# RNA-seq part II: quantification and models for assessing differential expression

Ian Dworkin

NGS 2013

# What we will cover today

- Why we use count data as input
- Introducing a bit of probability to why many RNA Differential analysis tools use a negative binomial.
- Why do care about variance/over-dispersion so much.
- How do we estimate over-dispersion with small sample sizes (and why edgeR and DGE give different results).

# Counting

- One of the most difficult issues has been how to count reads.

# Counting

- We are interested in transcript abundance.
- But we need to take into account a number of things.
- How many reads in the sample.
- Length of transcripts
- GC content and sequencing bias (influencing counts of transcripts within a sample).

# Seemingly sensible Counting (but ultimately not so useful).

- RPKM (reads aligned per kilobase of exon per million reads mapped) – Mortazavi et al 2008
- FPKM (fragments per kilobase of exon per million fragments mapped). Same idea for paired end sequencing.

Take home message:
Actual counts should be used as input for differential expression analysis, not (pre)scaled measures.

# RPKM

$$\mathrm{RPKM}_{\mathrm{G}} = \frac{r_g \times 10^9}{\mathrm{fl}_{\mathrm{g}} \times R}$$

$R$ = total # mapped reads from that sample

$$R = \sum_{g \in \mathrm{G}} r_g$$

$\mathrm{fl}_{\mathrm{g}}$ = feature length (i.e. transcript length)

# Problems with RPKM

- RPKM is not a consistent measure of expression abundance (or relative molar concentration).

- See
  - http://blog.nextgenetics.net/?e=51
  - Wagner et al 2012 Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci

# How about Transcripts per million (TPM)

$$\mathrm{TMP_G} = \frac{r_g \times \mathrm{rl} \times 10^6}{\mathrm{fl_g} \times T}$$

$R$ = total # mapped reads from that sample

$$T = \sum_{g \in \mathrm{G}} \frac{r_g \times \mathrm{rl}}{\mathrm{fl_g}}$$

rl = read length

While TPM is in general more (statistically) consistent, it is still generally not appropriate.

# Normalization (for DE) can be much more complicated in practice

- Why might scaling by total number of reads (sequencing depth) be a misleading quantity to scale by?

# Normalization (for DE) can be much more complicated in practice

- Scaling by total mapped reads (sequencing depth) can be substantially influenced by the small proportion of highly expressed genes.

(What might happen?)

- A number of alternatives have been proposed and used (i.e. using quantile normalization)

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics, 11, 94. doi:10.1186/1471-2105-11-94

# Counting (and normalizing) in practice

- In practice, we do not want to “pre-scale” our data as is done in F/R-PKM or TPM.

- Instead we are far better off using a model based approach for normalizing for read-length or library size in the data modeling per se.
- This is far more flexible.

Take home message:
Actual counts should be used as input for differential expression analysis, not (pre)scaled measures.

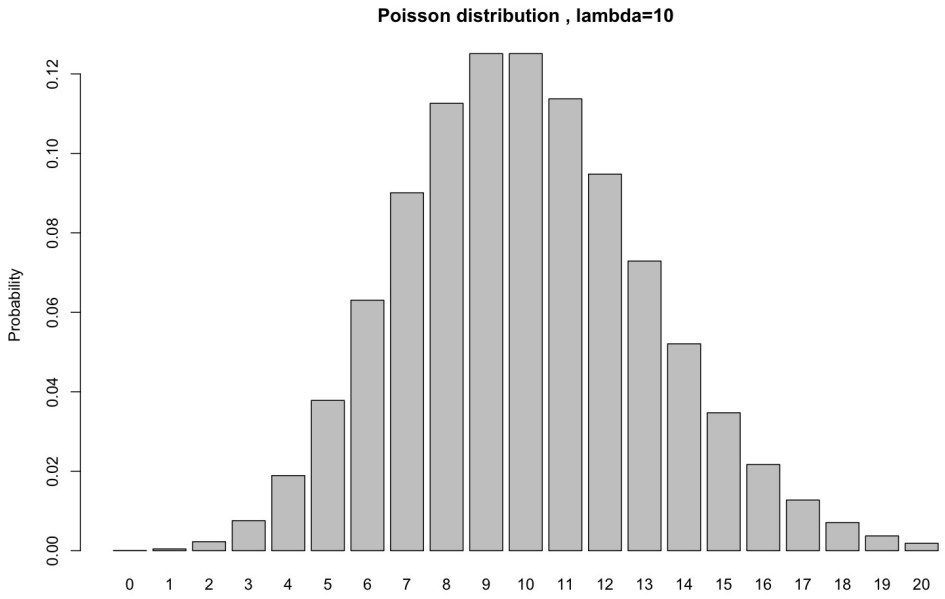
# A bit of background on probability.

- Fundamentally our observed measure of expression are the counts of reads.

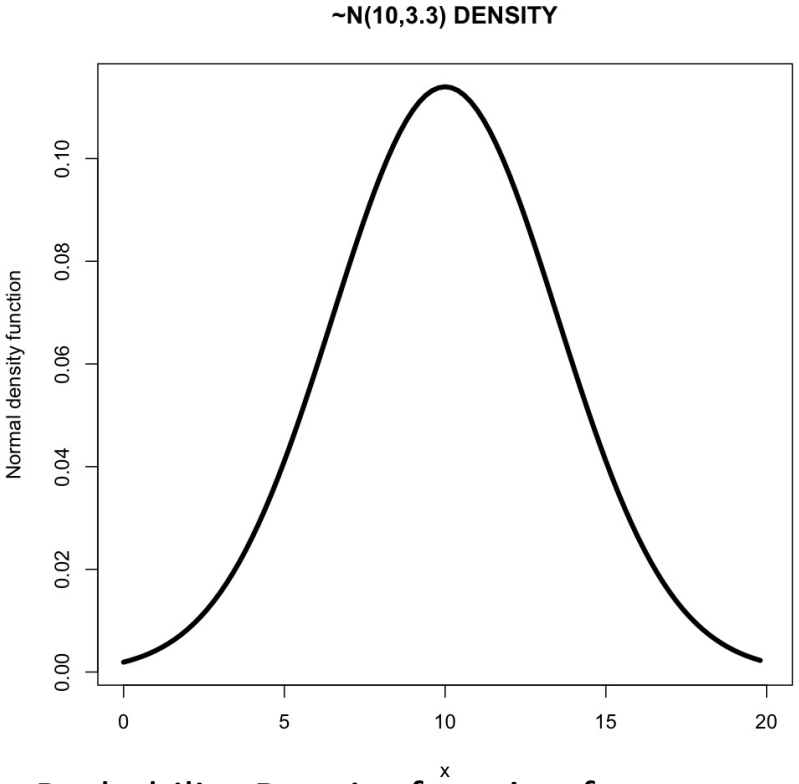
- Depending upon the data modeling framework we wish to use, we need to account for this, as these are not necessarily approximated well by normal (Gaussian) distributions that are used for “standard” linear models like t-tests, ANOVA, regression.

- This is not a problem at all, as it is easy to model data coming from other distributions, and is widely available in stats packages and programming languages alike.

# Probability Density vs. Mass function

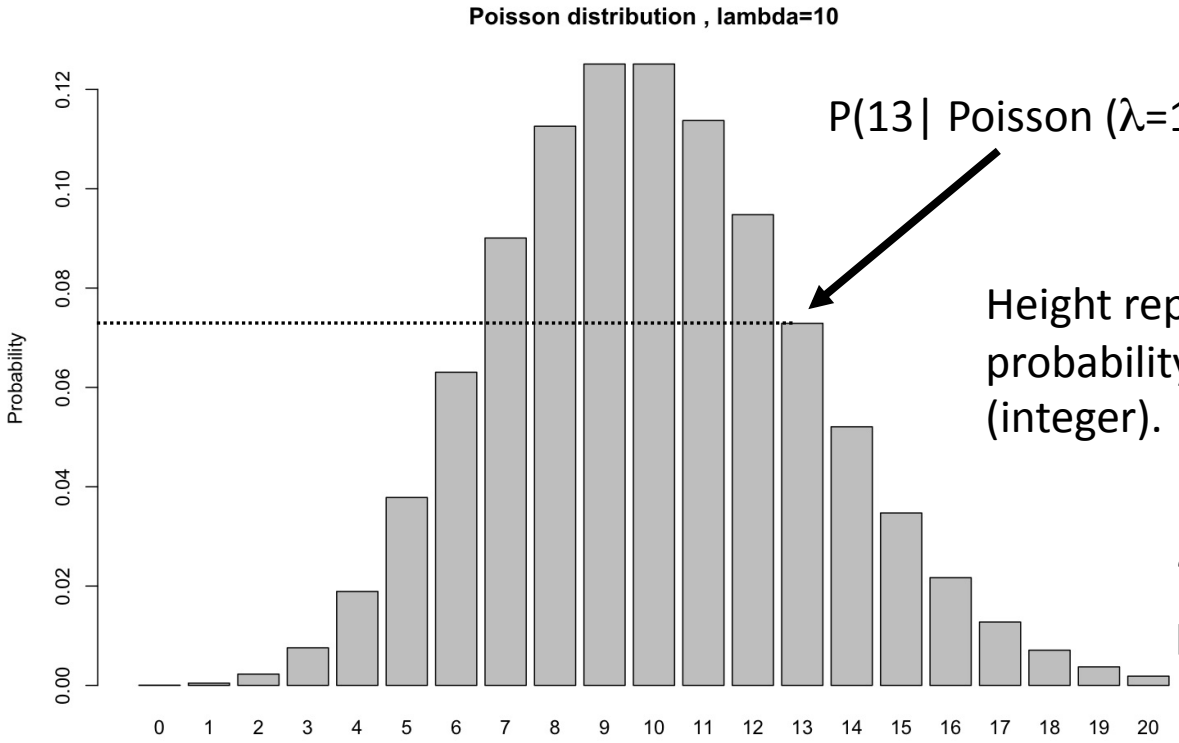



Probability Mass function for a discrete variable.




Probability Density function for a continuous variable.

# Probability Mass function (For discrete distributions, like read counts)




P(13| Poisson (λ=10)) =0.073

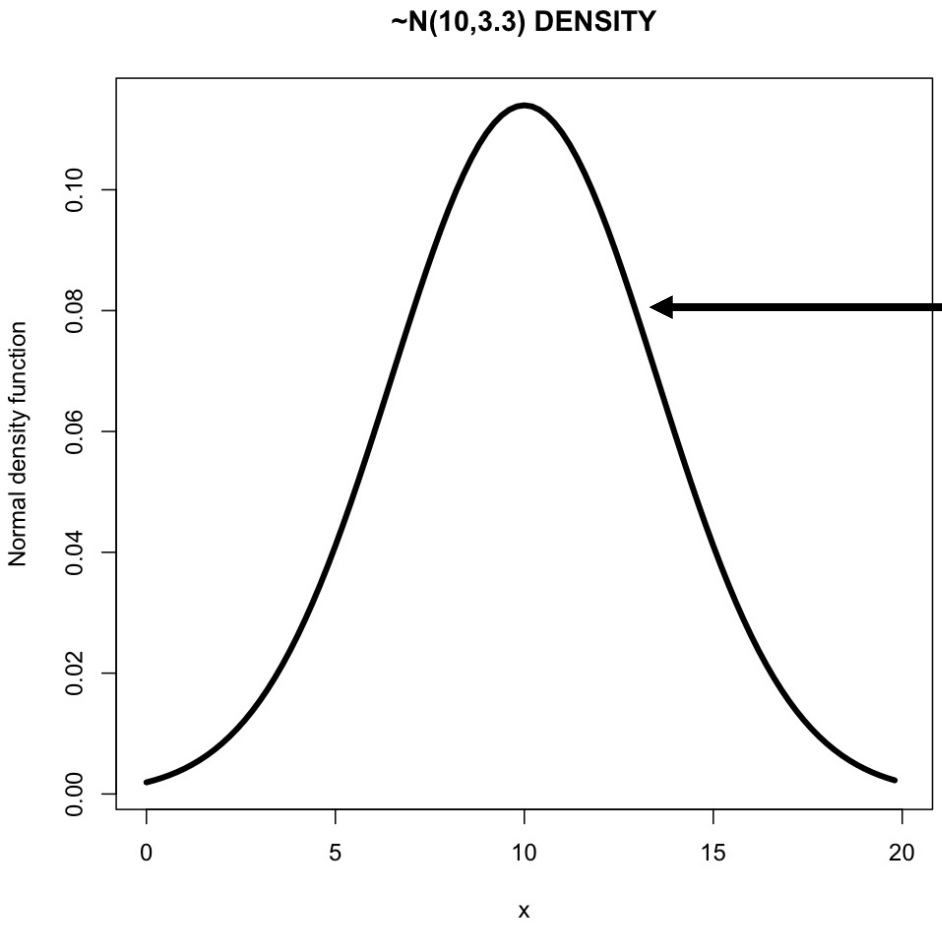
Height represents the probability at that point (integer).

"Area" of the box has no particular meaning.

P(integer) ≥ 0

P(non-integers) = 0.

# Probability Density function

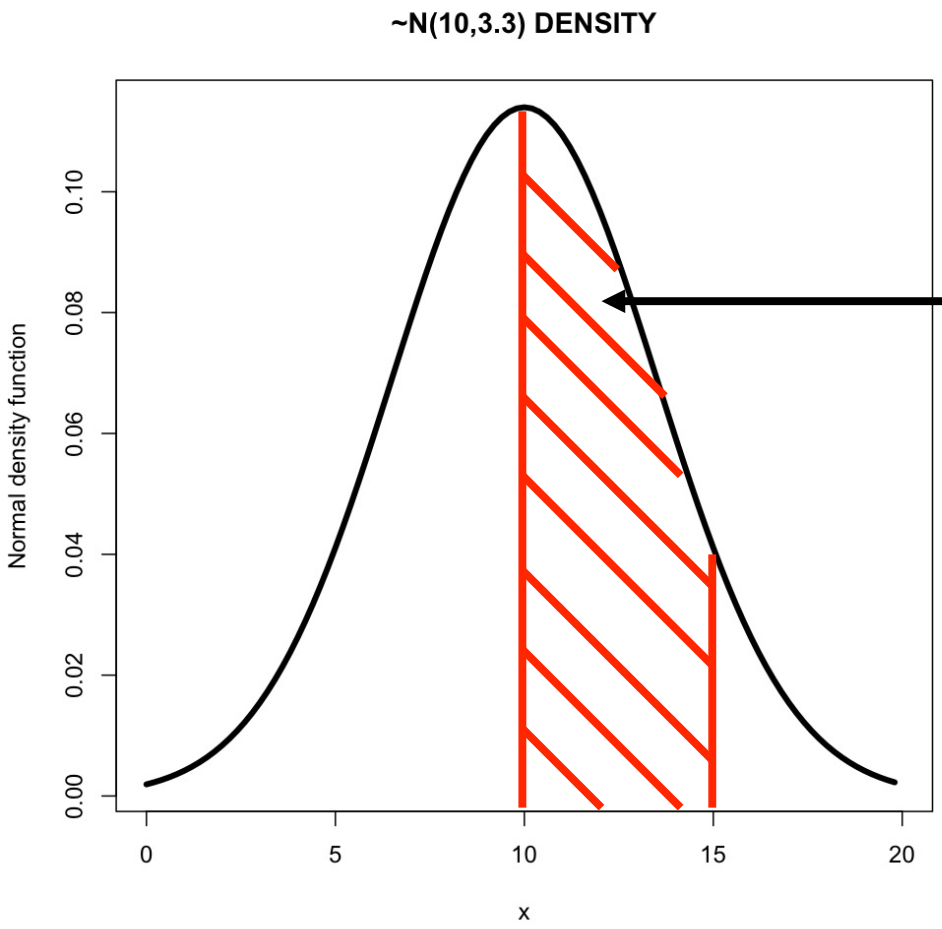



Height at x= 13 is 0.0799

This is not the probability at x=13, but the density.
i.e. f(13) = 0.0799, where f(x) is the normal distribution.

**P(x=13| N(mean=10,sd=3.3)) = 0**
**WHY?**

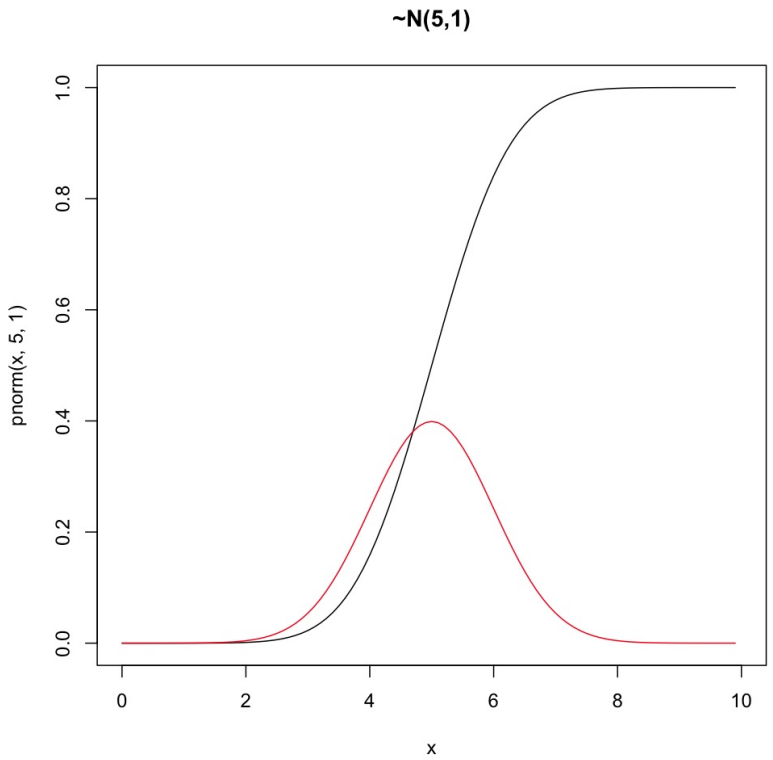
# Probability Density function




We can define the probability in the interval
$10 \leq x \leq 15$

$P(10 \leq x \leq 15 | N(10,3.3)) = 0.435$

# Clarifications on continuous distributions.




$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

$$P(X = x) = 0$$

AREA UNDER CURVE OF PDF =1

(The integral of the normal)

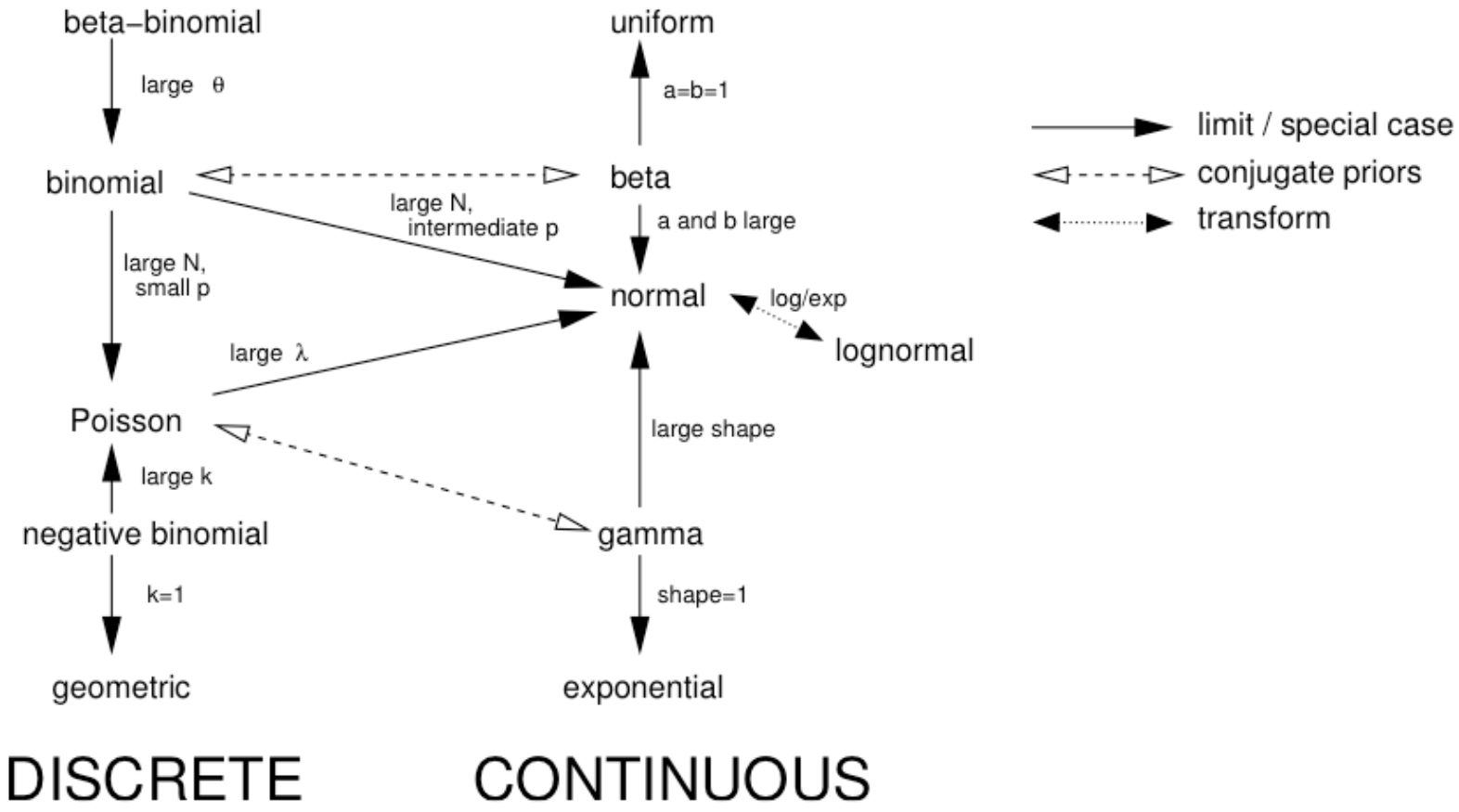



Figure 4.17 Relationships among probability distributions.

Bolker 2007 CH4 page 137

The multitude of probability distributions allow us to to choose those that match our data or theoretical expectations in terms of shape, location, scale.

Fitting a distribution is an art and science of utmost importance in probability modeling. The idea is you want a distribution to fit your data model "just right" without a fit that is "overfit" (*or underfit*). Over fitting models is sometimes a problem in modern data mining methods because the models fit can be too specific to a particular data set to be of broader use.

Seefeld 2007

# So why do we use them? It's all about shape and scale!

- Because they provide a usable framework for framing our questions, and allowing for parametric methods; i.e likelihood and Bayesian.
- Even if we do not know its actual distribution, it is clear frequency data is generally going to be better fit by a binomial than a normal distribution. Why?

# Why will it be a better fit?

- The binomial is **bounded** by zero and 1
- Other distributions (gamma, poisson, etc) have a lower boundary at zero.

- This provides a convenient framework for the relationship between means and variance as one approaches the boundary condition.

# Some discrete distributions (leading up to why we use negative binomial)

Binomial

Poisson

Negative-binomial

# Random variables

- This is what we want to know the probability distribution of.

- I.e. P(x|some distribution)

I will use “x” to be the random variable in each case.

# Binomial

Let's say you set up a series of enclosures. Within each enclosure you place 25 flies, and a pre-determined set of predators.
You want to know what the distribution (across enclosures) of flies getting eaten is, based on a pre-determined probability of success for a given predator species.

You can set this up as a binomial problem.

N ( R calls this size) = 25 (the total # of individuals or "trials" for predation) in the enclosure
p = probability of a successful predation "trial" (the coin toss)
x = # trials of successful predation. This is what we usually want for the probability distribution.

# Binomial

$$\binom{N}{x} p^x (1-p)^{N-x}$$

$$\binom{N}{x} = \frac{N!}{x!(N-x)!}$$

You can think of this in two ways.

A) A normalizing constant so that probabilities sum to 1.

B) # of different combinations to allow for x "successful" predation events out of N total.

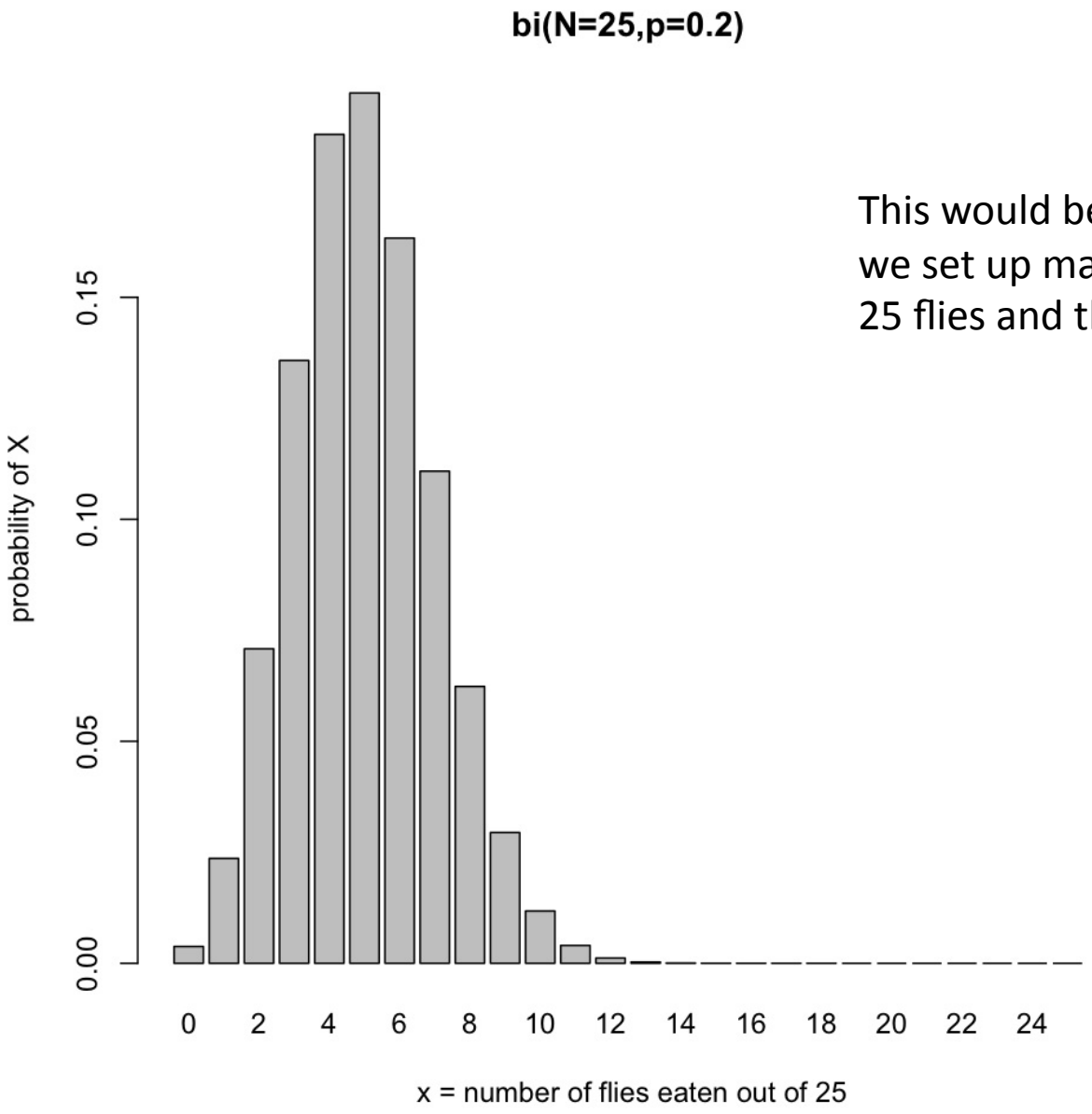
You will often see x=k and hear " N choose k"

# Example

- If predator species 1 had a per "trial" probability of successfully eating a prey item of 0.2, what would be the probability of exactly 10 flies (out of the 25) being eaten in a single enclosure.

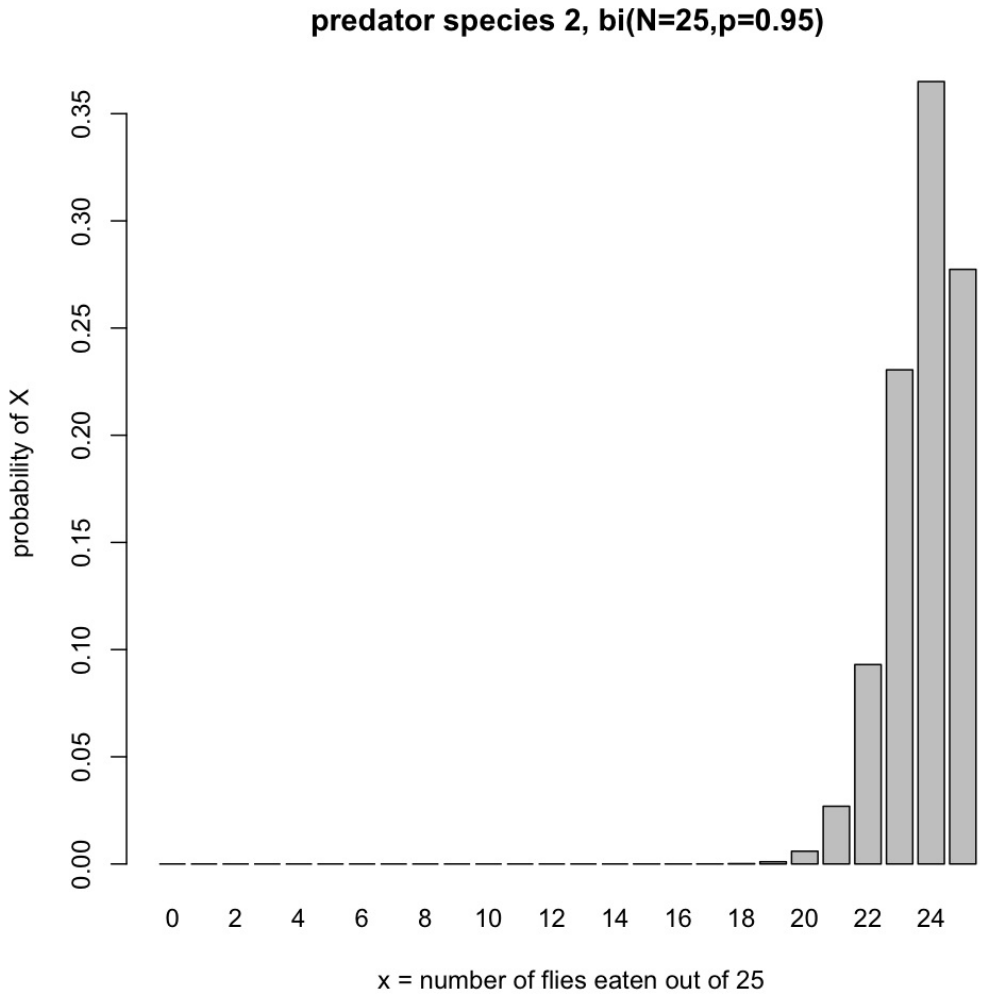
$P(x=10 \mid bi(N=25,p=0.2)) = 0.0118$

Not so high. We can look at the expected probability distribution for different values of x.




This would be the expected distribution if we set up many replicate enclosures with 25 flies and this predator.

# Predator species 2 is much hungrier….



# The rub…

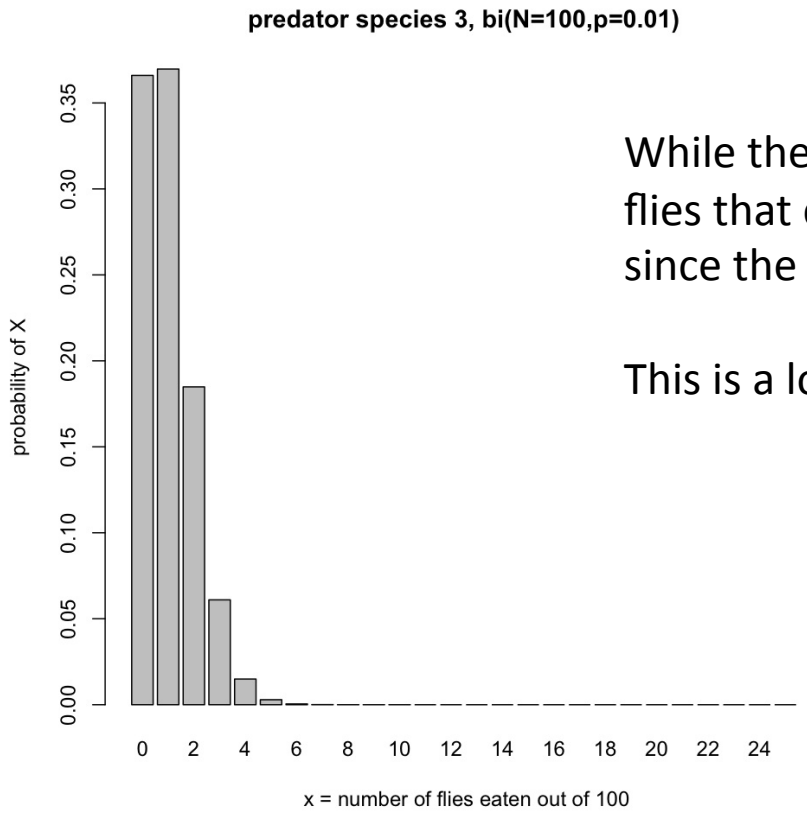
- Usually we are not interested in the probability of a given number of “successful” trials, but in estimating the parameter, p itself.

- P(D|H)
- P(x|bi(N=25,p=?)

# binomial

- $0 \leq x \leq N$
- Mean = Np (how do you estimate p)
- Var = Np(1-p)

# Let's say we had 100 flies per enclosure, and predator species 3 was really ineffective, p=0.01




While there may be a theoretical limit to the number of flies that can be eaten, practically speaking it is unlimited since the predation probability is so low.

This is a lot like the situation we have with RNA-seq data.

# Poisson

- When you have a discrete random variable where the probability of a “successful” trial is very small, but the theoretical (or practical) range is effectively infinite, you can use a poisson distribution.

- Useful for counting # of “rare” events, like new migrants to a population/year.
- # of new mutations/offspring..
- # counts of sequencing reads

# Poisson

- It is also (potentially) useful for RNA-seq data! (although we will see not very useful).

# Poisson

$$\frac{e^{-\lambda}\lambda^{x}}{x!}$$

x is our random variable (# events/unit sampling effort) – read counts for a gene in a sample
$\lambda$ Is the "rate" parameter. i.e. Expected number of reads (for a transcript) per sample
$\lambda$ is the mean and the variance!!!!

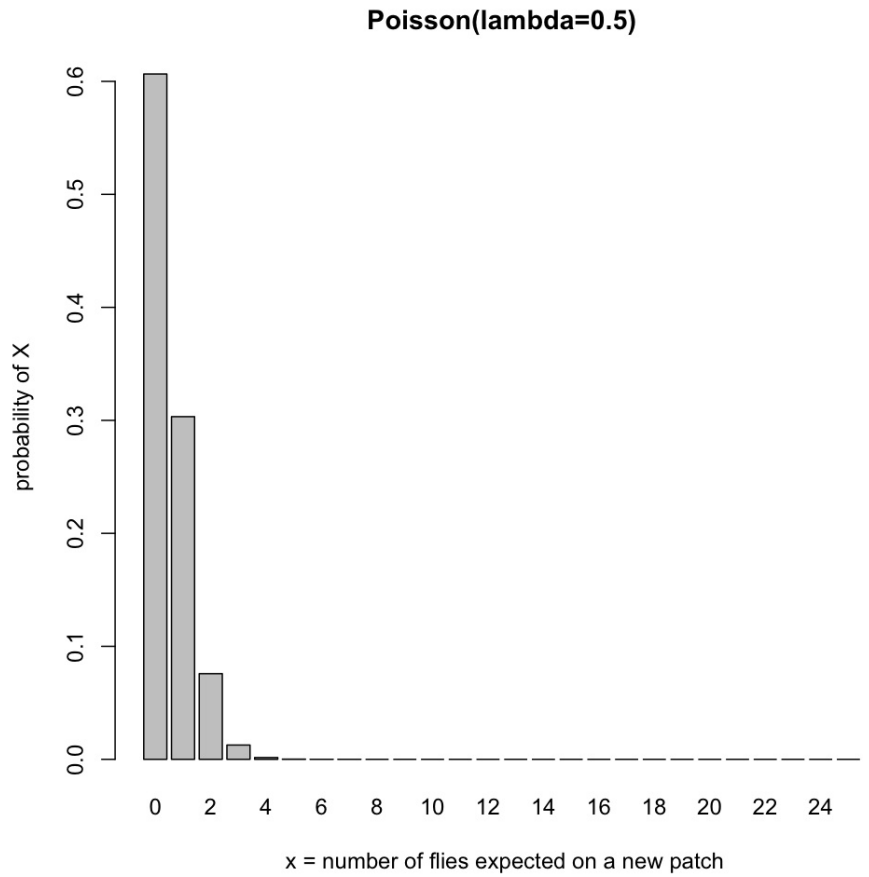
For its relation to a binomial when N is large and p is small
$\lambda = N*p$

# Poisson

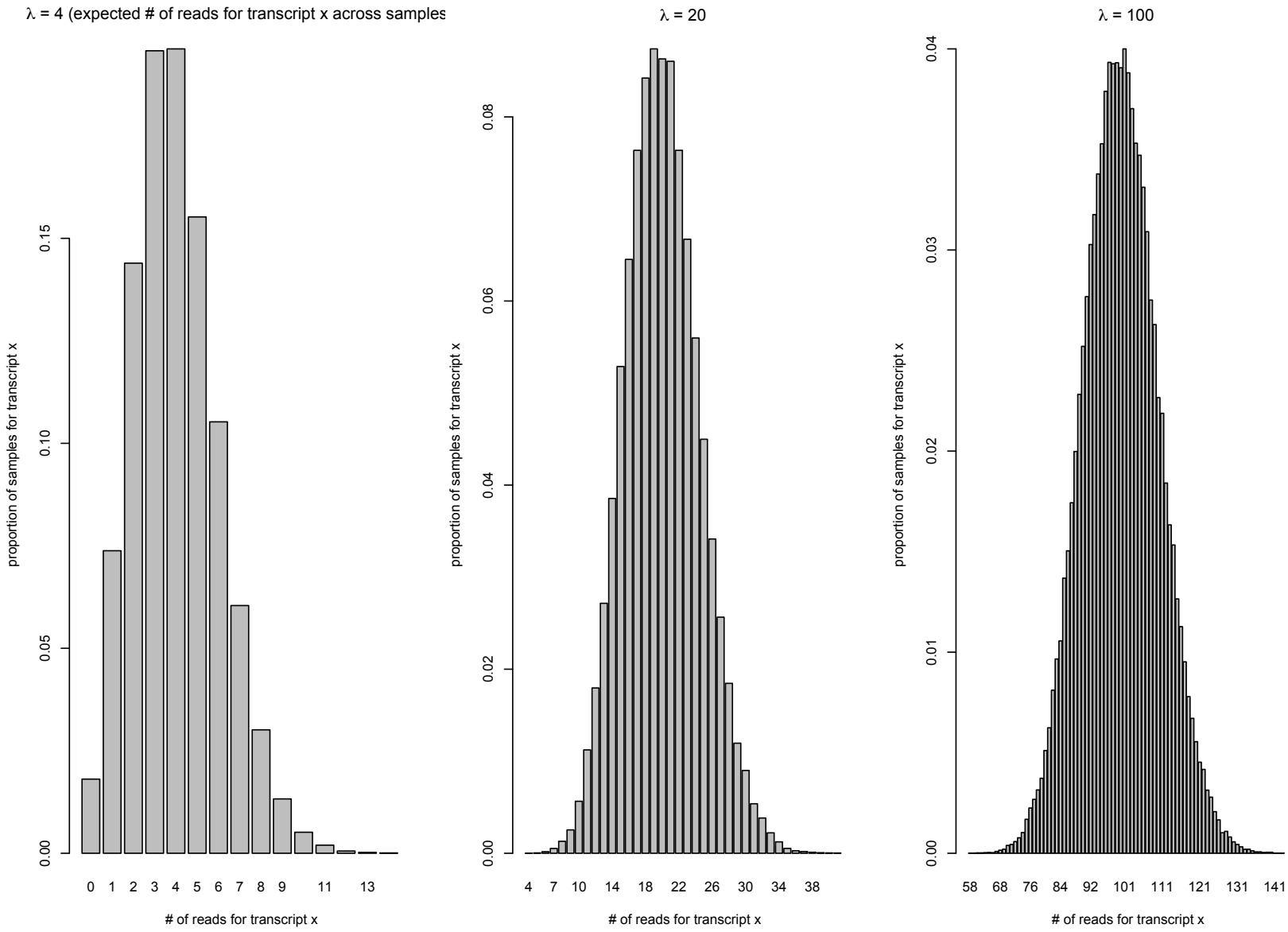
- Let's say flies disperse to colonize a new patch at a very low rate ( previous estimates suggest we will observe one fly for every two new patches we examine, $\lambda=0.5$).
- What is the probability of observing 2 flies on a new patch of land?

$P(x=2 \mid \text{poisson}(\lambda=0.5)) = 0.076$

# Probability of observing x number of flies on a patch given lambda=0.5



# What happens as lambda increases?



# Poisson mean and variance

- When lambda is small for your random variable, you will often find that your data is "over-dispersed".
- That is there is more variation that expected under Poisson (lambda).
- Similarly when lambda gets large, you will often find that there is less variation than expected under Poisson(lambda).

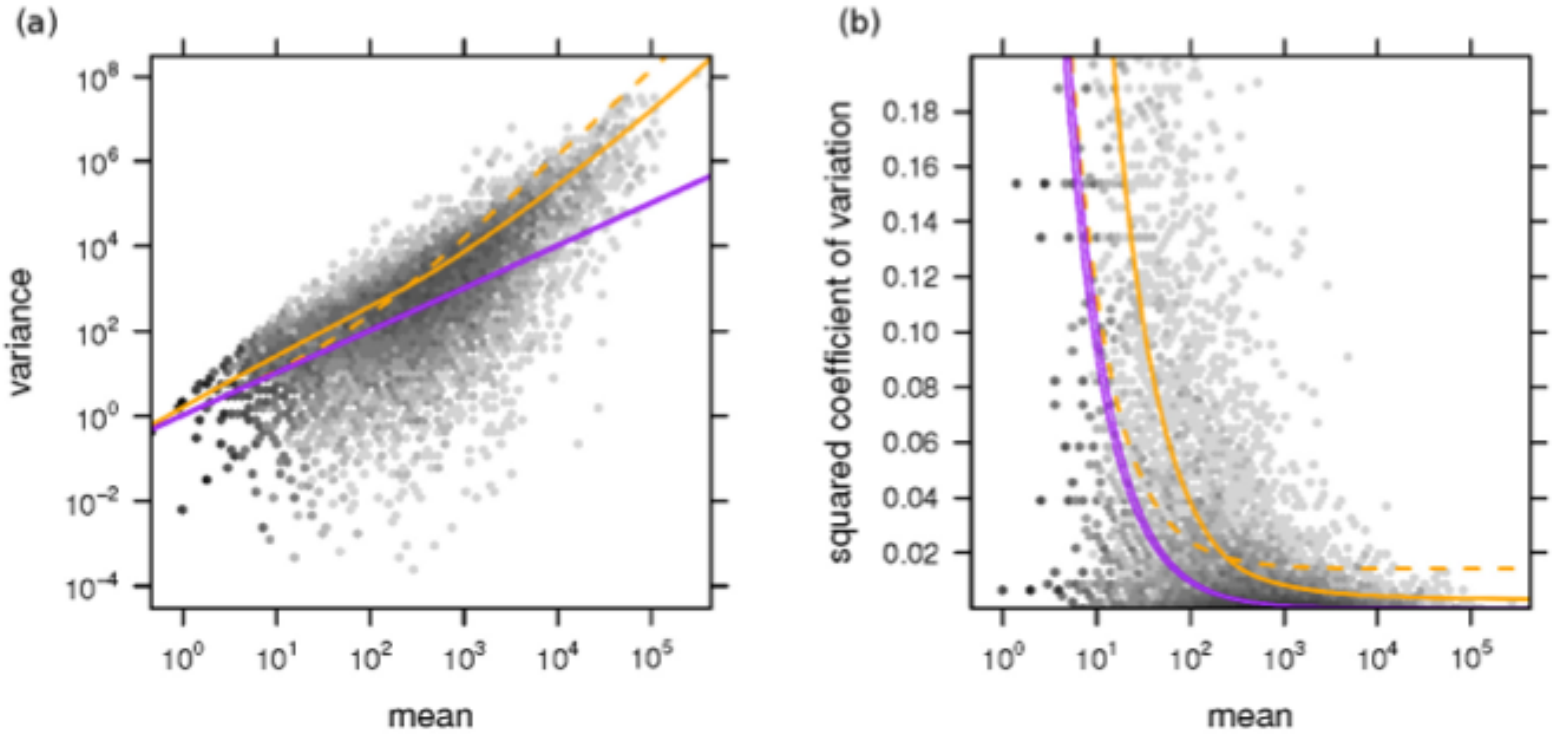



**Figure 1 Dependence of the variance on the mean for condition *A* in the fly RNA-Seq data**. (a) The scatter plot shows the common-scale sample variances (Equation (7)) plotted against the common-scale means (Equation (6)). The orange line is the fit $w(q)$. The purple lines show the variance implied by the Poisson distribution for each of the two samples, that is, $\hat{s}_j \hat{q}_{i,A}$. The dashed orange line is the variance estimate used by *edgeR*. (b) Same data as in (a), with the *y*-axis rescaled to show the squared coefficient of variation (SCV), that is all quantities are divided by the square of the mean. In (b), the solid orange line incorporated the bias correction described in Supplementary Note C in Additional file 1. (The plot only shows SCV values in the range [0, 0.2]. For a zoom-out to the full range, see Supplementary Figure S9 in Additional file 1.)

Anders and Huber 2010 Genome Biology

# Why poisson might not model sequence reads well

- Most RNA-seq data (and most count data in biology) is not modeled well by poisson because the relationships between means and variances tend to be far more complicated among (and within) biological replicates.
- It has been argued (Mortzavi et al 2008) that technical variation in RNA-seq is captured by Poisson. I have my doubts even on this.

# Quasi-poisson

- Since over-dispersion is such a common issue, a number of approaches have been developed to account for it with count data.
- One is to use a quasi-poisson.
- Instead of variance(x) = $\lambda$, it is

- Variance (x) =$\lambda\theta$
- Where $\theta$ is the (multiplicative) over-dispersion parameter.

# Negative binomial

- In biology the Neg. Binomial is mostly used like a poisson, but when you need more dispersion of x (it needs to be spread out more).

- The negative binomial is a Poisson distribution where lambda itself varies according to a Gamma distribution.

# Negative binomial

$$\text{Negative Binomial Distribution} = \frac{\Gamma(k+x)}{\Gamma(k)x!}\left(\frac{k}{k+\mu}\right)^{k}\left(\frac{\mu}{k+\mu}\right)^{x}$$

Expected number of counts =μ
Over-dispersion parameter = k

For our purposes all we care about is that

$$\text{var}(x) = \mu + k\mu^2$$

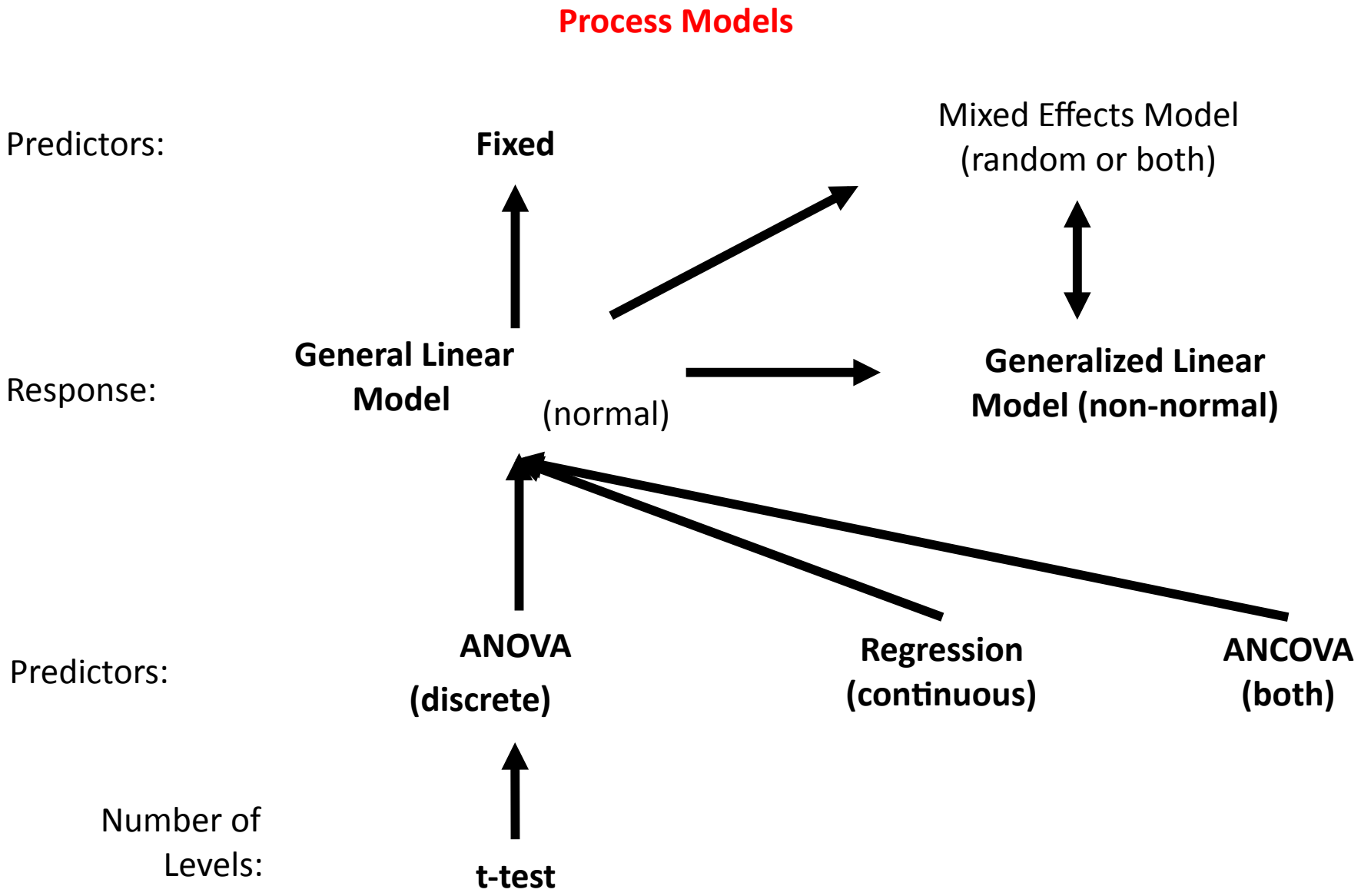
# General(ized) linear models

- For response variables that are continuous, you are likely familiar with approaches that come from the general linear model.

$$y \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right)$$

A standard linear regression (if x is continuous).
If x is discrete this would be a t-test/Anova.

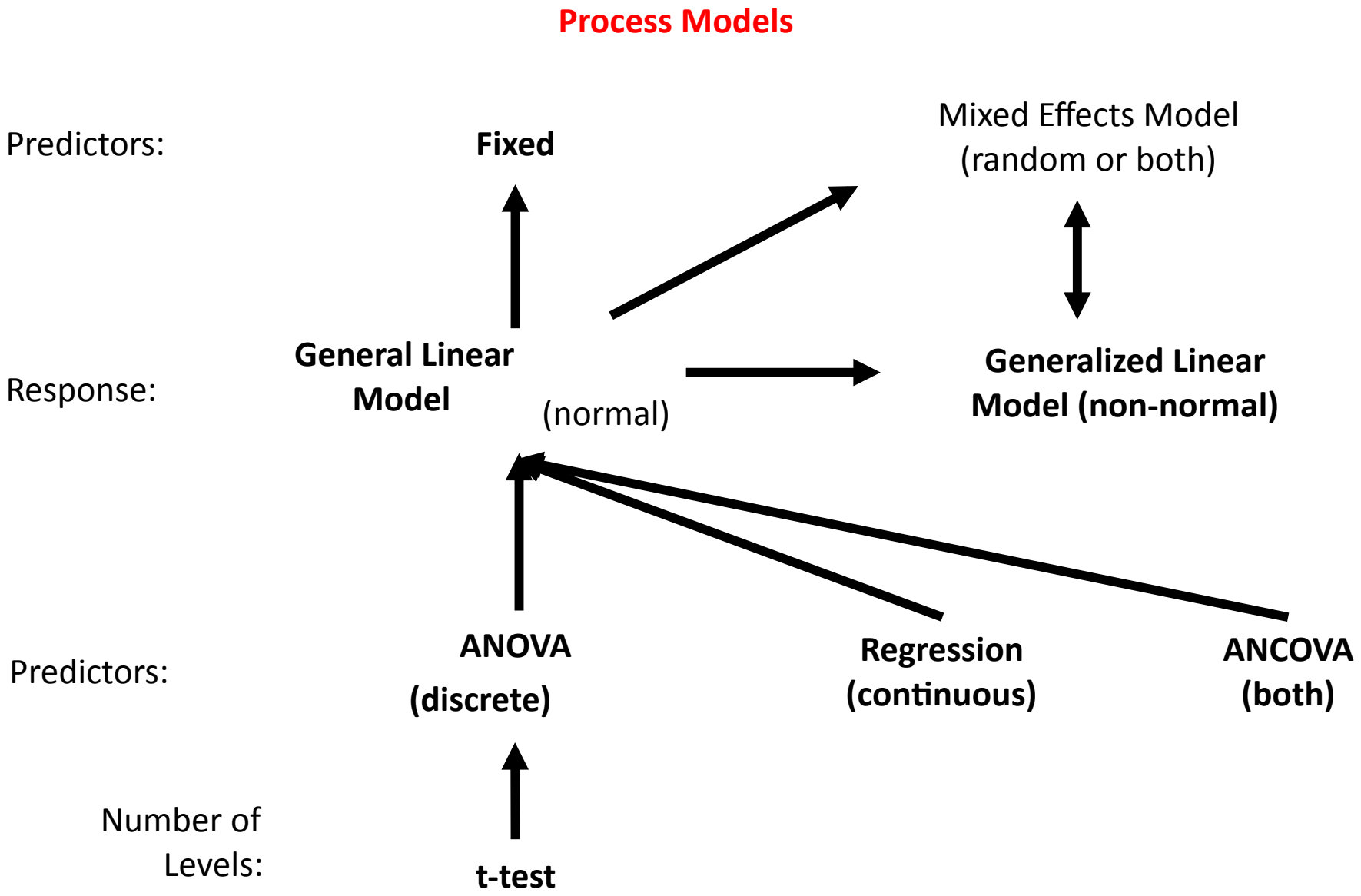
# Continuity of Statistical Approaches



# Generalized linear models

- But what do you do when your response variable is not normally distributed?
- The framework of the linear model can be extended to account for different distributions fairly easily (one major class of these is the generalized linear models).

# Continuity of Statistical Approaches



# Generalized Linear Models (GLiM)

- In many cases a **general linear model** is not appropriate because values are bounded
  - e.g. counts > 0, proportions between 0 and 1
- A generalization of linear models to include any distribution of errors from the exponential family of distributions
    - Normal, Poisson, binomial, multinomial, exponential, gamma, NOT negative binomial
- General Linear Model is just a special case of GLiM in which the errors are normally distributed
- Example, logistic regression
- We will use likelihood for parameter estimation and inference

# Generalizations of GLM

- Instead of a simple linear model:

$$Y = b_0 + b_1x_1 + b_2x_2 + e$$

  - Assume that e's are independent, normally distributed with mean 0 and **constant** variance $s^2$
  - Can solve for b's by minimizing squared e's

- GLiM considers some adjustment to the data to linearize Y - a ***link*** function

$$Y = g(\ b_0 + b_1x_1 + b_2x_2 + e)$$

or $$f(Y) = b_0 + b_1x_1 + b_2x_2 + e$$

  - For example for count data which are always positive

$$f(Y) = \log(Y) \qquad \textbf{\textit{log link}}$$

# What is a link function?

- The link function is a way of transforming the **observed response variable** (LHS).
- Goals
- 1) linearize observed response
- 2) Alter the boundary conditions of the data.
- 3) To allow for an additive model in the covariates (RHS)

# Poisson Family

- Data are counts of something (i.e. 0, 1, 2, 3, 4…)
- Number of occurrences of an event over a fixed period of time or space
- Examples…

- If the mean value is high then counts can be log-normal or normally distributed
- When mean value is low then there starts to be lots of zeros and variance depends on the mean
- If upper end is also bounded then binomial would be better

- Default link is the *log* link, variance function = $\mu$
  - *i.e., family = poisson (link = "log", variance = "mu")*
  - Other option might be the *sqrt* link

# Poisson Family

- Frequency Tables (log-linear or multinomial models)
  - Comparison of counts among categories or cells
  - Like a G-test (or $\chi^2$ test)

# Poisson and nb Family

$$\log(\hat{y}) = \beta_0 + \beta_1 x$$

*or*

$$\mu = e^{\beta_0 + \beta_1 x}$$

Essentially it means you can log transform the sequence counts and use a poisson, quasi-poisson or negative binomial to fit it
(most links are more complicated, this is nice and simple).

i.e. counts are modeled as

$$counts_{ij} \sim pois(\lambda = \mu,\ \sigma^2 = \lambda)$$

$$counts_{ij} \sim qpois(\lambda = \mu,\ \sigma^2 = \lambda\theta)$$

$$counts_{ij} \sim nb(\lambda = \mu,\ \sigma^2 = \mu + \mu^2 k)$$

# Methods using nb glm

- edgeR (but it is not default, so beware!)
- DESeq (maybe DEXseq as well?)
- BaySeq
- Limma (voom – kind of sort of…).

- However these all model the variance quite differently (how they borrow information across genes to estimate mean-variance relationships).

See Yu, Huber & Vitek 2013 (Bioinformatics) for discussion of this issue.

# Methods using poisson and quasi-poisson

- tspm (two stage poisson model)
  - Fits models with poisson first. If over-dispersed then uses a quasi-poisson.
  - Thus there are essentially two groups of genes.

# Why this is so awesome

- Since we can fit these as a generalized linear model, we can fit arbitrarily complex designs (if we have sufficient sampling to estimate the parameters).

- We can incorporate all aspects of read length, library size, lane, flow cell in addition to all of the important biological predictors (your treatments).

- NO t-tests for you!!!

# Estimating over-dispersion (variance) (or why programs seemingly doing the same thing give different results)

# Variances require lots of data to estimate well (not just for count data)

- It turns out that to estimate variances, you need a lot more replication than you do for means.
- However most RNA-seq experiments still have small numbers of biological replicates.
- So how to go about estimating variances?

# IF sample sizes are large (within and between treatments).

- Most methods do well (based on NB, quasi-P or non-parametric approaches).

- They can model individual level variances (and potentially can use resampling approaches to avoid having to make parametric assumptions).

# But if sample sizes (in terms of biological replication) is small.

- Then we have a problem.
- This is where the software really tends to differ, as they all make different assumptions about the variance, and how best to model it.

- In particular edgeR and DEseq use some methods to borrow information across genes (and have options to change this process).
- This can dramatically change the results.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology, 11(10), R106. doi:10.1186/gb-2010-11-10-r106

Anders et al (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nature Protocols, 8(9), 1765–1786

**Table 1.** Existing and proposed approaches for differential analysis of RNA-seq experiments with two conditions

| | Probability model | Estimation of dispersion | Testing | $n=1$ | Time |
|---|---|---|---|---|---|
| (a) sSeq (proposed) (this manuscript) | $X_{gij} \sim \mathcal{NB}(s_{ij}\mu_{gi}, \phi_g/s_{ij})$ | $\hat{\phi}_g^{sSeq} = \delta\xi + (1-\delta)\hat{\phi}_g^{MM}$, where $\xi$ is a common dispersion and $\delta$ is a weight | $H_0: \mu_{gA} = \mu_{gB}$ Exact test | Yes | min |
| (b) edgeR (Robinson and Smyth, 2008) | $X_{gij} \sim \mathcal{NB}(m_{ij}p_{gi}, \phi_g)$ | $\hat{\phi}_g^{edgeR}$ maximize linear combination of per-gene and common-dispersion conditional likelihoods | $H_0: p_{gA} = p_{gB}$ Exact or GLM-based test | Yes* | min |
| (c) DESeq (Anders and Huber, 2010) | $X_{gij} \sim \mathcal{NB}(s_{ij}\mu_{gi}, \phi_{gi})$ | $\hat{\phi}_{gi}^{DESeq} = \left(\hat{V}_{gi} - \hat{\mu}_{gi}\frac{1}{n_i}\sum_j \frac{1}{s_{ij}}\right)/\hat{\mu}_{gi}^2$ $\hat{V}_{gi}$ is estimated as function of the mean | $H_0: \mu_{gA} = \mu_{gB}$ Exact or GLM-based test | Yes | min |
| (d) baySeq (Hardcastle and Kelly, 2010) | $X_{gij} \sim \mathcal{NB}(N_{ij}p_{gi}, \phi_g)$ Empirical priors on sets of parameters | $\hat{\phi}_g^{baySeq}$ maximize per-gene integrated quasi-likelihood | $H_0: p_{gA} = p_{gB}$ Posterior probability cutoff | Yes | h |
| (e) BBSeq (Zhou *et al.*, 2011) | $X_{gij} \sim \mathcal{B}inom(p_{gi}, N_{ij})$ $p_{gi} \sim \mathcal{B}eta$, $logitE\{p_{gi}\} = Z\beta$, $\mathrm{V}(p_{gi}) = \mathrm{E}(p_{gi})(1-\mathrm{E}(p_{gi}))\phi_g$ | $\hat{\phi}_g^{BBSeq}$ maximize per-gene marginal likelihood; is a free parameter or a function of the mean | $H_0: \beta = 0$ Wald test | Yes | h |
| (f) SAMseq (Li and Tibshirani, 2011) | Non-parametric | | $H_0$: same distributions A and B Wilcoxon test & resampling | No | min |

(a) $s_{ij}$ is the size factor for sample $j$ in condition $i$ as defined in (Anders and Huber, 2010). $\mu_{gi}$ is the expected normalized expression of gene $g$ for a sample in condition $i$. $\hat{\phi}_g^{MM}$ is the per-gene dispersion estimate using the method of moments in Equation (6).
(b) $m_{ij}$ is the 'effective' library size. $p_{gi}$ is the probability that a read in $i$ maps to gene $g$. *Up to v2.4.6.
(c) $\phi_{gi}$ is gene- and condition-specific dispersion. $\hat{\mu}_{gi}$ and $\hat{V}_{gi}$ can be estimated by the method of moments or by the Cox-Reid corrected Maximum Likelihood.
(d) $N_{ij}$ is the size of the library $i$ from condition $j$. $p_{gi}$ is as in (b).
(e) $p_{gi}$ is as in (b). $N_{ij}$ is as in (d). $\beta$ is the coefficient of the linear predictor associated with an indicator $Z$ of conditions. Column 'Time' is the run time for the experimental datasets in Section 4 on a laptop computer.

Yu et al (2013). Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics, 29(10), 1275–1282.*

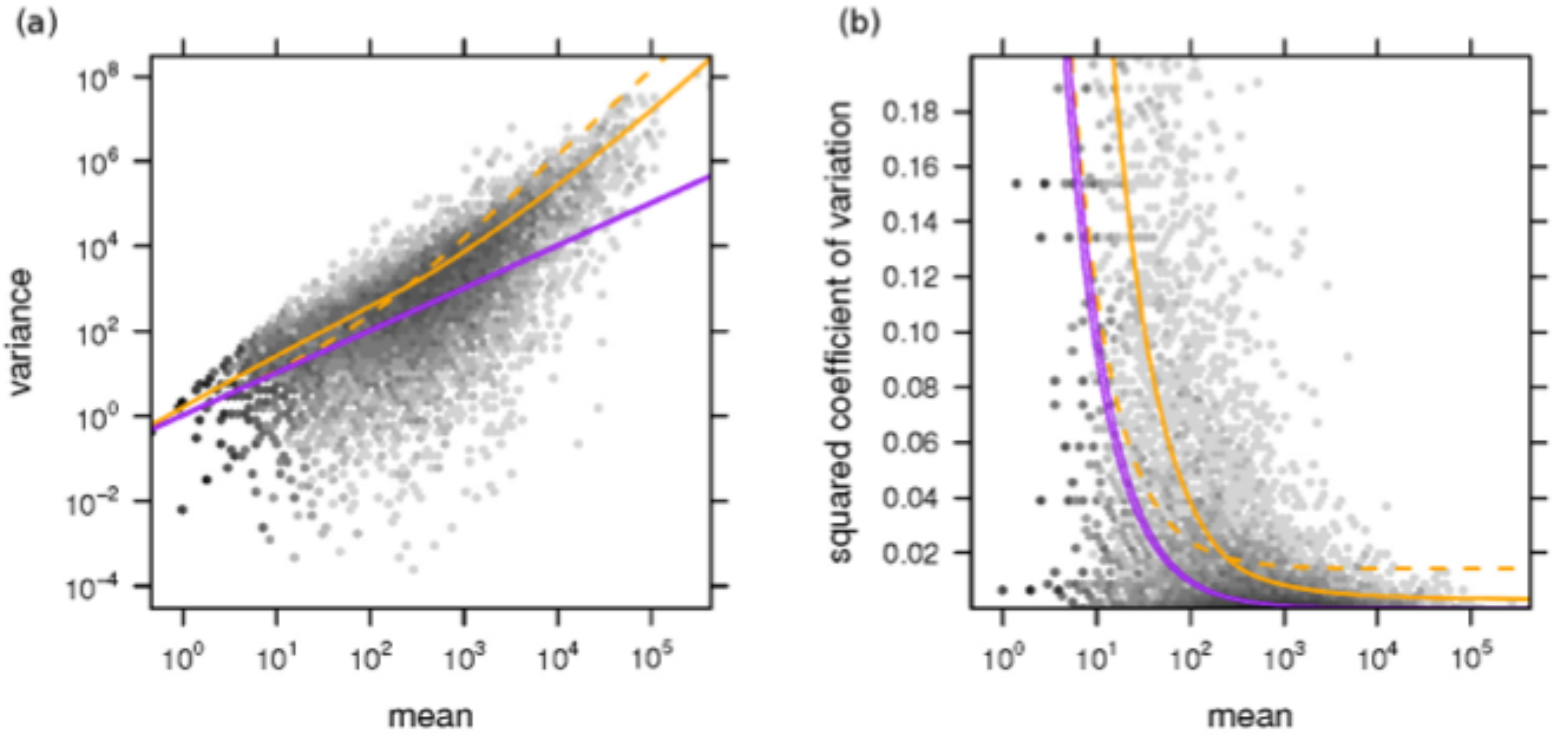
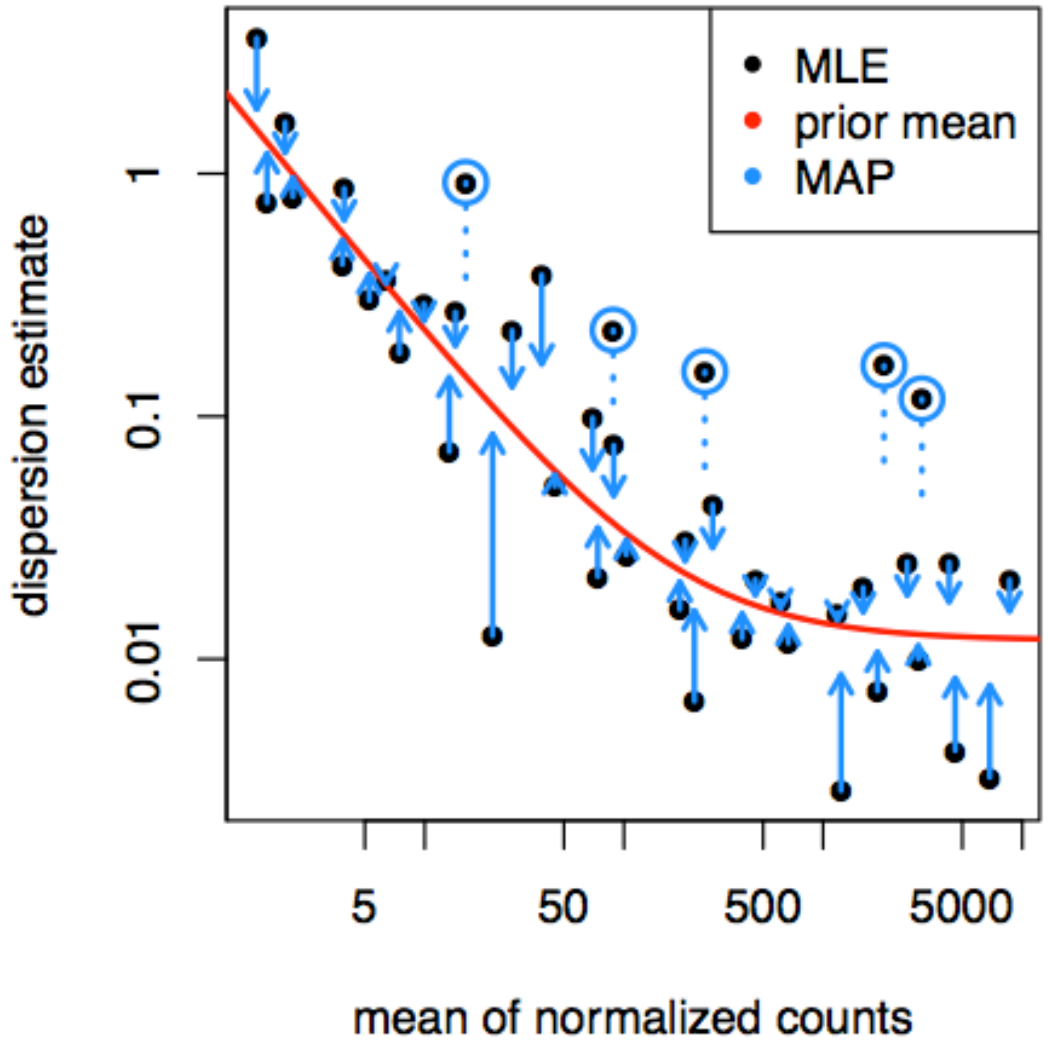



**Figure 1 Dependence of the variance on the mean for condition *A* in the fly RNA-Seq data**. (a) The scatter plot shows the common-scale sample variances (Equation (7)) plotted against the common-scale means (Equation (6)). The orange line is the fit $w(q)$. The purple lines show the variance implied by the Poisson distribution for each of the two samples, that is, $\hat{s}_j \hat{q}_{i,A}$. The dashed orange line is the variance estimate used by *edgeR*. (b) Same data as in (a), with the *y*-axis rescaled to show the squared coefficient of variation (SCV), that is all quantities are divided by the square of the mean. In (b), the solid orange line incorporated the bias correction described in Supplementary Note C in Additional file 1. (The plot only shows SCV values in the range [0, 0.2]. For a zoom-out to the full range, see Supplementary Figure S9 in Additional file 1.)

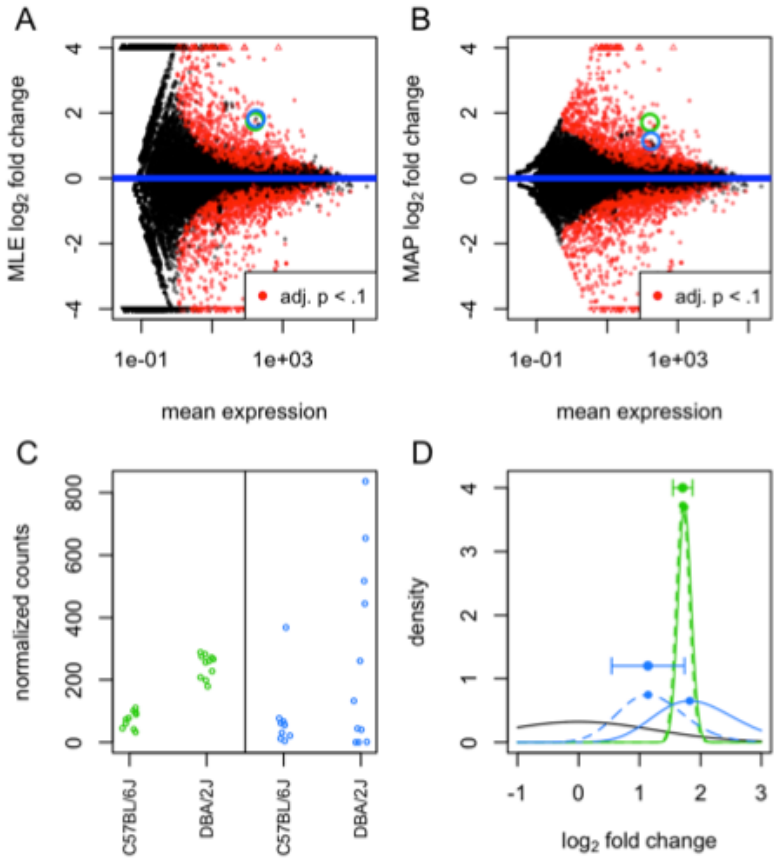
Anders and Huber 2010

# Let’s think about this.




Love, Huber & Anders 2014 BioRXiV doi: 10.1101/002832

# We can also “shrink” estimates based on over-dispersion….



# Take home

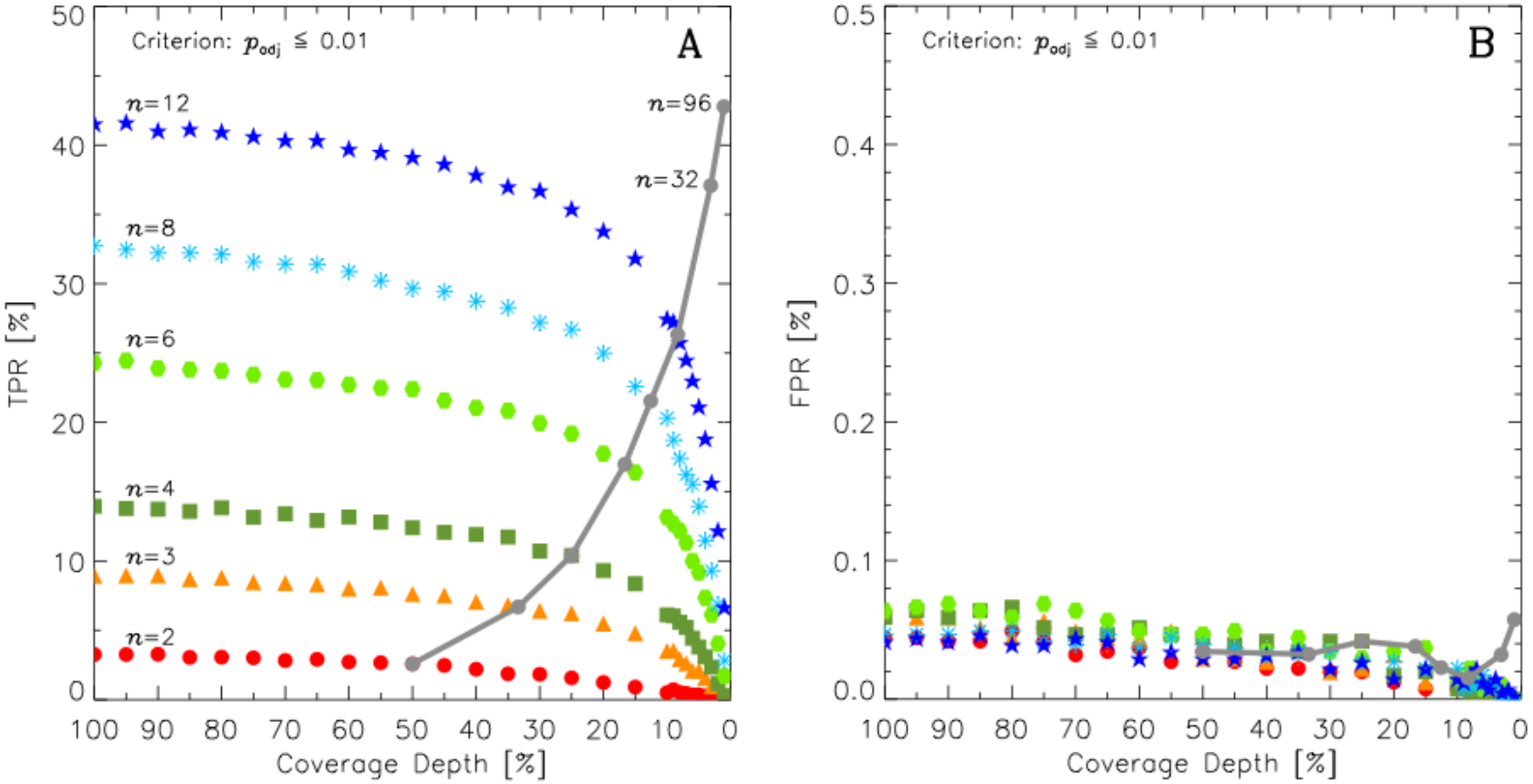
- With small sample sizes, the methods use different approaches to get gene-wise over-dispersion (based on all data).

- EdgeR is more powerful (more significant hits) than DGE generally. But much more susceptible to false positives due to outliers.
- DGE2 “should” be somewhere in the middle.

# Biological replication gives far more statistical power than increased sequencing depth within a biological sample!!!!

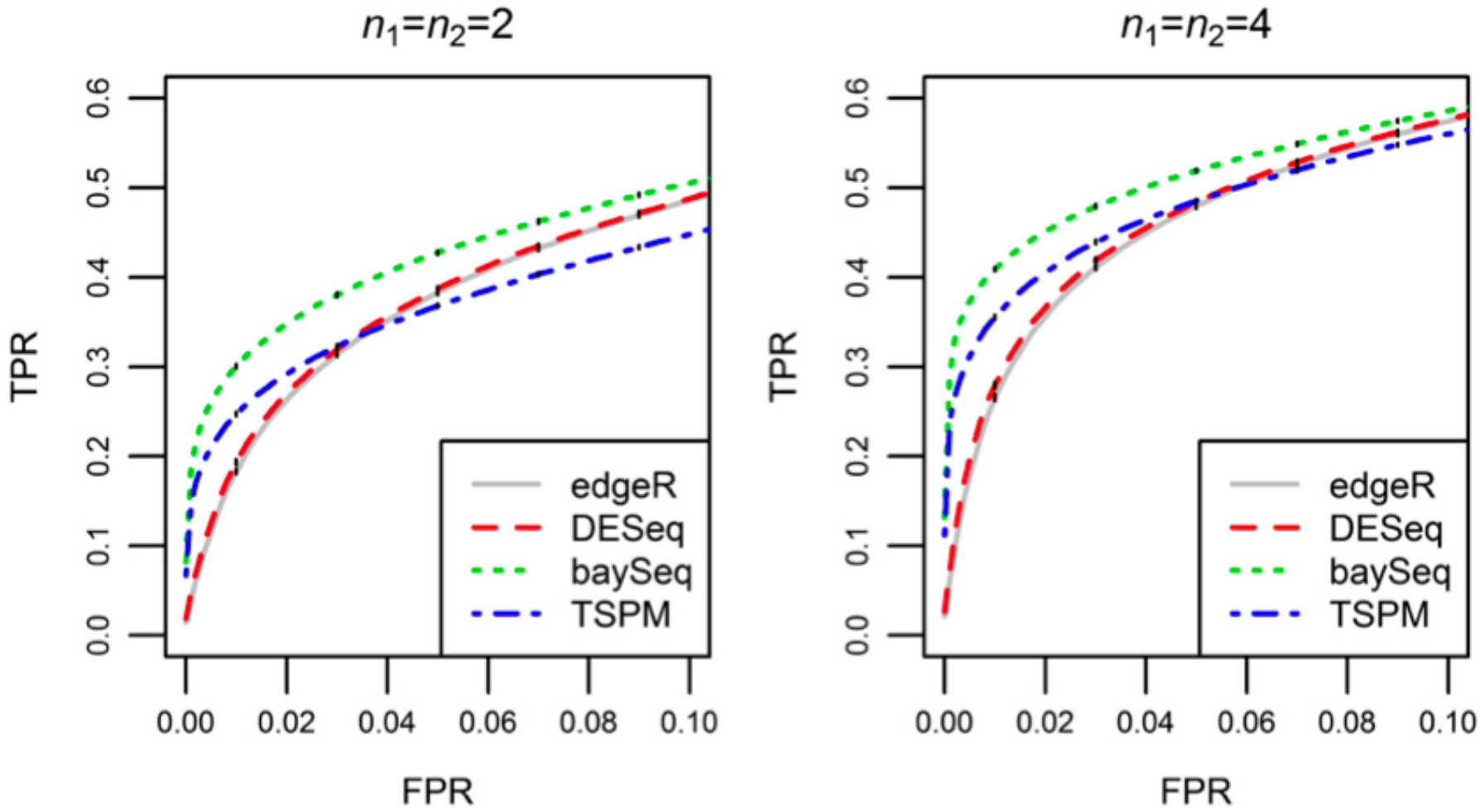
- Sequencing (and library prep) costs are still sufficiently expensive that most experiments use small numbers of biological replicates.
- Given the additional costs of library costs (~225$/ sample at our facility), many folks go for increased depth instead of more samples.
- For a given level of sequencing depth (total) for a treatment, it is far better to go for more biological replicates, each at lower sequencing depth (rather than fewer replicated at higher sequencing depth).

# Biological replication gives far more statistical power than increased sequencing depth within a biological sample!!!!



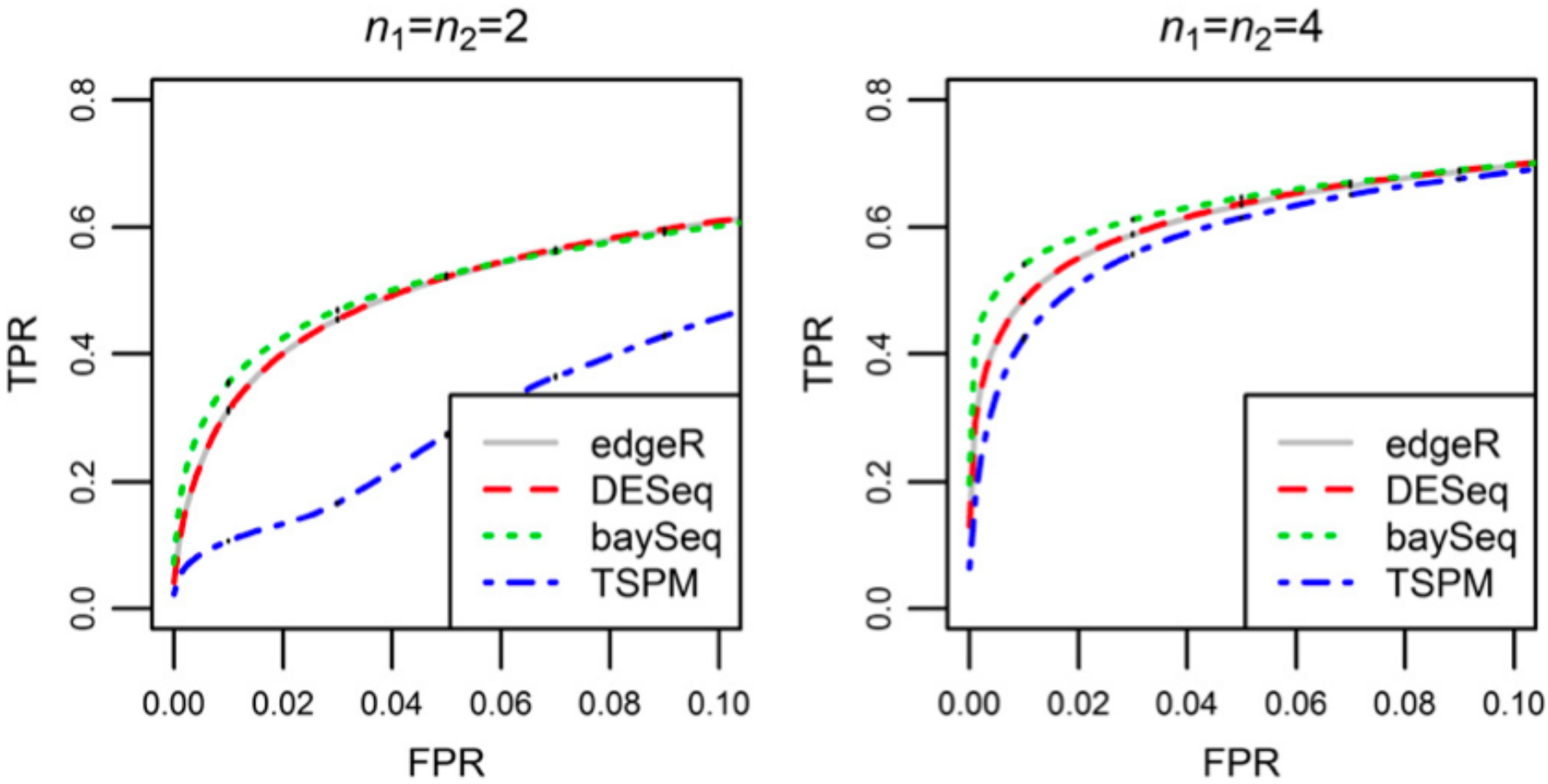

Robles et al. 2012

# How do the methods compare in simulation?



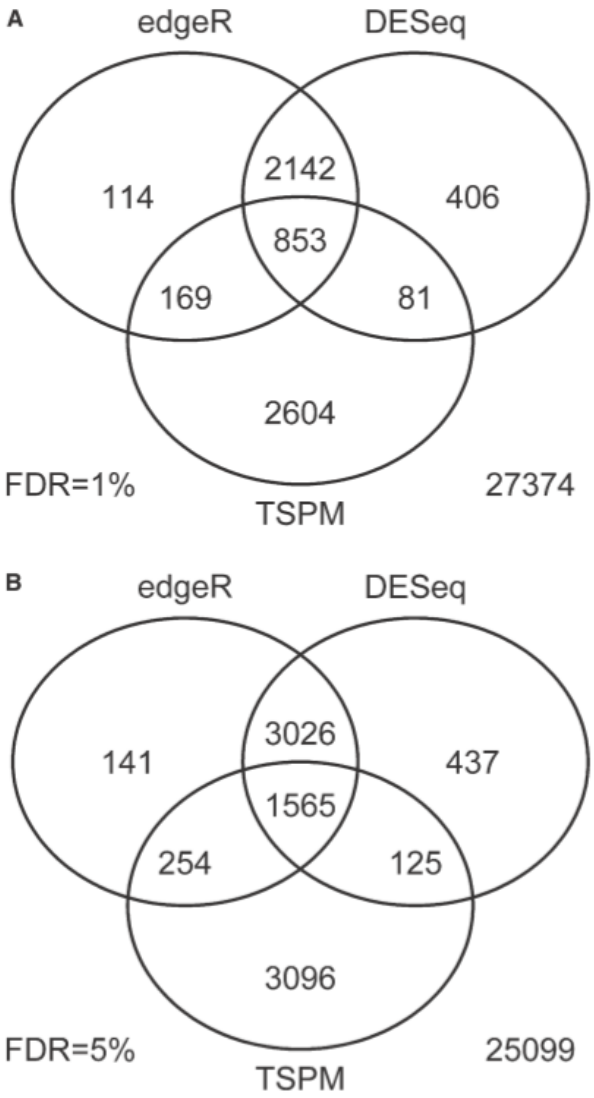

Kvam et al. 2012

# How do the methods compare in



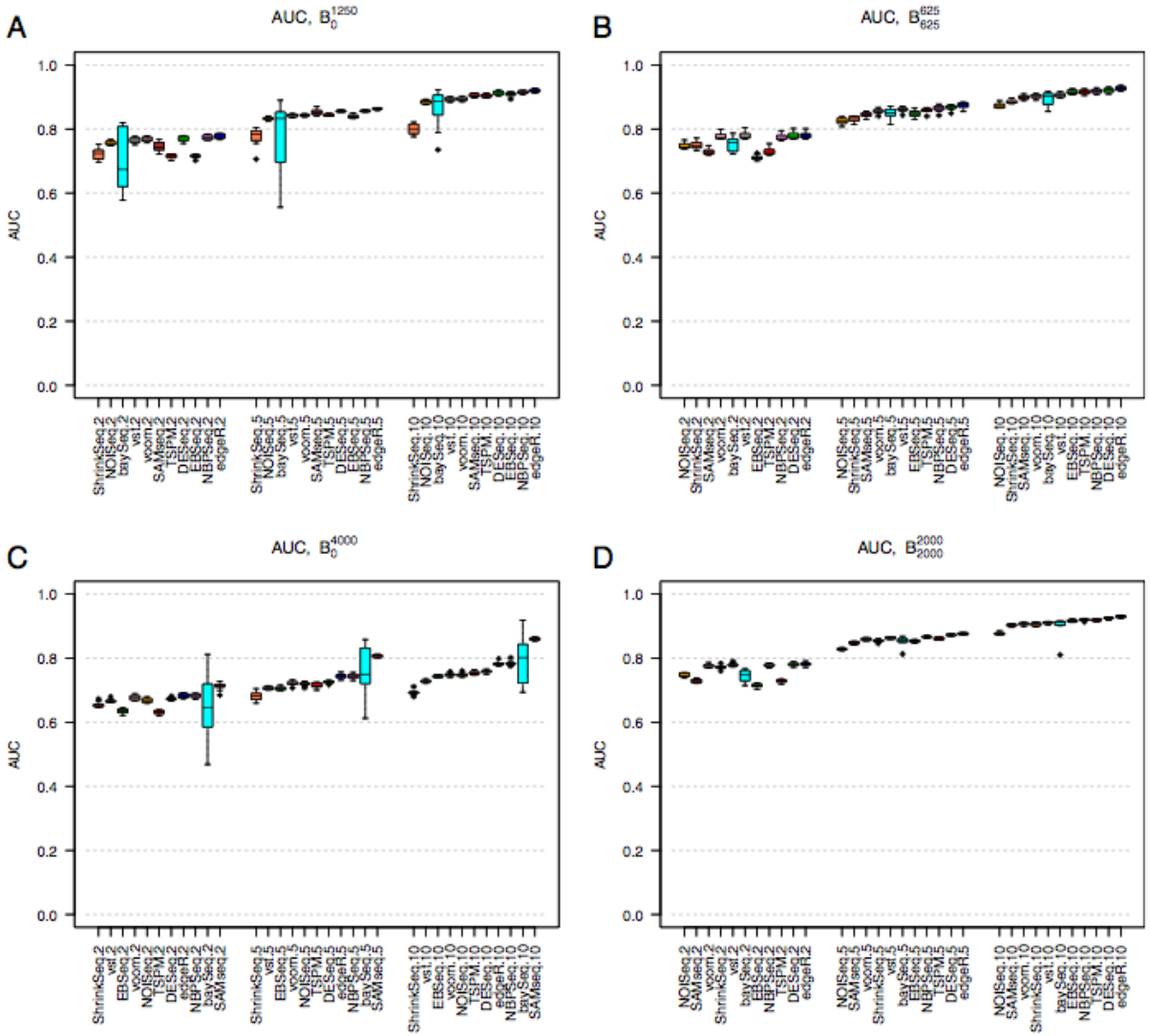

Kvam et al. 2012

# How do the methods compare for real data?




Kvam et al. 2012

# How do the methods compare in a different set of simulations?




Will explain ROC (receiver operator curves) and the area under curves on board.

Soneson 2012

# Differential expression (subset, see my github page)

- DEseq (http://www.ncbi.nlm.nih.gov/pubmed/20979621)
- DeSeq2
- Limma/voom
- EDGE-R
- Sailfish (kmer approach)
- EBseq (RSEM/EBseq)
- Beers simulation pipeline(http://www.cbil.upenn.edu/BEERS/)
- DEXseq (http://bioconductor.org/packages/release/bioc/html/DEXSeq.html)

# References


- Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., & Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics, 13, 484. doi: 10.1186/1471-2164-13-484

- Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics, 11, 94. doi:10.1186/1471-2105-11-94

- Kvam, V. M., Liu, P., & Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. American Journal Of Botany, 99(2), 248–256. doi:10.3732/ajb.1100340

- Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics, 14, 91. doi:10.1186/1471-2105-14-91

- Wagner, G. P., Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in biosciences = Theorie in den Biowissenschaften, 131(4), 281–285. doi:10.1007/ s12064-012-0162-3

- Vijay, N., Poelstra, J. W., Künstner, A., & Wolf, J. B. W. (2012). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Molecular Ecology. doi:10.1111/mec.12014