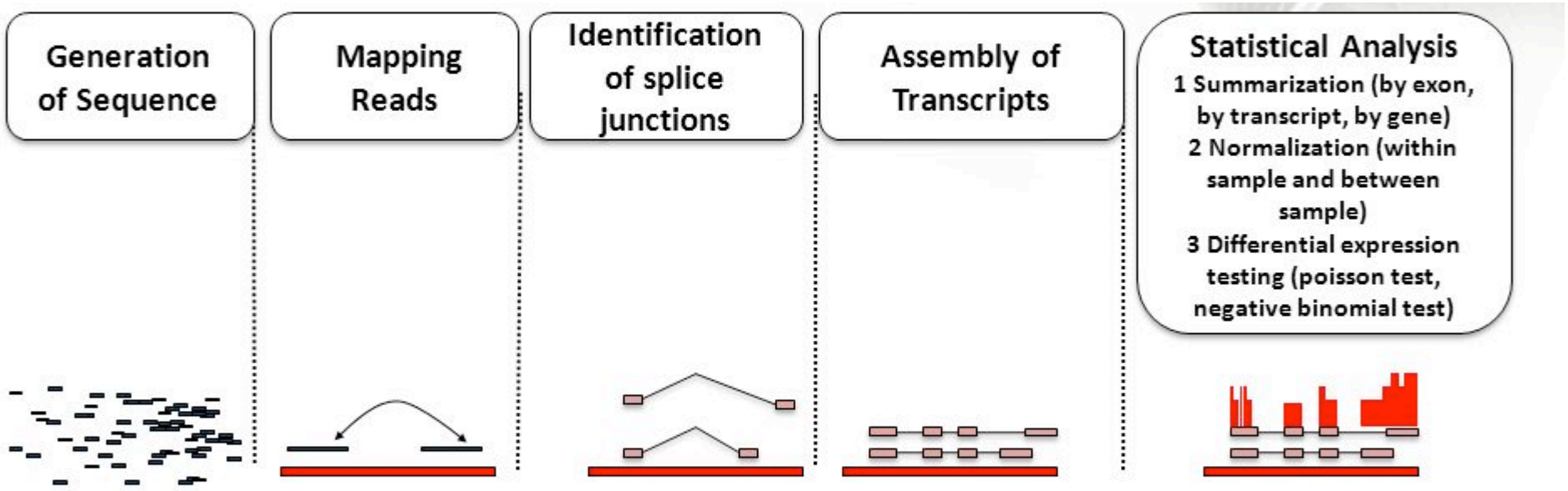
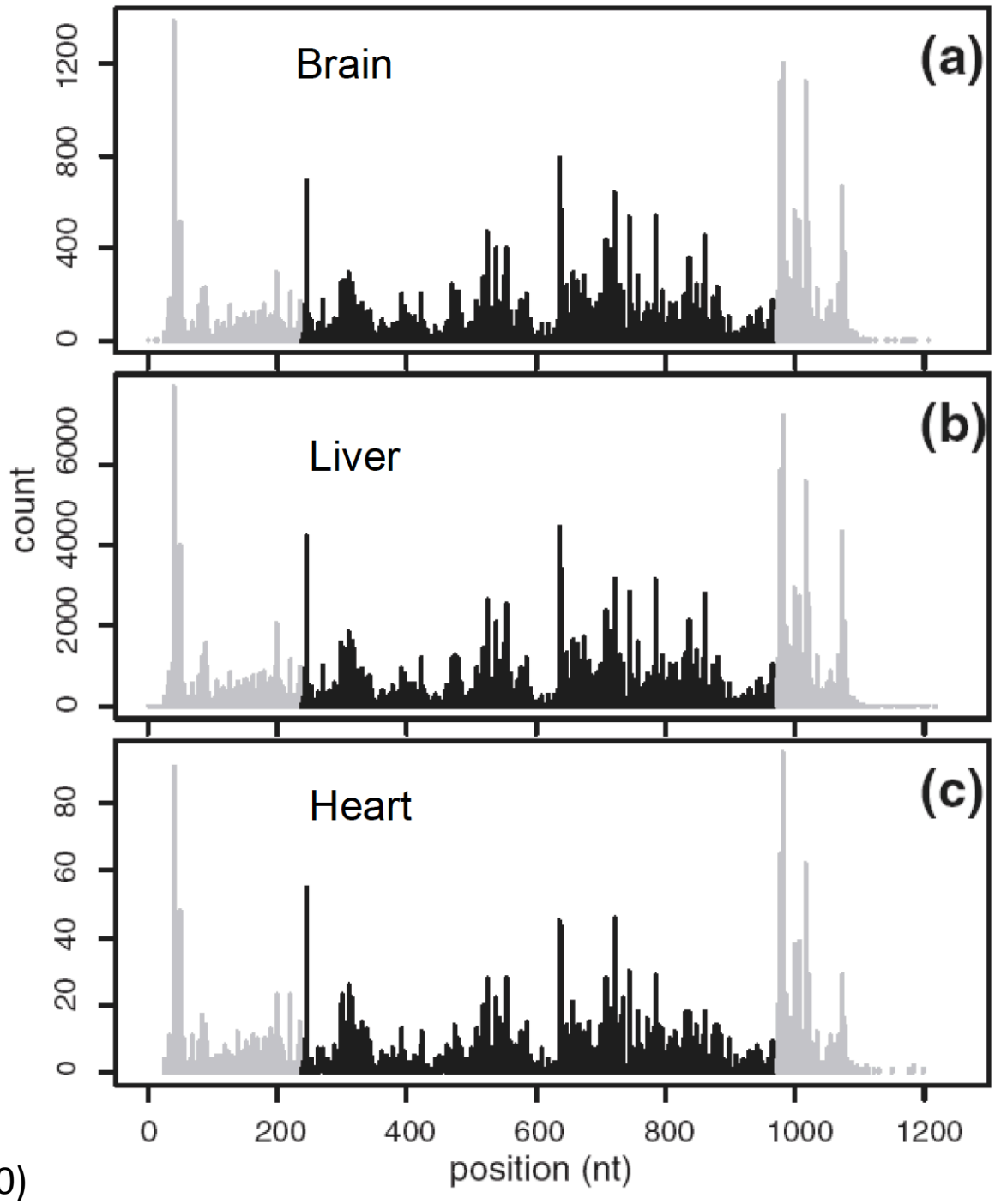
"Intelligence is the ability to adapt to change."
-- Steven Hawking



Read are non-uniformly distributed, but same pattern across tissues with large differences in expression levels.

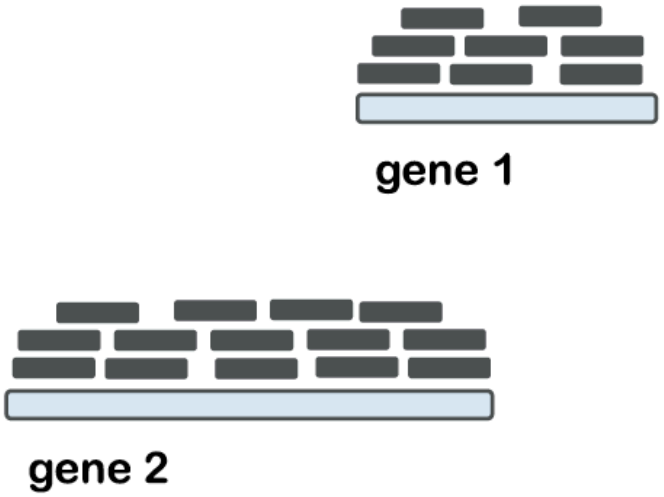


Example: read counts along the transcript of the Apoe gene in mouse.

Local sequence predicts read rate variation. (see also Hansen et al 2010)

Revisiting normalization

We can count reads per GENE (ignore isoforms), per isoform, or per exon.



	sample A	sample B
gene 1	50	40
gene 2	50	40
…	…	…
gene 99	50	40
gene 100	10	1000

d_A = 4'960 d_B = 4'960

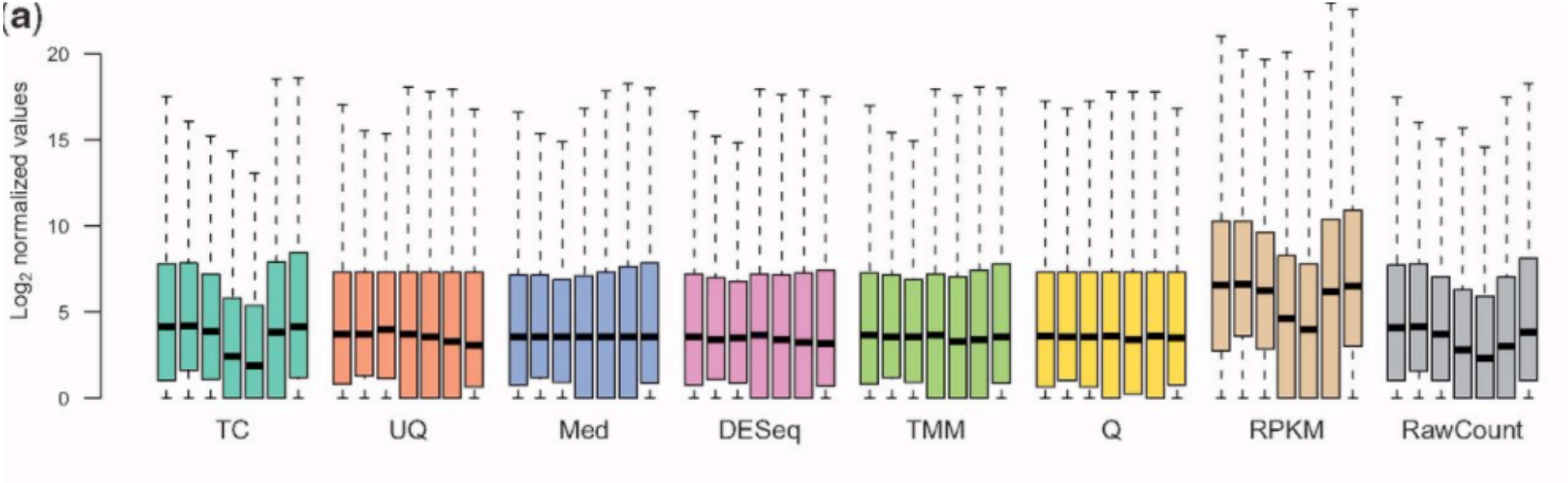
The problem with simple RPKM/FPKM/RPM metrics (global scaling techniques) is they technically only capture and correct for differences related to high-count genes.

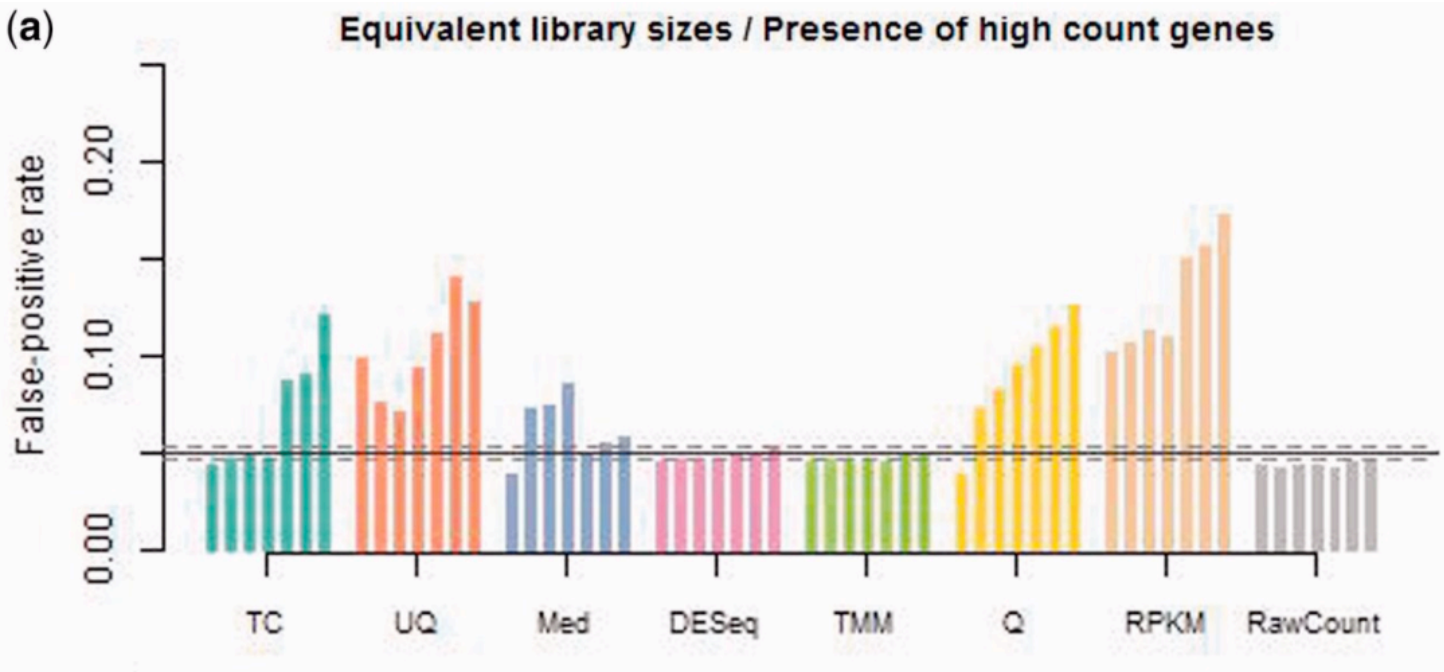
Alternative: adjusts counts distributions with respect to the middle quartiles, so to reduce the effect of high-counts genes.

Trimmed Mean of M-values

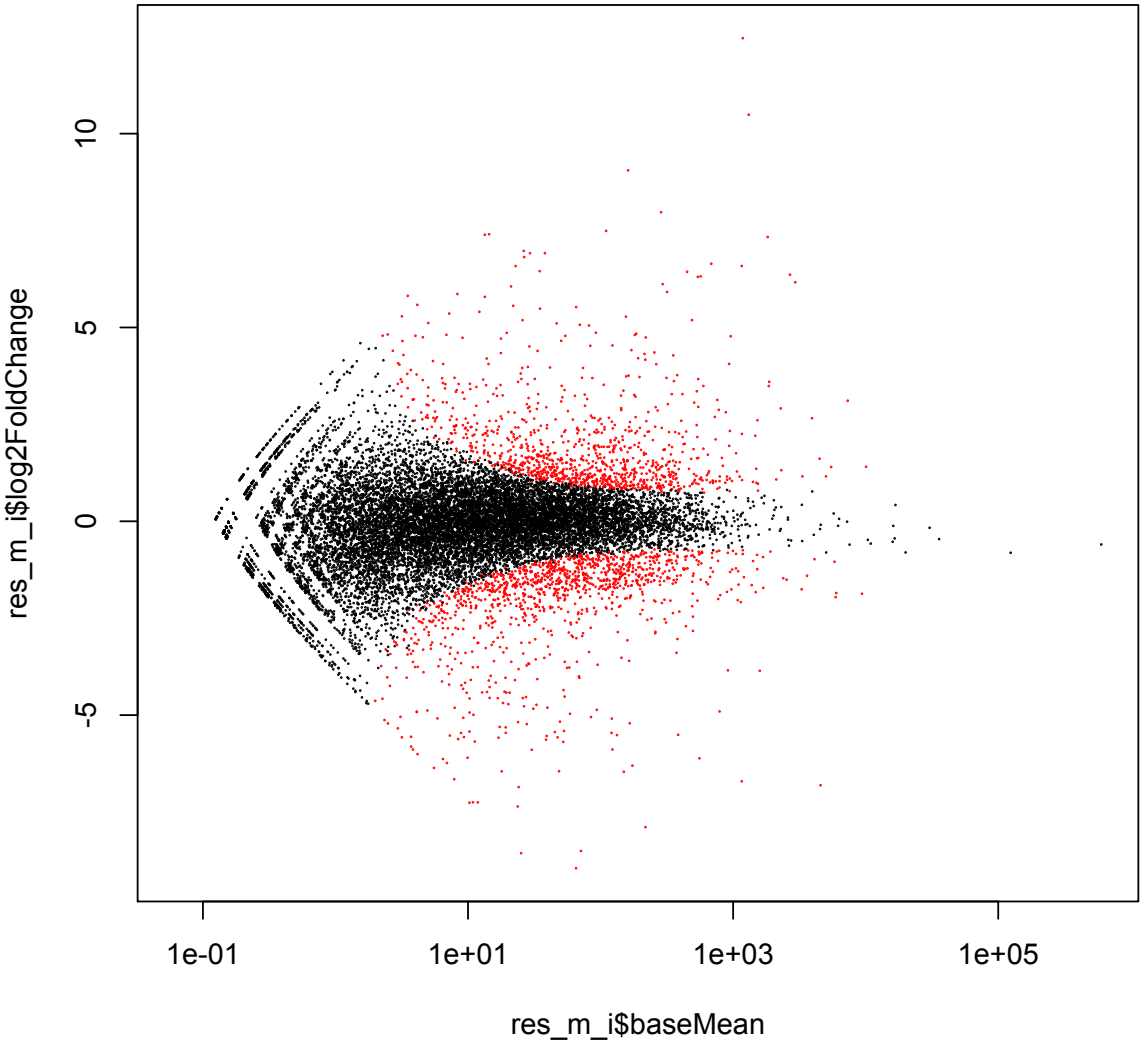
- Trimmed Mean of M-values (TMM) removes the 30% of genes with the most extreme values.
- Similar to quantile normalization but easily accounts for differences in library composition between samples. Many similar approaches.
- However, if MANY things are change or read depth is dramatically different then really best approach is to leverage spike-in RNA.

Dillies et. al. **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. (2013)**





Differential Expression



Basic ideas

1. To determine the genes whose read count differences between two conditions are greater than expected by chance, DGE tools must make assumptions about the distribution of read counts.
2. The null hypothesis { that the mean read counts of the samples of condition A are equal to the mean read counts of the samples of condition B } is tested for each gene individually.

Because tests are essentially looing for deviations from expected variance … ***A good estimate of variance for each gene is essential to determine whether the changes are due to chance.***

What distribution are read data?

The binomial distribution works when we have a fixed number of events n, each with a constant probability of success p.

- e.g., a series of $n = 10$ coin flips, each of which has a probability of $p = 5$ of heads
- The binomial distribution gives us the probability of observing k heads

$$p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Event: An RNAseq read "lands" in a given gene (success) or not (failure)

BUT …

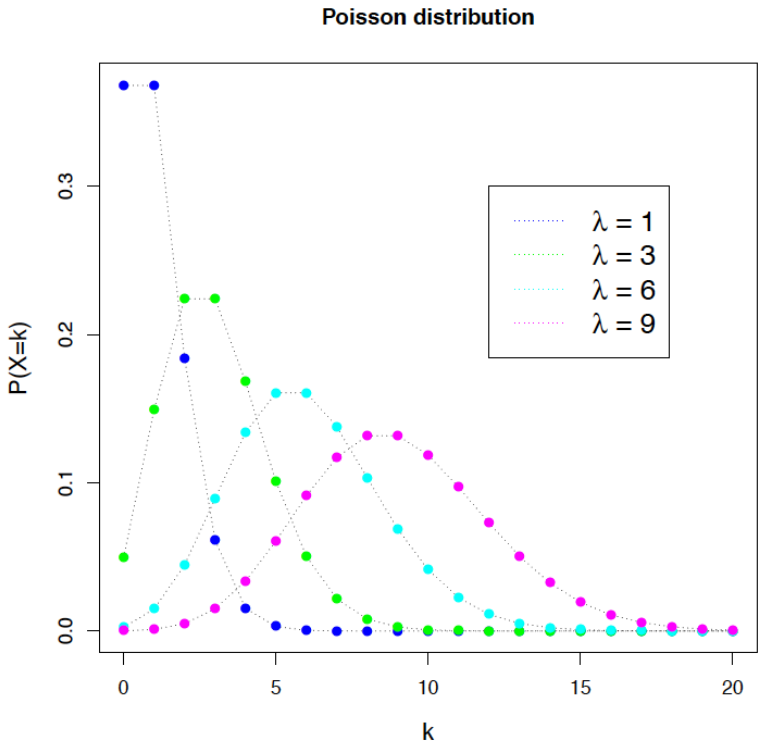
- In sequencing data we simply don't know the number of trials (n) that will happen but only the number of successes (reads) observed.

- E.g. We do not know how many times success "did not happen".

Poisson

- Individual reads can be interpreted as binary data (Bernoulli trials): they either originate from gene i or not.
- We are trying to model the discrete probability distribution of the number of successes (success = read is present in the sequenced library).
- the pool of possible reads that could be present is large, while the proportion of reads belonging to gene i is quite small.

Hence RNA-seq is often modeled as a Poisson distribution



Remember we used the Poisson when talking about coverage!

- For $X \sim \text{Poisson}(\lambda)$, both the mean and the variance are equal to λ

Replicates

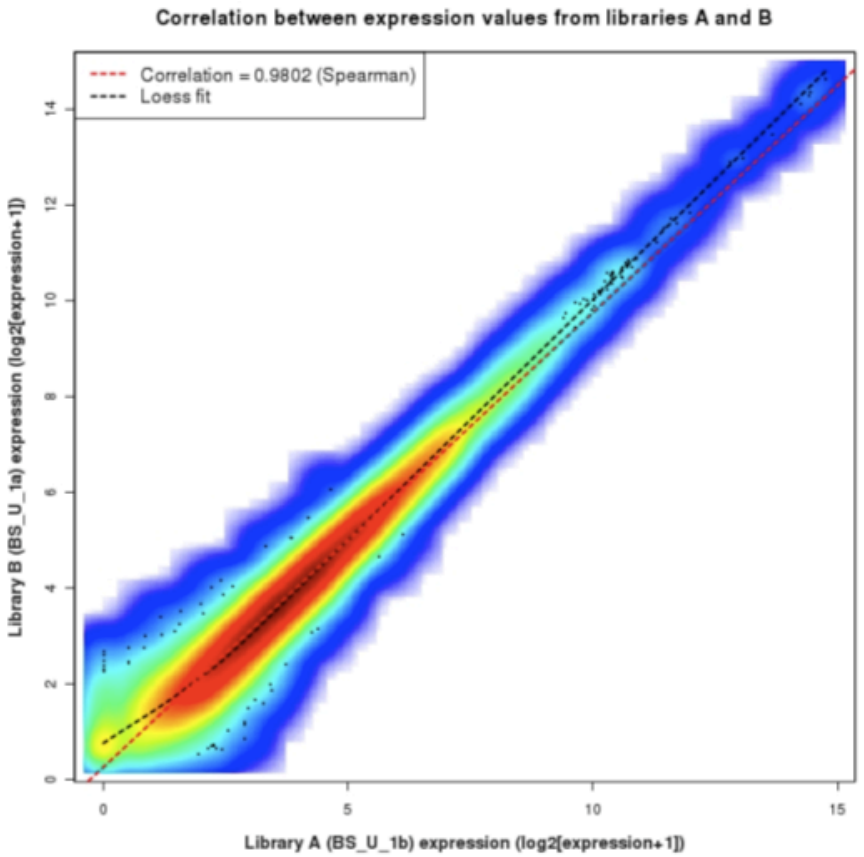
Fundamental problem with generalizing results gathered from unreplicated data is a **complete lack of knowledge about biological variation**.

Without an estimate of variability within the groups, there is no sound statistical basis for inference of differences between the groups.

More replicates = Better Estimates!!

Replicates

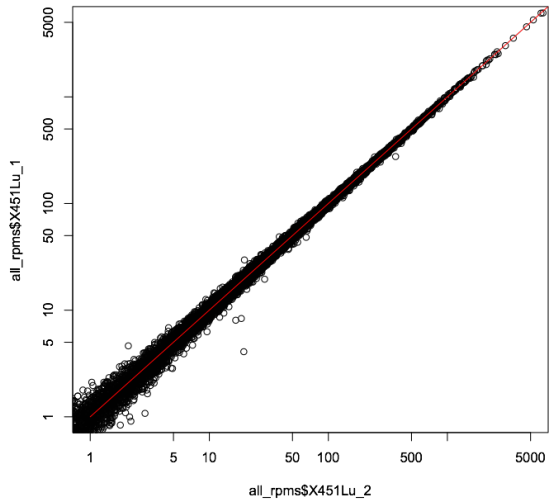
- Technical Replicate
  - Multiple instances of sequence generation
    - Flow Cells, Lanes, Indexes
- Biological Replicate
  - Multiple isolations of cells showing the same phenotype, stage or other experimental condition
  - Some example concerns/ challenges:
    - Environmental Factors, Growth Conditions, Time
  - Correlation Coefficient 0.92-0.98



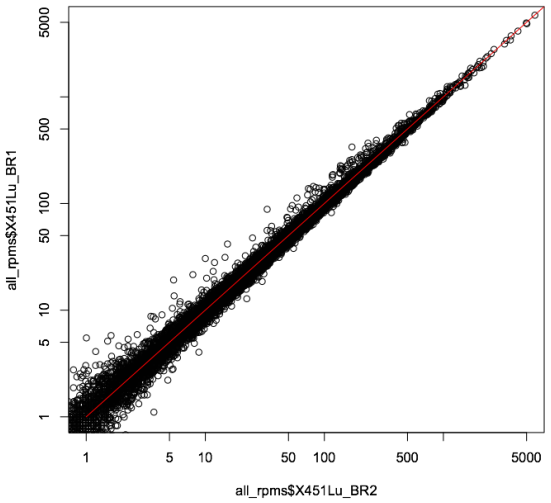
Overdispersion

- Technical replicates are well estimated by Poisson ***but*** biological replicates show overdispersion (e.g. greater variance (noise) than expected).

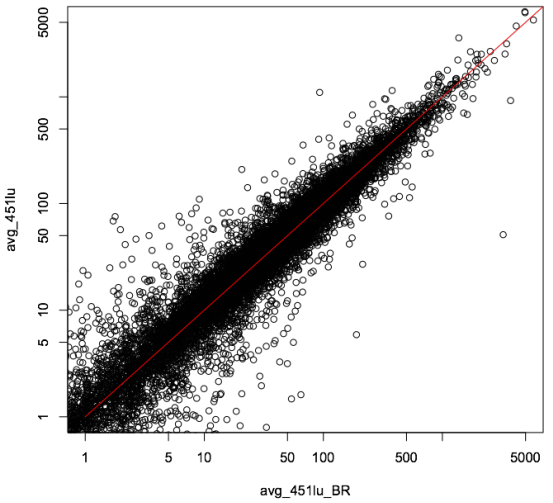
Replicates, Condition 1



Replicates, Condition 2



Condition 1 Vs. Condition 2



All axes show RPM
(reads per million mapped)

Overdispersion

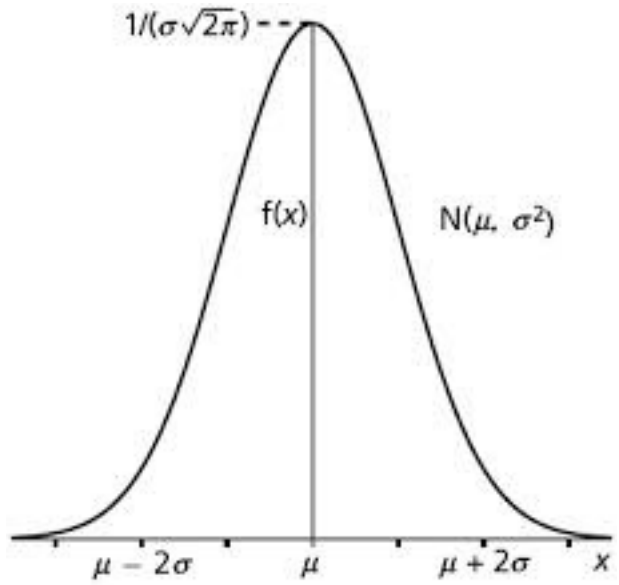
- Technical replicates are well estimated by Poisson ***but*** biological replicates show overdispersion (e.g. greater variance (noise) than expected).

- **Overdispersion** can be captured with the negative binomial distribution, which is a more general form of the Poisson distribution that allows the variance to exceed the mean

But now need mean and dispersion from data!

Statistical Distributions

gaussian, poisson, negative binomial -- what does all this mean?



$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gaussian (normal) distributions

- -nice and easy to work with
- -describe smooth distributions
- -worked well for microarrays
- -underlie the t-test (among others)

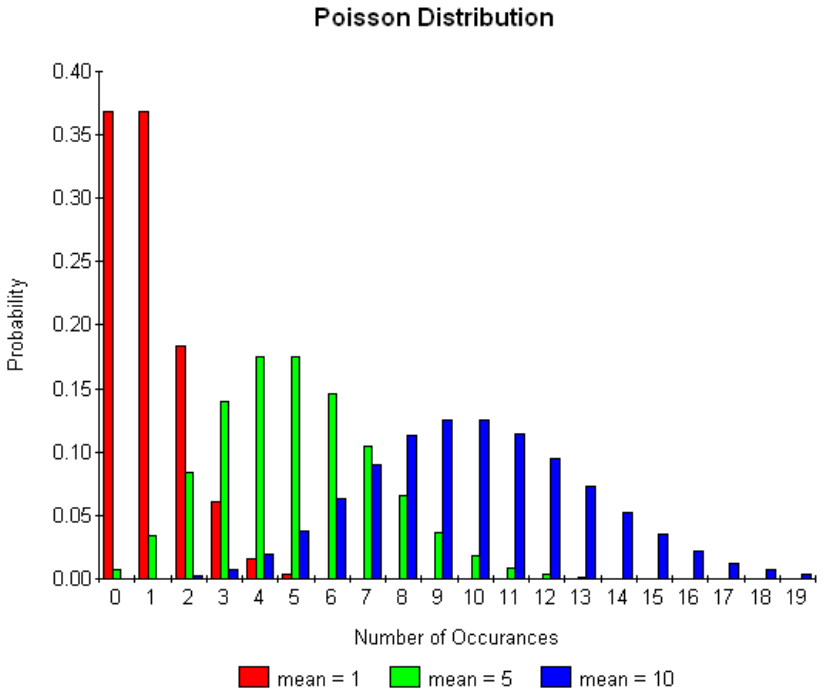
The **dispersion** in this case is equal to the standard deviation

You completely specify this distribution by the mean (μ) and the standard deviation (σ).

BUT … since reads are count based, they can't be normally distributed (you can't have -3 counts, or 12.2 counts)

Statistical Distributions

gaussian, **poisson**, negative binomial -- what does all this mean?



$$f(k;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

Poisson distributions

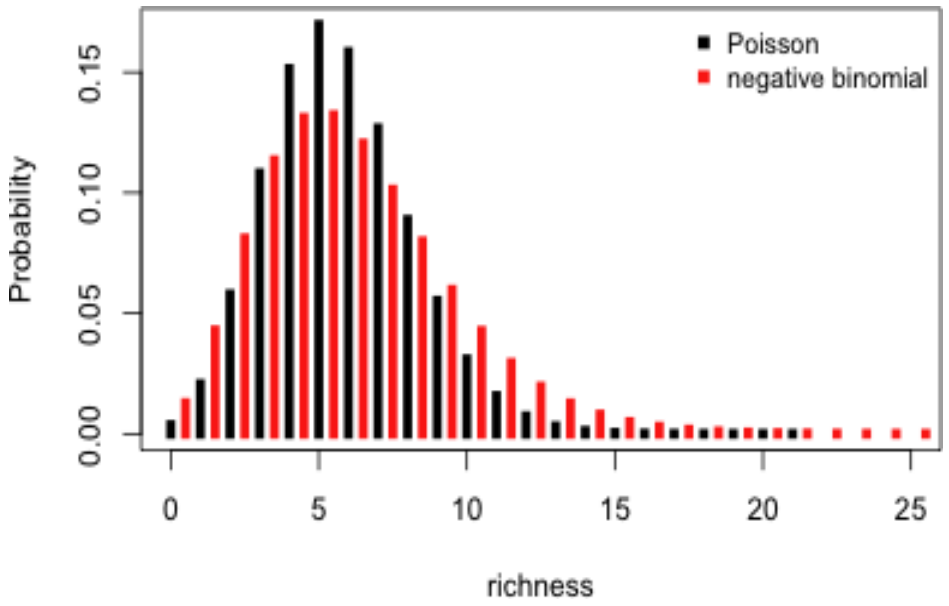
- like a gaussian for non-smooth distributions
- describes things like stars in a small area on the sky
- for very large numbers, this looks like a gaussian distribution

The **dispersion** in this case is equal to the mean (λ). You completely specify this distribution by the mean.

e.g., For a set of samples where your gene has an average of λ=100 reads overall, how likely is it that your gene randomly has k=10 reads in the mutant samples?

Statistical Distributions

gaussian, poisson, **negative binomial** -- what does all this mean?



Negative Binomial distributions

- like a poisson but allows the variance to be different from the mean
- often called "over-dispersed" poisson distribution
- intuitively this is a poisson where the mean itself is associated with uncertainty

The **dispersion** (variance σ^2) in this case is measured empirically from the data.

$$\sigma^2 = \mu + \alpha\mu^2$$

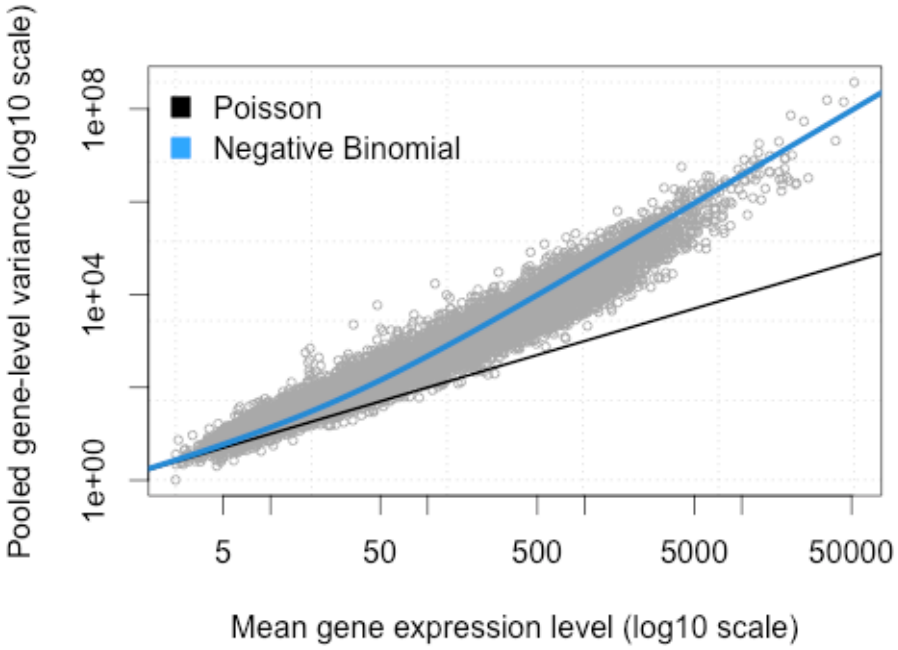
From this formula it is evident that the dispersion is always greater than the mean for $\alpha > 0$. If $\alpha = 0$, the NB distribution is a Poisson distribution.

Statistical Distributions

gaussian, poisson, **negative binomial** -- what does all this mean?

RNA-seq data fits a Negative Binomial (NB) distribution.
But really, that's just saying that RNAseq looks like "counts" data with more variation than just statistical fluctuations– i.e. it also has biological variation in it.

How do we know? Because, when you measure variance (per gene, between replicates), it's not equal to the mean, and it's not even a good linear fit



the variance of counts is generally greater than their mean, especially for genes expressed at a higher level.

* Unforunately for real data, even the NB fit isn't always great

Anders & Huber, Genome Biology 2010

Estimating dispersion

- Unfortunately, we have to be content with few biological replicates per condition due to the high costs associated with sequencing experiments and the large amount of time that goes into library preparations.

- This makes the gene-wise estimates of dispersion rather unreliable.

Compensating for lack of replicates

- Some tools therefore compensate for the lack of replication by incorporating **PRIOR** information.

- E.g. “borrowing” information across all genes with similar expression values.

  Bayesian “shrinkage”

- These fitted values of the mean and dispersion are then used instead of the raw estimates to test for differential gene expression.

General summary of major methods

Feature	DESeq2	edgeR	limmaVoom	Cuffdiff
Seq. depth normalization	Sample-wise size factor	Gene-wise trimmed median of means (TMM)	Gene-wise trimmed median of means (TMM)	FPKM-like or DESeq-like
Assumed distribution	Neg. binomial	Neg. binomial	*log*-normal	Neg. binomial
Test for DE	Exact test (Wald)	Exact test for over-dispersed data	Generalized linear model	*t*-test
False positives	Low	Low	Low	High
Detection of differential isoforms	No	No	No	Yes
Support for multi-factored experiments	Yes	Yes	Yes	No
Runtime (3-5 replicates)	Seconds to minutes	Seconds to minutes	Seconds to minutes	Hours


SPRING BREAK IS HERE!!