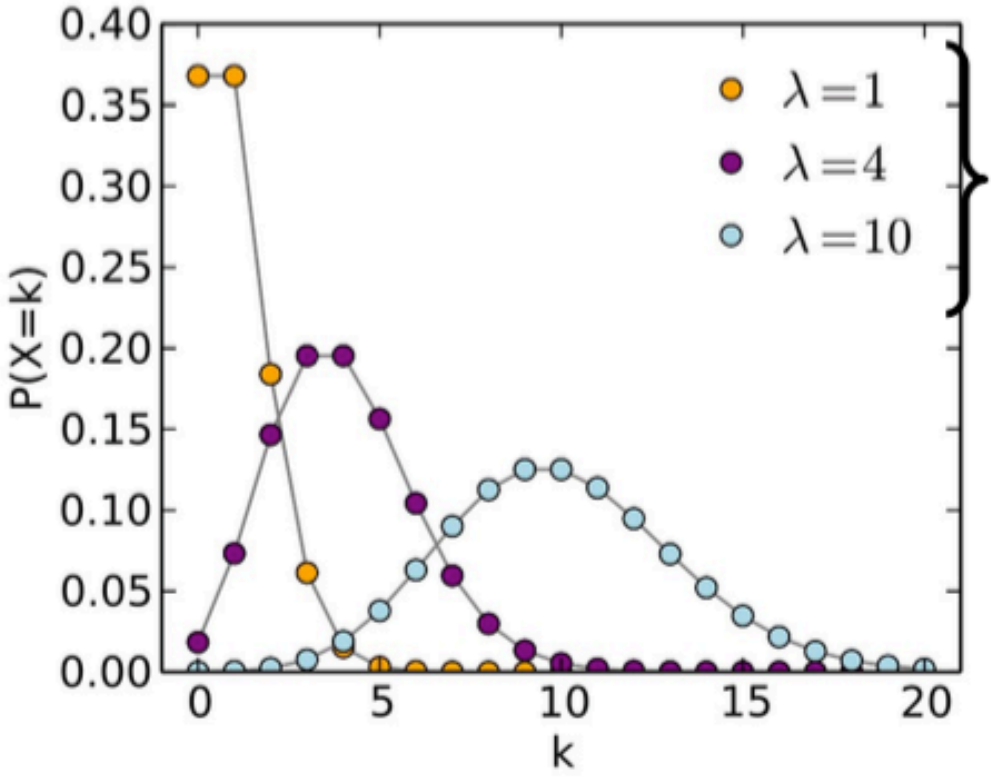
- Used to quantify biological and technical variation in your dataset to reveal variation of interest
- What model best fits RNA-seq data?

BITS Bioinformatics and Training Services (https://www.bits.vib.be)

- Organize bioinformatics courses/workshops
- Much course material available online

The poisson distribution

The counts of technical replicates follow a **poisson distribution** (Marioni et al 2008). The Poisson distribution can be applied to systems with a large number of possible events, each of which is rare.



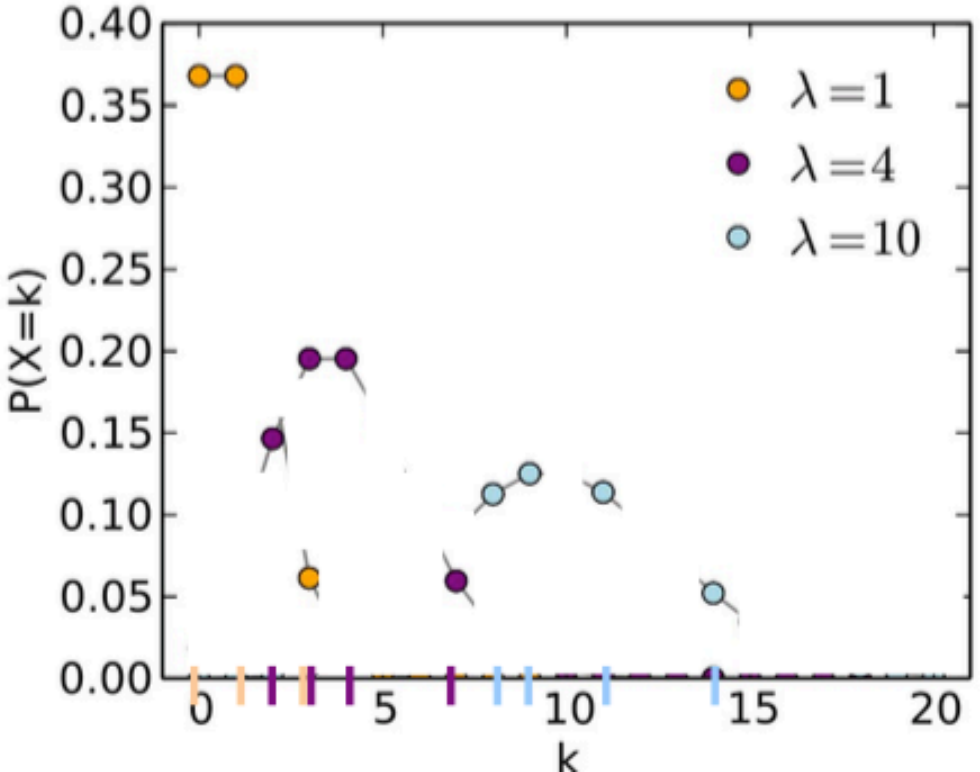
From Wikipedia. Can be 3 different genes, each with their own poisson distribution. Lambda is the mean of the gene's distribution, with a certain number of reads.

Y=axis: chance to pick that number of reads.

VIB

The poisson distribution

So when we have 4 technical replicates sequenced up to a big depth (say 10 M reads). We can get **by chance**, these numbers for 3 different genes.



GeneA **0, 0, 1, 3**

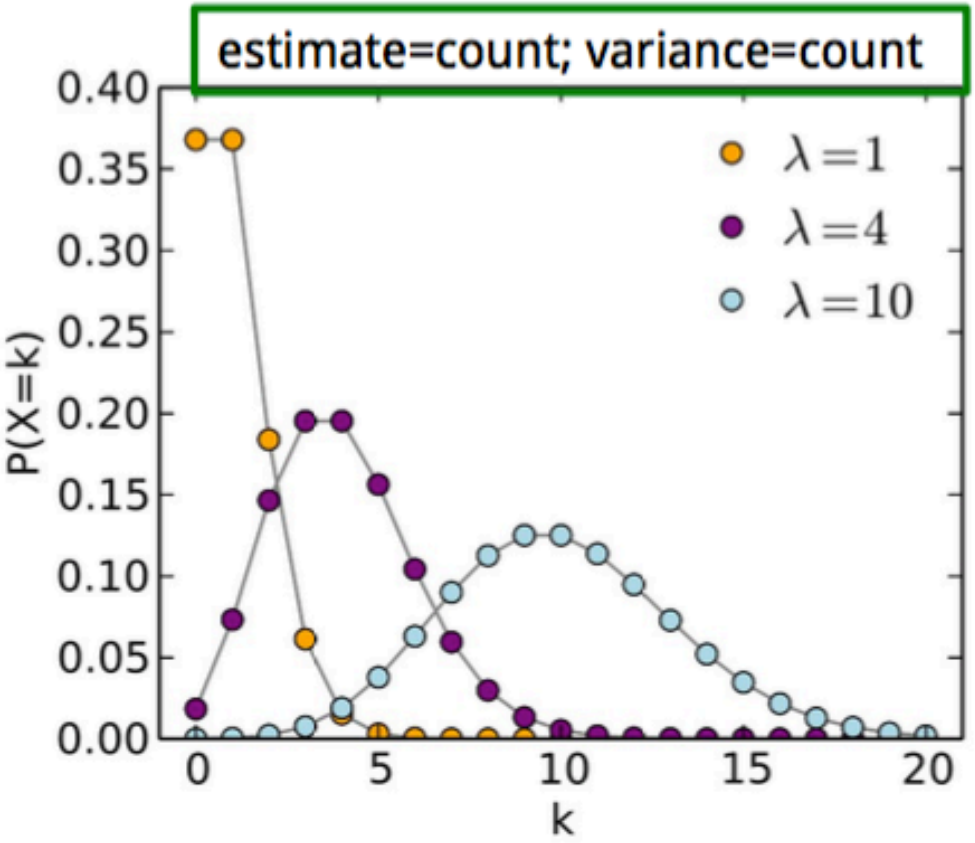
GeneB **2, 3, 4, 7**

GeneC **8, 9, 11, 14**

VIB

Conclusion of the experiment

How bigger the fraction in the pool, how quicker (i.e. with less sequencing depth) we are **certain** about the estimate of that fraction.



For lower counts, the variance is **relatively bigger** than the variance for higher counts.

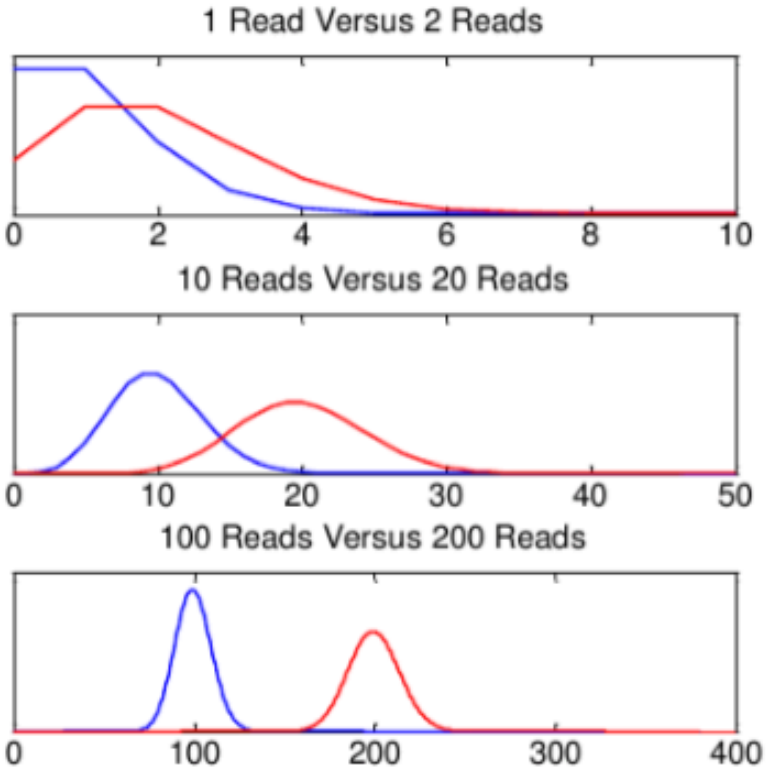
CV (coëfficient of variation) = sqrt(count)/count

Genes with **lower expression** need much deeper sequencing than genes with higher expression levels.

Comparing counts

"Here we show the overlap of Poisson distributions of single measurements at different read counts. Because **relative Poisson uncertainty is high at low read counts**, a count of 1 versus 2 has very little power to discriminate a true 2X fold change, though at higher counts a 2X fold change becomes significant.

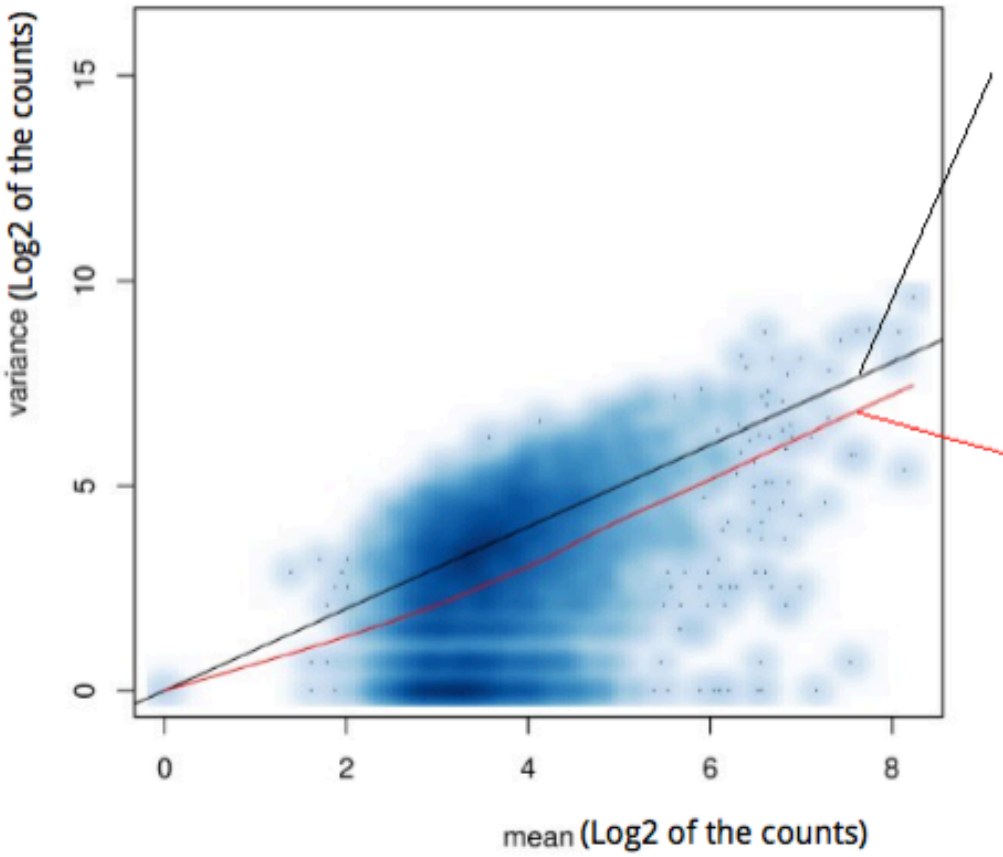
In an actual experiment, the width of the distribution would be greater due to additional biological and technical uncertainty, but the **uncertainty** to the mean expression **would narrow with each additional replicate**."



VIB

Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. Bioinformatics (2013) doi: 10.1093/bioinformatics/btt015

Comparing technical replicates



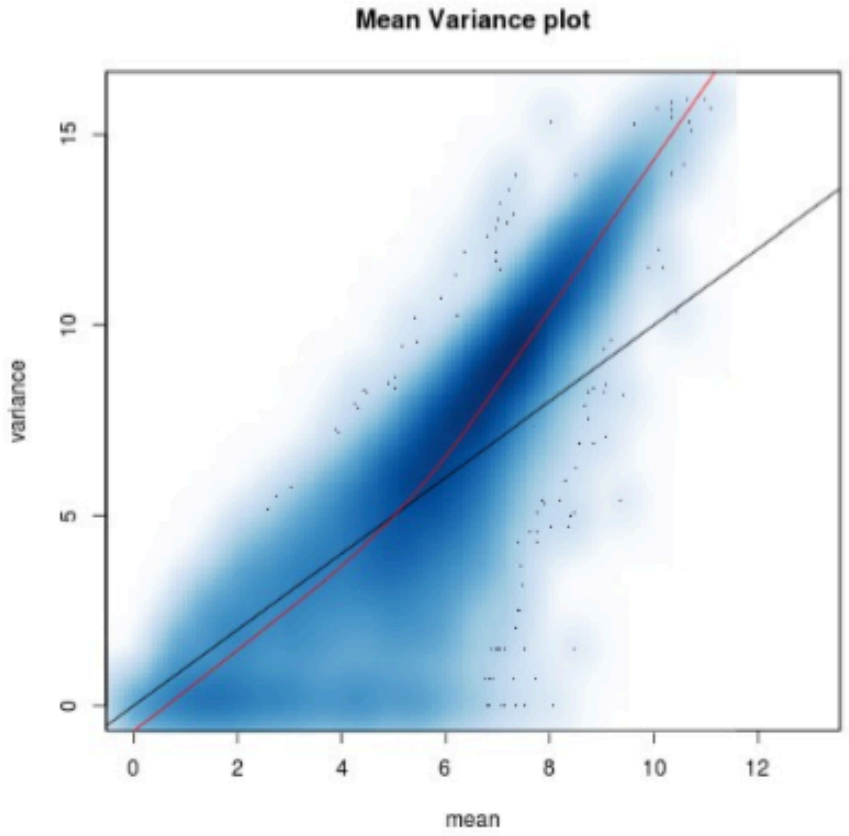
Correlation between mean and variance according to Poisson

Lowess fit through the data

Risso et al. "GC-Content Normalization for RNA-Seq Data"
BMC Bioinformatics 2011, 12:480
http://www.biomedcentral.com/1471-2105/12/480 - EDASeq package (R)

But poisson does not seem to fit

Extending the samples to real biological samples, this mean variance relationship does not hold...



Plotted using EDASeq Package in R.

VIB

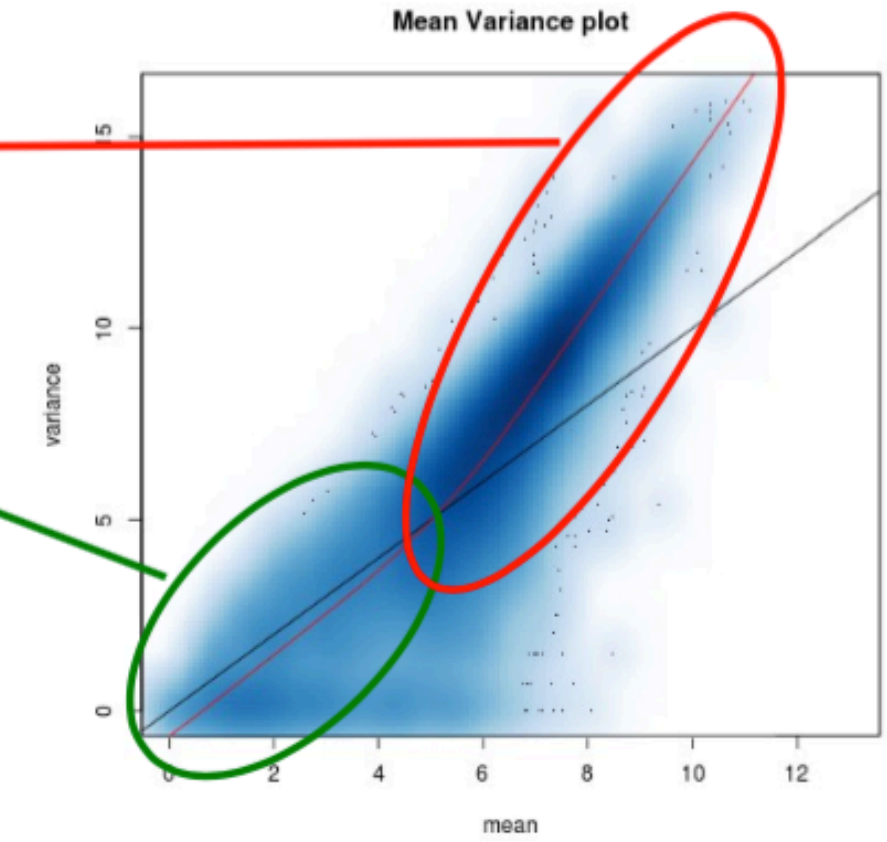
But poisson does not seem to fit

Extending the samples to real biological samples, this mean variance relationship does not hold!

Something is going on!

Reasonable fit

Plotted using EDASeq Package in R.

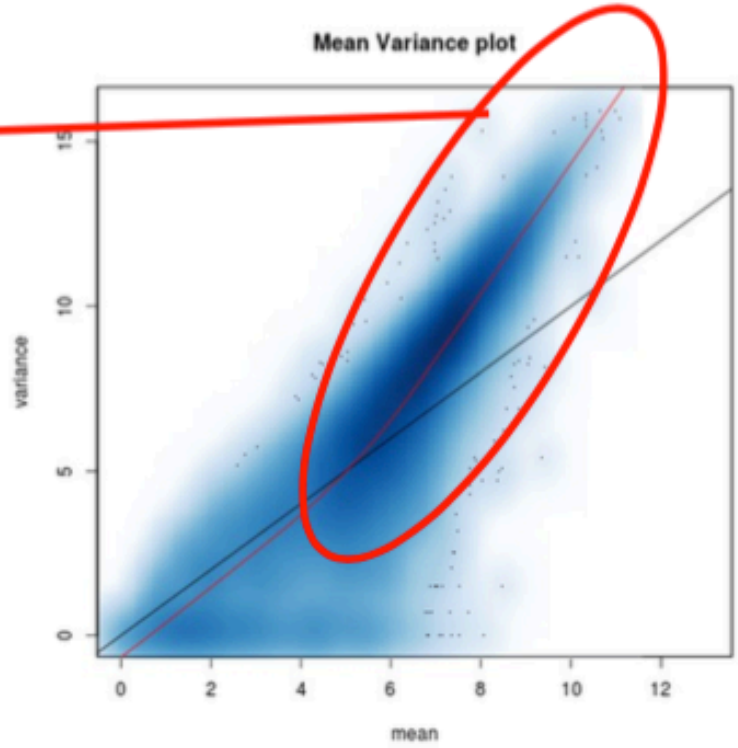


VIB

An extra source of variation

The Poisson distribution has an '**overdispersed**' variance: the variance is bigger than expected for higher counts between biological replicates.

Something is going on!



Plotted using EDASeq Package in R.

VIB

An extra source of variation

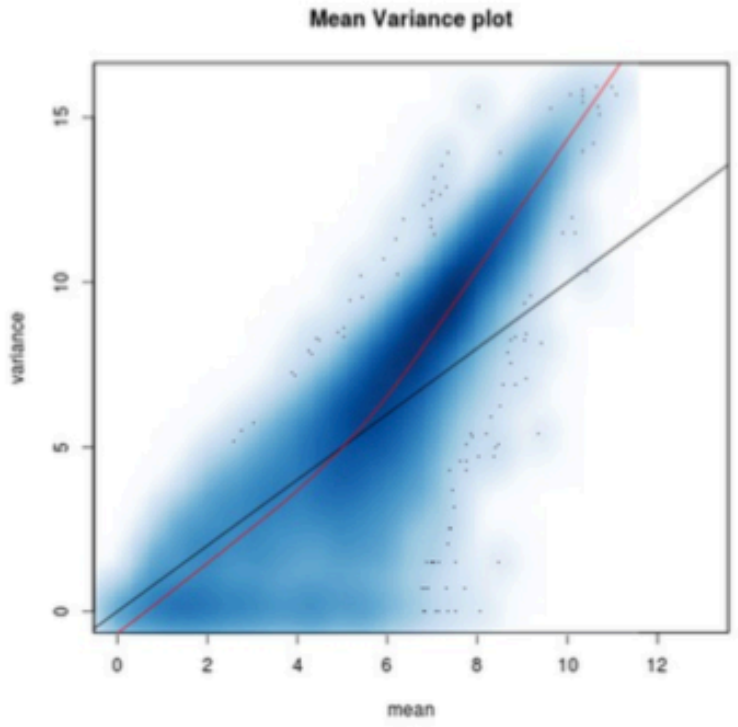
Where Poisson: CV = std dev / mean => $CV^2 = 1/\mu$
If an additional distribution is involved (also dependent on π, the fraction of the gene in the cDNA pool), we have a

mixture of distributions:

$CV^2 = 1/\mu + \varphi$

Low counts!

dispersion



Generalization of Poisson with this extra parameter: the **Negative Binomial Model** fits better!

VIB

Variation summary, intuitively

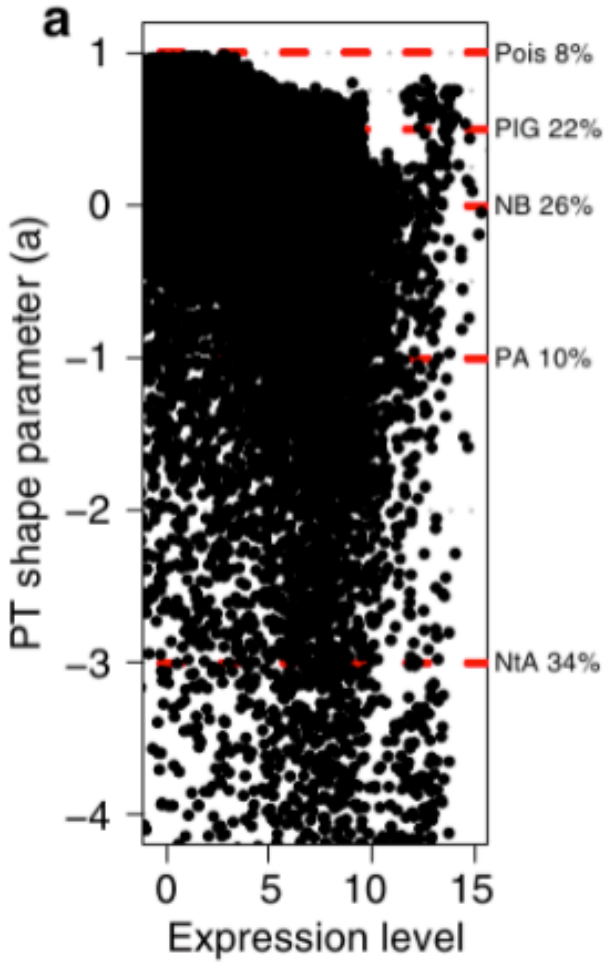
Total CV^2 = Technical CV^2 + Biological CV^2

For **low counts**, the Poisson (technical) variation or the measurement error is dominant.

For **higher counts**, the Poisson variation gets smaller, and another source of variation becomes dominant, the **dispersion** or the **biological variation**. Biological variation does not get smaller with higher counts.

VIB

Beyond the NB model



It appears from analysis of many biological replicates (#=69) that not every gene can be modeled as NB: the **Poisson-Tweedie** model provides a further generalisation and a better fit for many genes (with an additional shape parameter).

Left figure: raw data shows that about 26% of the genes fit a NB model. Depending on the estimated shape parameter, other distributions fit better.

VIB

Esnaola et al. BMC Bioinformatics 2013, 14:254
http://www.biomedcentral.com/1471-2105/14/254

Multiple hypothesis testing

- Thousands of genes = thousands of hypothesis tests (simultaneously)
- Increased chance of false positives! (Type I error)
  - e.g. you test for differential expression in 1000 genes that are not differentially expressed
  - You would expect 1000 x 0.05 = 50 of them to have a *P*-value < 0.05
- Individual *P*-values not useful
  - Need multiple testing statistic instead

False discovery rate
(Benjamini & Hochberg 1995)

- The expected proportion of Type I errors among the rejected hypotheses
  - i.e. the proportion of false positives
- Tends to be conservative if many genes are DE
  - FDR = 0.05 common for exploratory/broad scope studies
  - FDR < 0.05 common for medical applications and hunts for candidate genes