RNA-seq: quantification and models for assessing differential expression (at least for some approaches)

Ian Dworkin
NGS2016
@IanDworkin

What we will cover today

- Absolute fundamentals of experimental design
- 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).
- A bit about dealing with multiple comparisons (if we have time).

Goals

I am not planning on trying to provide any sort of overview of statistical methods for genomic data. Instead I am going to provide a few short ideas to think about.

Statistics (like bioinformatics) is a rapidly developing area, in particular with respect to genomics. Rarely is it clear what the "right way" to analyze your data is.

Instead I hope to aid you in using some common sense when thinking about your experiments for using high throughput sequencing.

Caveats

- There are whole courses on proper experimental design and statistics. Great books too. This material in Bio720 is not enough!

- For experimental design I highly recommend:
  - Quinn & Keough: Experimental Design and data analysis for biologists.

http://www.amazon.com/Experimental-Design-Data-Analysis-Biologists/dp/0521009766/

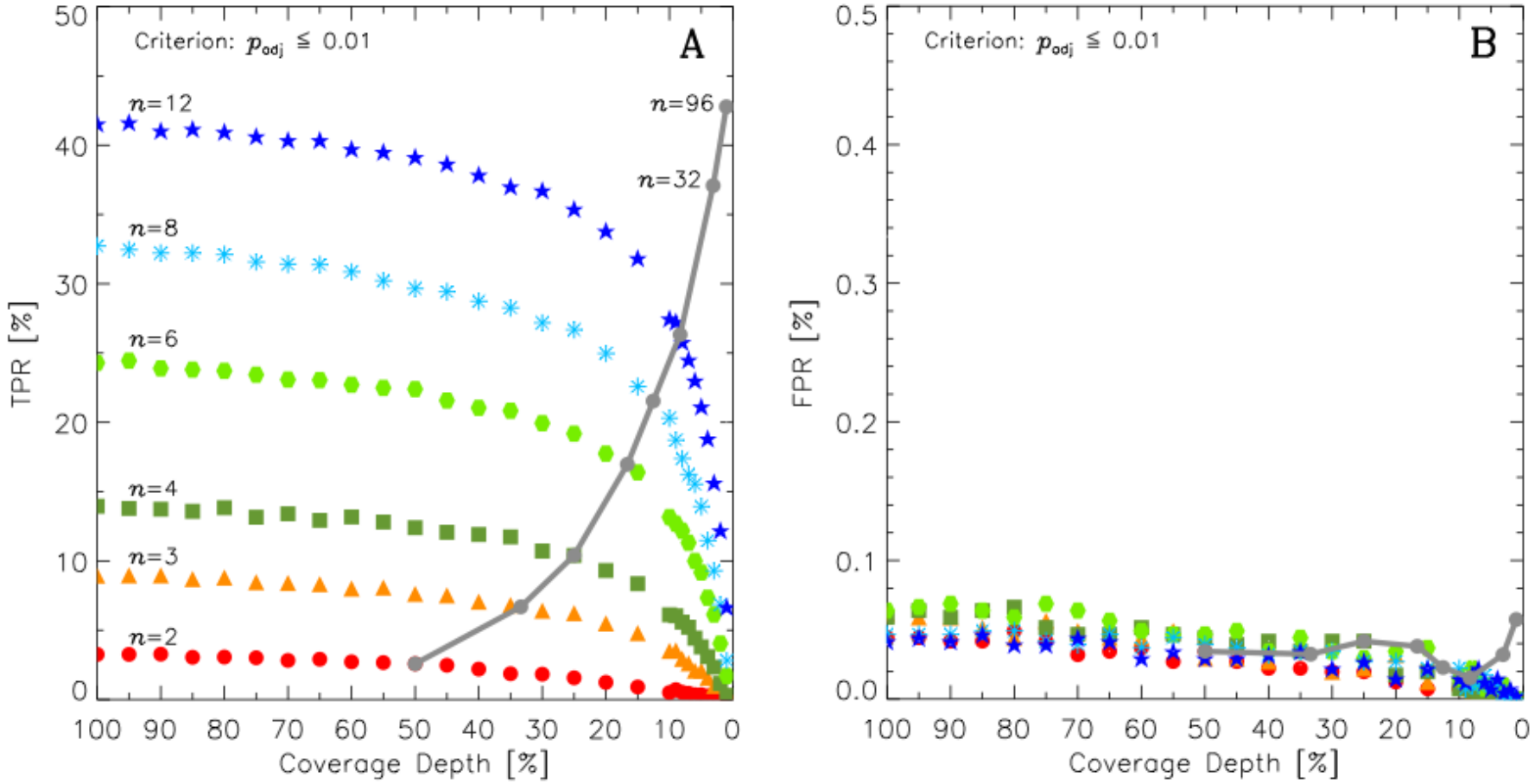
The basics of experimental design

- There are a few basic points to always keep in mind:
  - Biological replication (as much as you can afford) is extremely important. To robustly identify differentially expressed (DE) genes requires statistical powers.
    - (note: this is not how many reads you have for a gene within a sample, but how many biologically/statistically independent samples per treatment).
  - Technical replication does not help with statistical power (i.e. don’t split a single sample and run as two libraries).

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

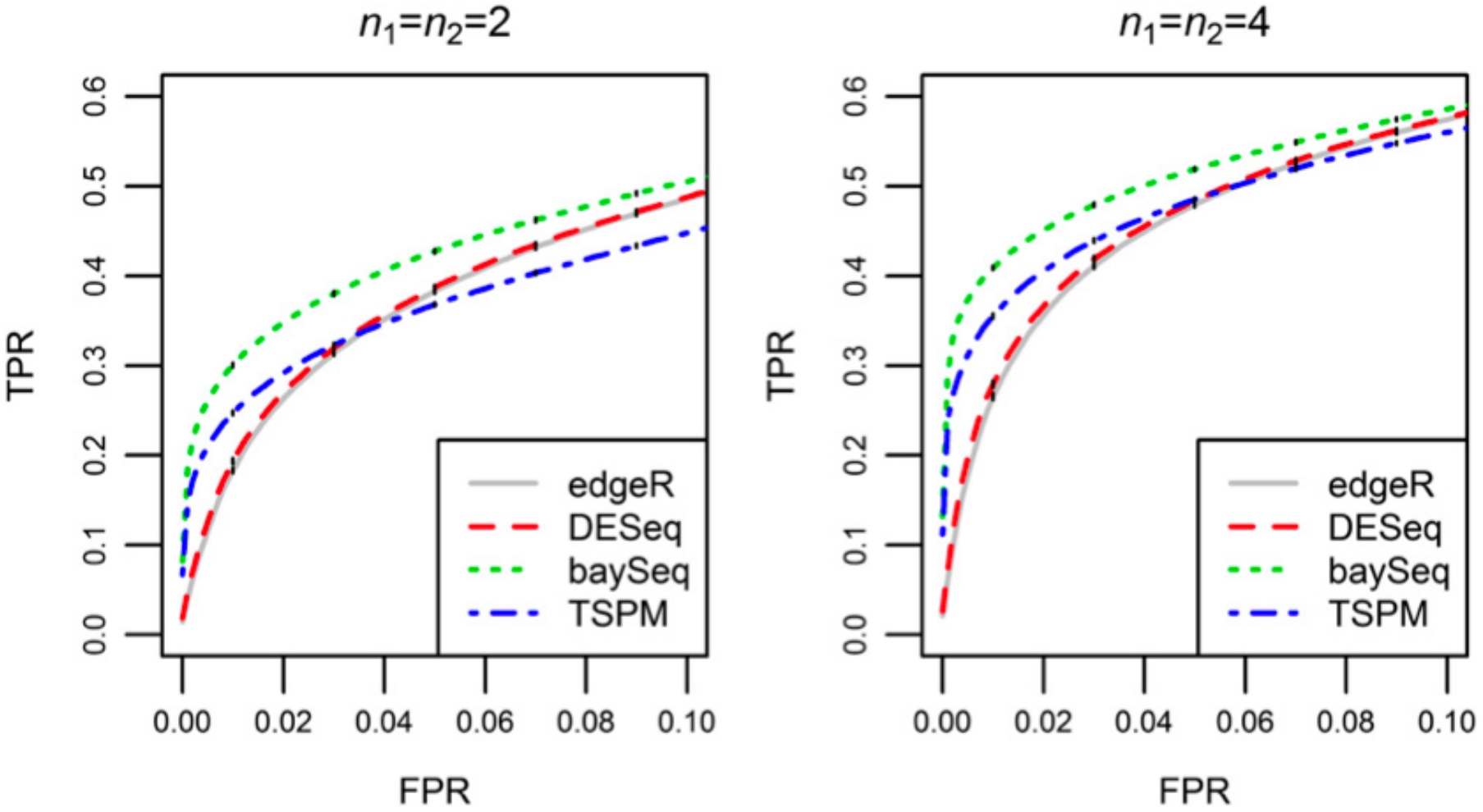
- Sequencing (and library prep) costs are still sufficiently expensive that most experiments use small numbers of biological replicates.
- Given the additional costs of library costs (~225$/sample at our facility), many folks go for increased depth instead of more samples.
- For a given level of sequencing 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

The basics of experimental design

- There are a few basic points to always keep in mind:
  - Biological replication.
  - Design your experiment to avoid ***confounding*** your different treatments (sex, nutrition) with each other or with technical variables (lane within a flow cell, between flow cell variation).
    - Make diagrams/tables of your experimental design, or use a randomized design.

The basics of experimental design

- There are a few basic points to always keep in mind:
  - Biological replication.
  - Design experiment to avoid *confounding* variables.
  - Sample individuals (within treatment) randomly!

Useful references

Paul L. Auer and R.W. Doerge 2010. Statistical Design and Analysis of RNA-Seq Data. Genetics. 10.1534/genetics.110.114983
PMID: 20439781

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

Designing your experiment before you start.

Sampling

Replication

Blocking

Randomization

Over all we are going to be thinking about how to **avoid Confounding** sources of variation in the data.

All of these are larger topics that are part of **Experimental Design**.

Sampling

Sampling

Replication

Blocking

Randomization

Sampling design is all about making sure that when you “pick” (sample) observations, you do so in a **random** and **unbiased** manner.

Proper sampling aims to control for unknown sources of variation that influence the outcome of your experiments.

This seems reasonable, and often intuitive to most experimental biologists, but it can be very insidious.
Whiteboard…

Sampling

Sampling

Replication

Blocking

Randomization

Biological replicates Not technical ones.

- There is little purpose in using technical replication (i.e. same sample, multiple library preps) from a given biological sample UNLESS part of your question revolves around it.

- Focus on biological variability. While you are confounding some sources of technical and biological variability, we already know a lot about the former, and little about the latter (in particular for your system).

Replication

Sampling

R**eplication**

Blocking

Randomization

Imagine you have an experiment with one factor (sex), with two treatment levels (males and females).

You want to look for sex specific differences in the brains of your critters based on transcriptional profiling, so you decide to use RNA-seq.

Perhaps you have a limited budget so you decide to run one sample of male brains, and one sample of female brains, each in one lane of a flow cell.

What (useful) information can you get out of this?

Not much (but there may be some). Why?

Replication

Sampling

R**eplication**

Blocking

Randomization

Why?

No replication. How will you know if the differences you observe are due to differences in males and females, random (biological) differences between individuals, or technical variation due to RNA extraction, processing or running the samples on different lanes.

All of these sources of variation are confounded, and there are no particularly good ways of separating them out.

But there are lots of sources of variation, so how do we account for these?

Replication

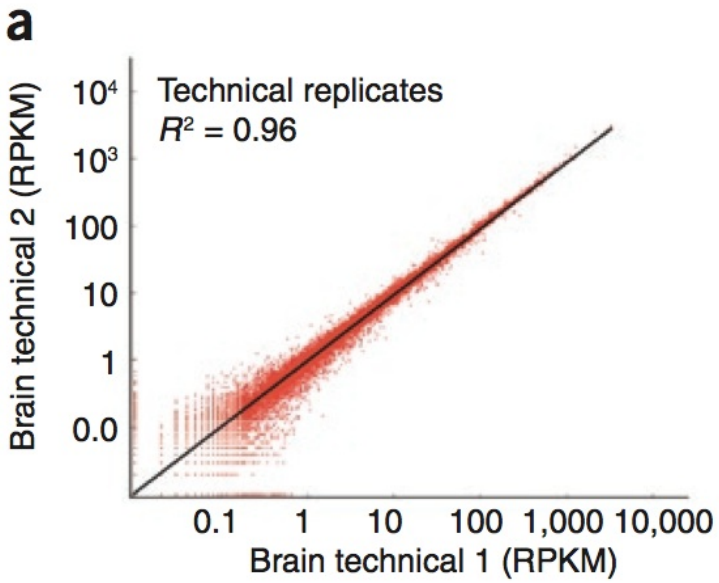
Sampling

Replication

Blocking

Randomization

To date, several studies have suggested that “technical” replicates for RNA-seq show very little variation/ high c



Mortazavi et al. 2008

How might such a statement be misleading about variation?

Replication

Sampling

Replication

Blocking

Randomization

This study looked at a single source of technical variation.

Running exactly the same sample on two different lanes on a flow cell.

This completely ignores other sources of "technical variation"
variation due to RNA purification
variation due to fragmentation, labeling, etc..
lane to lane variation
flow cell to flow cell variation

All of these may be important (although unlikely interesting) sources of variation…

However…..

Replication

Sampling

Replication

Blocking

Randomization

Many studies have ignored the BIOLOGICAL SOURCES of VARIATION between replicates. In most cases biological variation between samples (from the same treatment) are generally far more variable than technical sources of variation.

While it would be nice to be able to partition various sources of technical variation (such as labeling, RNA extraction), it often too expensive to perform such a design (see white board).

IF you have limited resources, it is generally far better to have biological replication (independent biological samples for a given treatment) than technical replication.

Does these lead to confounded sources of variation?

Blocking

Sampling

Replication

Blocking

Randomization

Blocks in experimental design represent some factor (usually something not of major interest) that can strongly influence your outcomes. More importantly it is a factor which you can use to group other factors that you are interested in.

For instance in agriculture there is often plot to plot variation. You may not be interested in the plot themselves but in the variety of crops you are growing.

But what would happen if you grew all of strain 1 on plot 1 and all of strain 2 on plot 2?

Whiteboard.

These plots would represent blocking levels

Blocking

Sampling

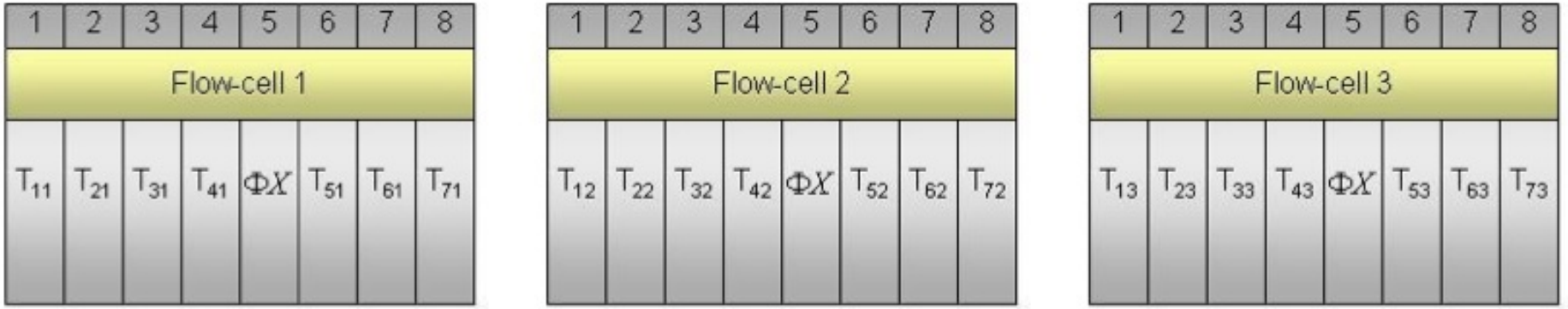
Replication

Blocking

Randomization

In genomic studies the major blocking levels are often the slide/chip for microarrays (i.e. two samples /slide for 2 color arrays, 16 arrays/slide for Illumina arrays).

For GAII/HiSeq RNA-seq data the major blocking effect is the flow cell itself and lanes within the flow cell.



Auer and Doerge 2010

Blocking

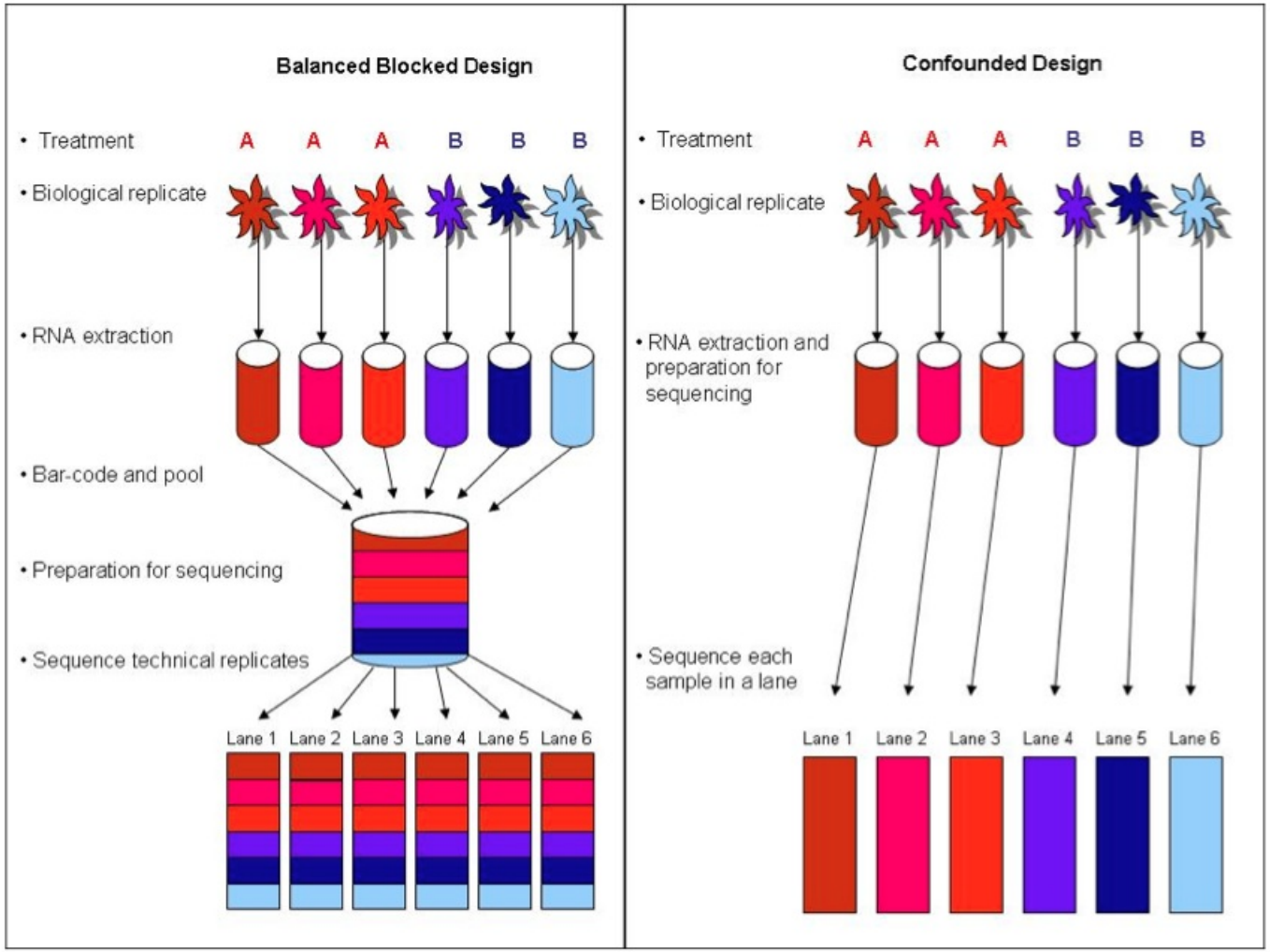
Incorporating lanes as a blocking effect

Sampling

Replication

Blocking

Randomization



Auer and Doerge 2010

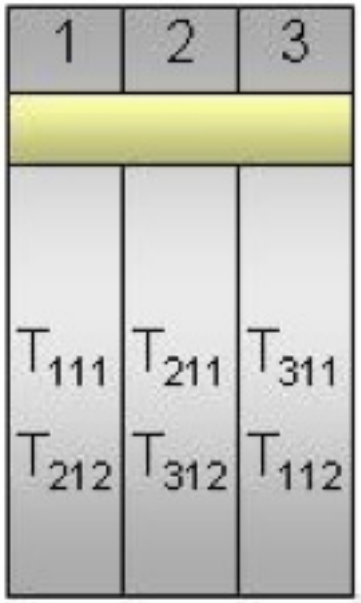
Blocking designs

Sampling

Replication

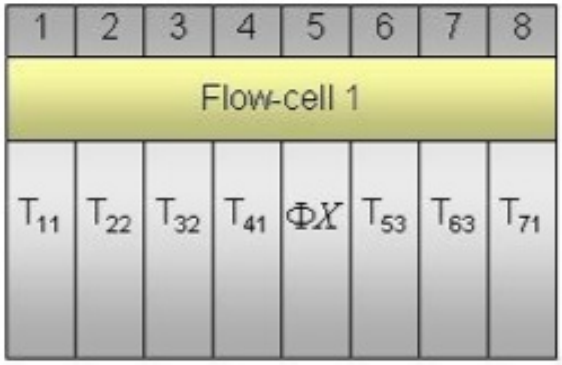
Blocking

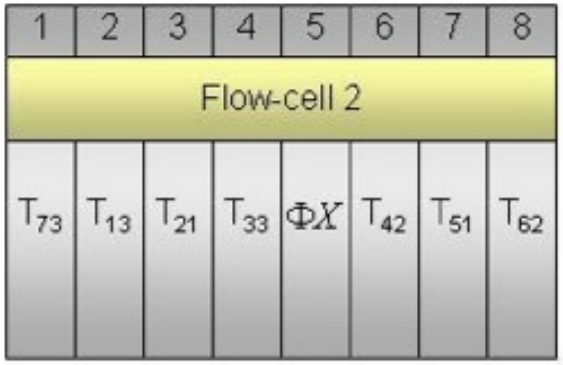
Randomization

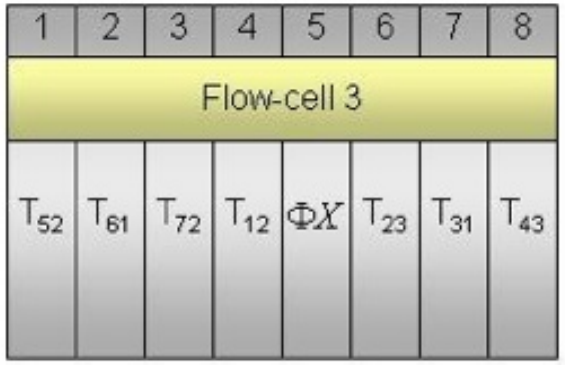


Balanced **I**ncomplete **B**locking **D**esign (BIBD)

Let's dissect these subscripts.







Balanced for treatments across flow cells.. Randomized for location

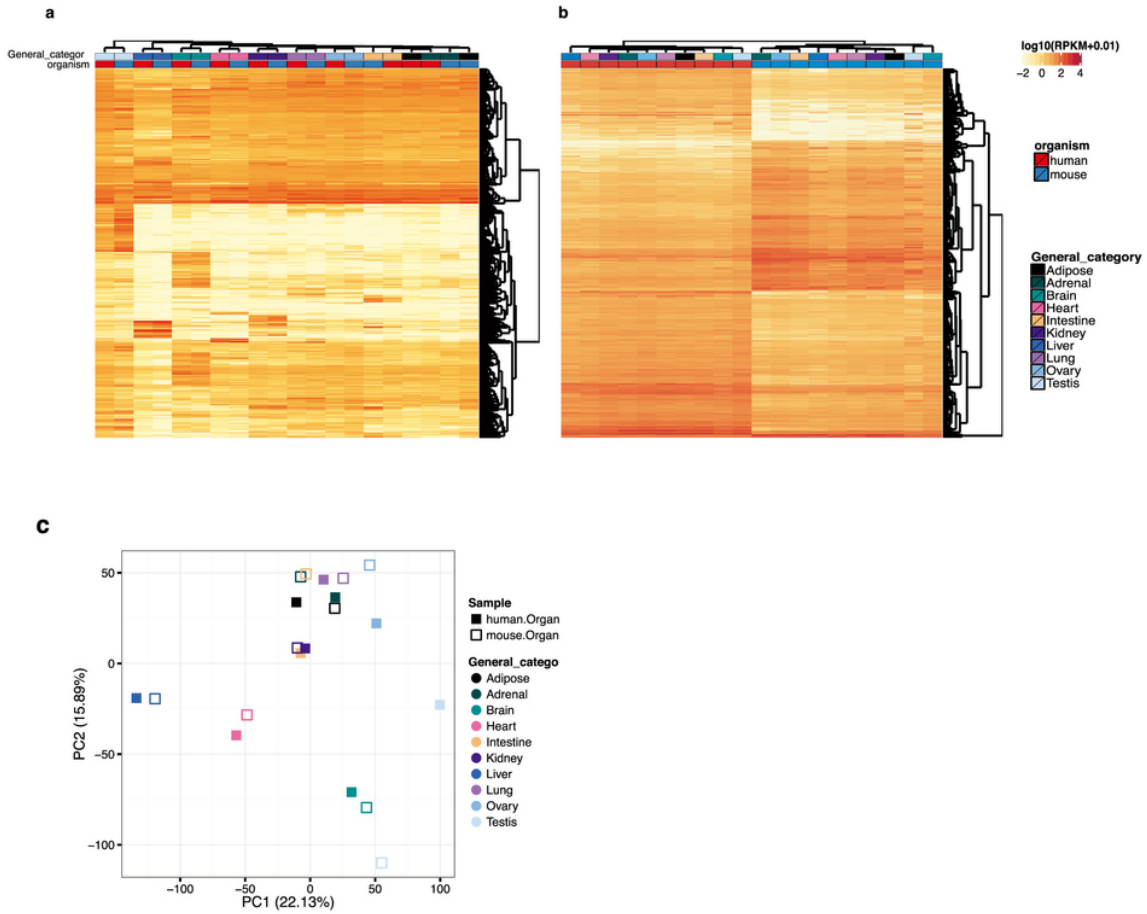
Auer and Doerge 2010

What standard technical issues should you consider for blocking:

- Flow Cell
- Lane
- Adaptors
- Library prep
- Same instrument
- People!
- RNA extraction/purification

What happens when you fail to block (or replicate)?

In a recent analysis of the mod-encode data, RNAseq data suggested that clustering (for gene expression) more by species than by tissue. This was an unusual finding.



Yue F, Cheng Y, Breschi A, et al.: A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014; 515(7527): 355–364

Lin S, Lin Y, Nery JR, et al.: Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A. 2014; 111(48): 17224–17229

A new re-analysis demonstrated some potentially serious issues with the experimental design

Gilad Y and Mizrahi-Man O. A reanalysis of mouse ENCODE comparative gene expression data [v1; ref status: indexed, http://f1000r.es/5ez] F1000Research 2015, 4:121 (doi: 10.12688/f1000research.6536.1)

Figure 1. Study design for :
Yue F, Cheng Y, Breschi A, et al.: A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014; 515(7527): 355–364
Lin S, Lin Y, Nery JR, et al.: Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A. 2014; 111(48): 17224–17229

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	● Human
testis		pancreas		● Mouse

Gilad Y and Mizrahi-Man O 2015 [v1; ref status: awaiting peer review, http://f1000r.es/5ez]
F1000Research 2015, 4:121 (doi: 10.12688/f1000research.6536.1)

F1000Research

Differential expression

- Probably the single most common use of RNA-Seq data is examine differential expression of transcripts (transcriptional profiles).

Differential expression

- But differential expression of what?

Differential expression

- But differential expression of what?
  - Genes
  - Transcripts (alternative transcripts)
  - Allele specific expression
  - Exon level expression

Your primary goals of your experiment should guide your design.

- The exact details (*# biological samples*, sample depth, read_length, strand specificity) of how you perform your experiment needs to be guided by your primary goal.

- Unless you have all the $$, no single design can capture all of the variability.

Your goals matter

- For instance: If your primary interest in discovery of new transcripts, sampling deeply within a sample is probably best.

- For differential expression analyses, you will almost never have the ability to perform Differential expression analysis on very rare transcripts, so it is rarely useful to generate more than 15-20 million read pairs per biological sample.

A simple truth:
There is no technology nor statistical wizardry that can save a poorly planned experiment. The only truly failed experiment is a poorly planned one.

To consult the statistician after an experiment is finished is often merely to ask him(her) to conduct a post mortem examination. He(she) can perhaps say what the experiment died of.

Ronald Fisher

Counting

- One of the most difficult issues has been how to count.
- We first need to ask what ***features*** we want to count.

What Features could we count?

What Features could we count?

- Counting at the level of genes (reads mapped to gene regardless of transcript).
- Counting at the level of transcript.
- Counting at the level of exons.
- Counting at the level of kmers within one of the above
- Counting at the level of nucleotides within exon/transcript/gene.

Counting

- We are interested in transcript abundance.
- But we need to take into account a number of things.

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.
- TPM, TMM… etc…

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

BUT: Not everyone agrees with this approach though. Nor with my arguments about counting.

Lior Patcher's blog is a good place to watch the debate.
Also check out some comments in the vignette and paper on limma/voom.

RPKM

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

R = total # mapped reads from that sample

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

fl_{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, etc..)

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.

The issue is that getting unambiguous counts is hard (Rob).

Differential Expression analysis. A Primer.

- I am assuming that we have already decided on an appropriate method to count and convert mapped reads to discrete values…
- There is a bit we need to know to help us understand what to do next.

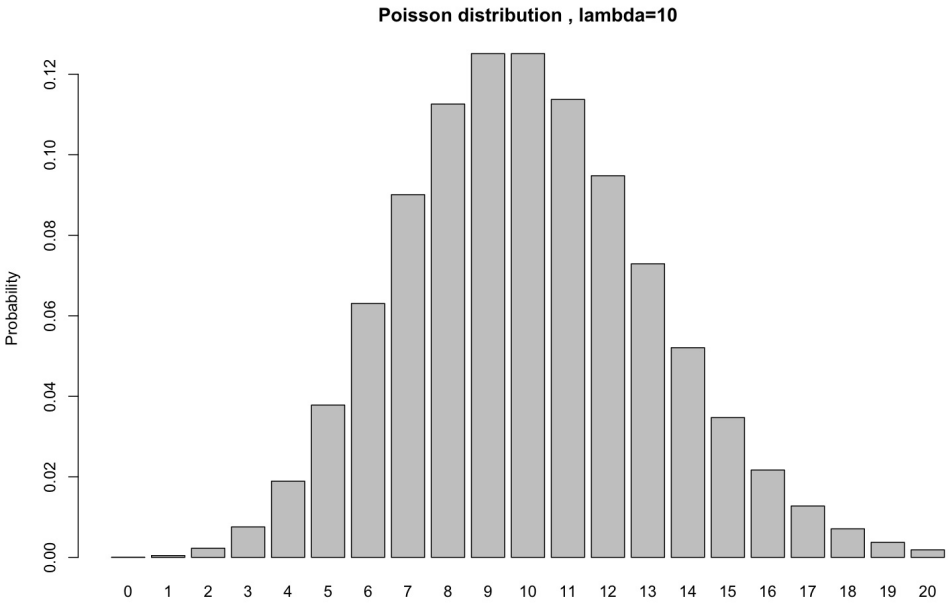
A bit of background on probability.

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

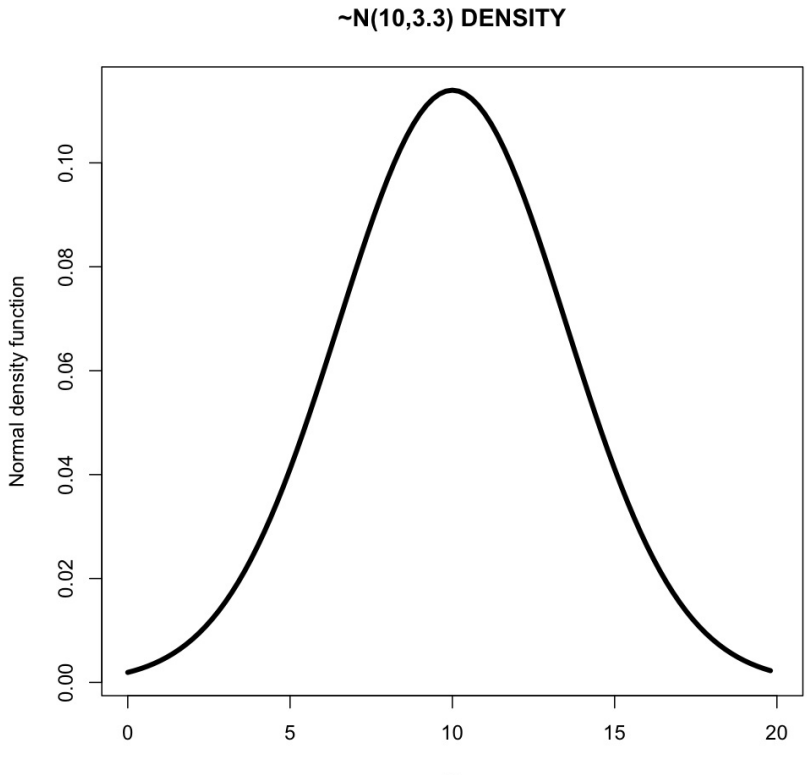
- Depending upon the data modeling framework we wish to use, we need to account for this, as these are not necessarily approximated well by normal (Gaussian) distributions that are used for “standard” linear models like t-tests, ANOVA, regression.

- 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

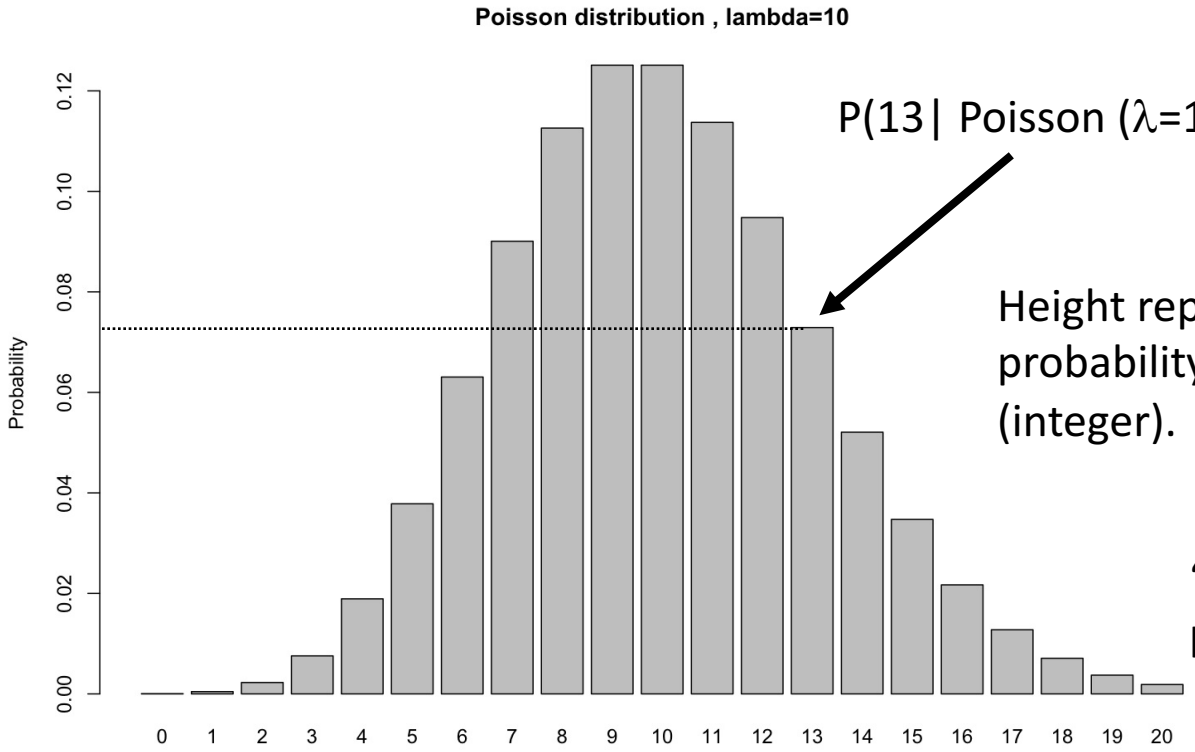


Probability Mass function for a discrete variable.



Probability Density function for a continuous variable.

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



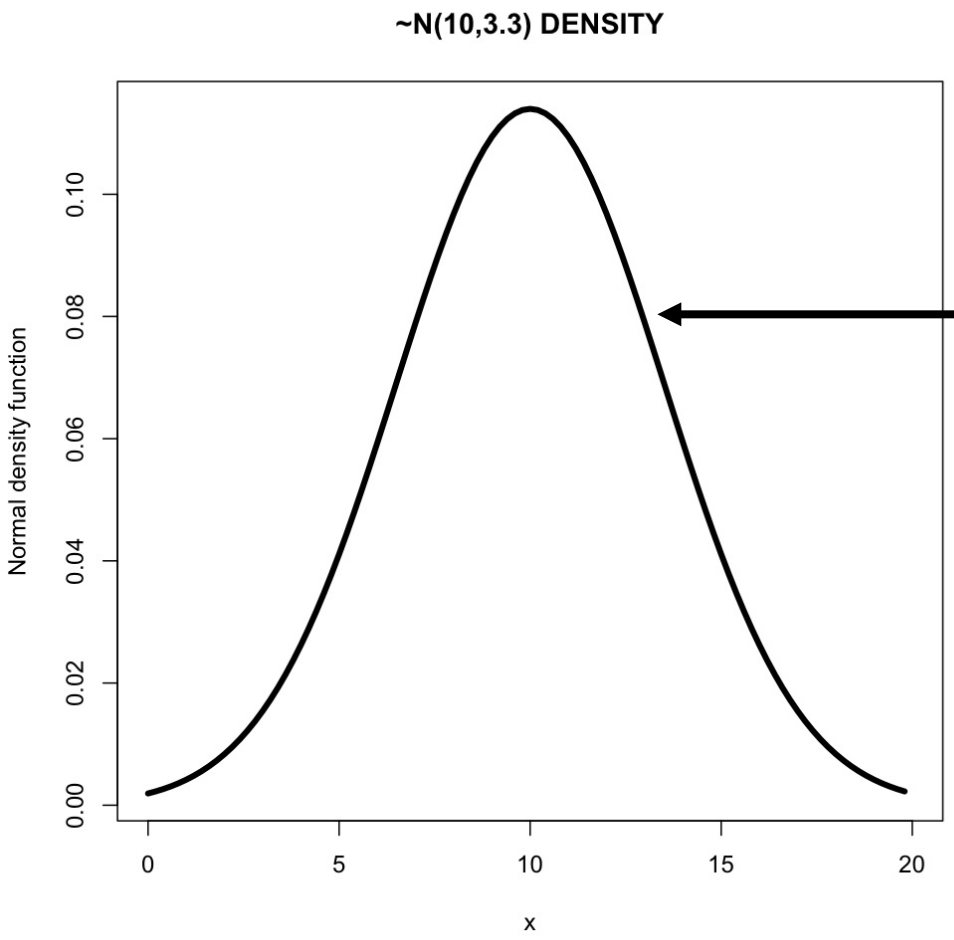
P(13| Poisson (λ=10)) =0.073

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

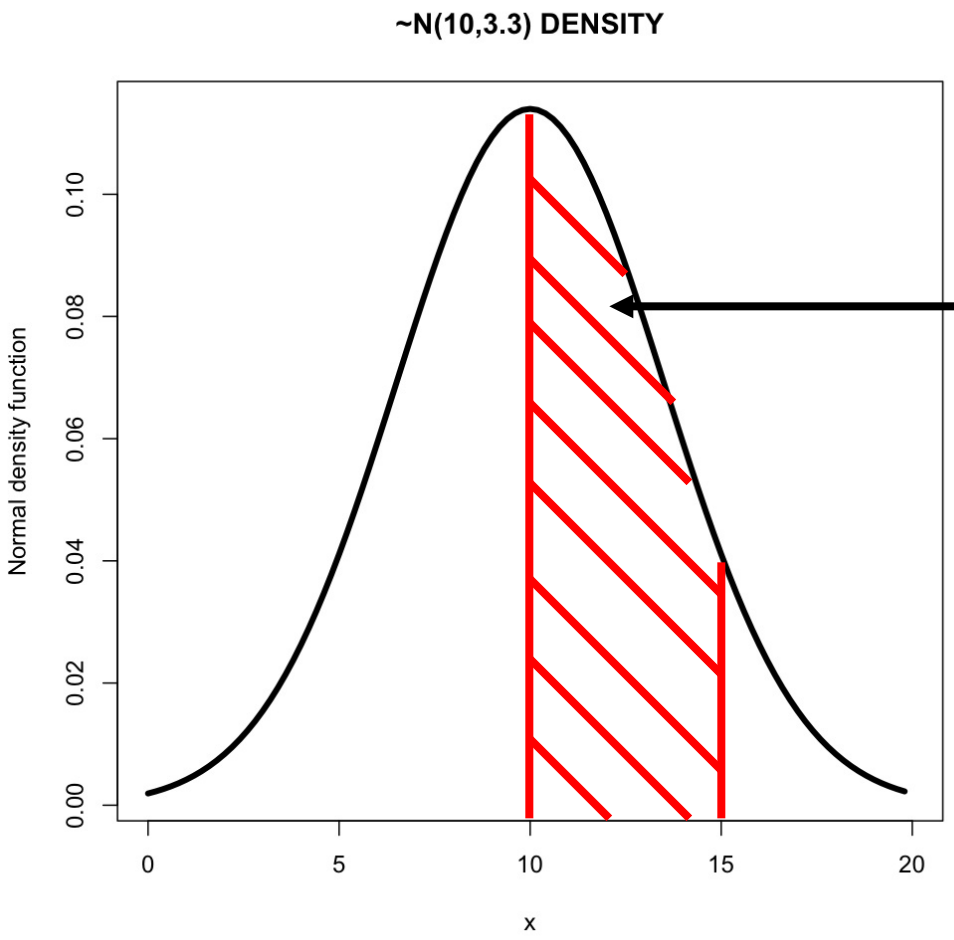


Height at x= 13 is 0.0799

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

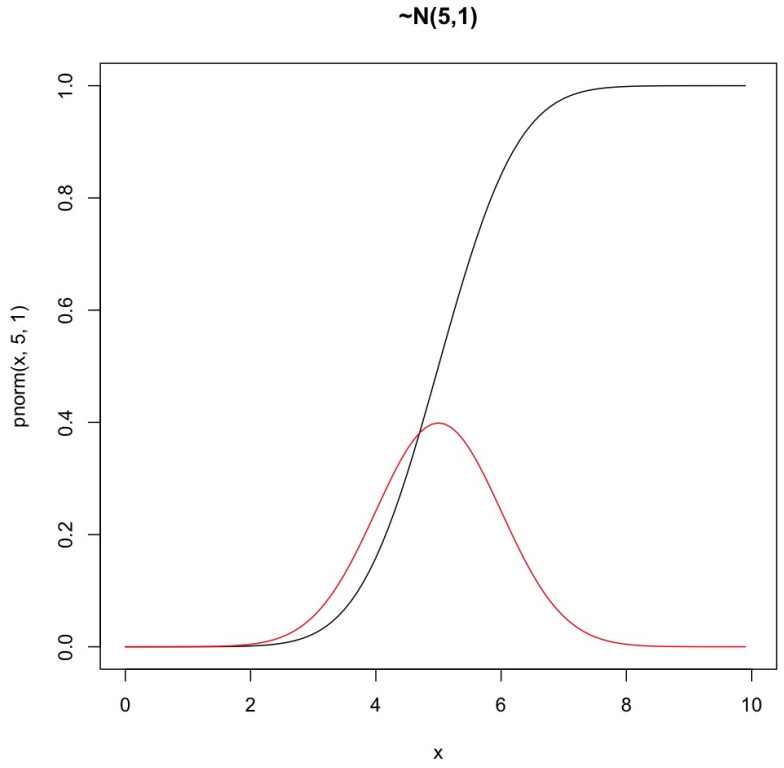


We can define the probability in the interval

$10 \leq x \leq 15$

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

Clarifications on continuous distributions.



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

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

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

AREA UNDER CURVE OF PDF =1

(The integral of the normal)

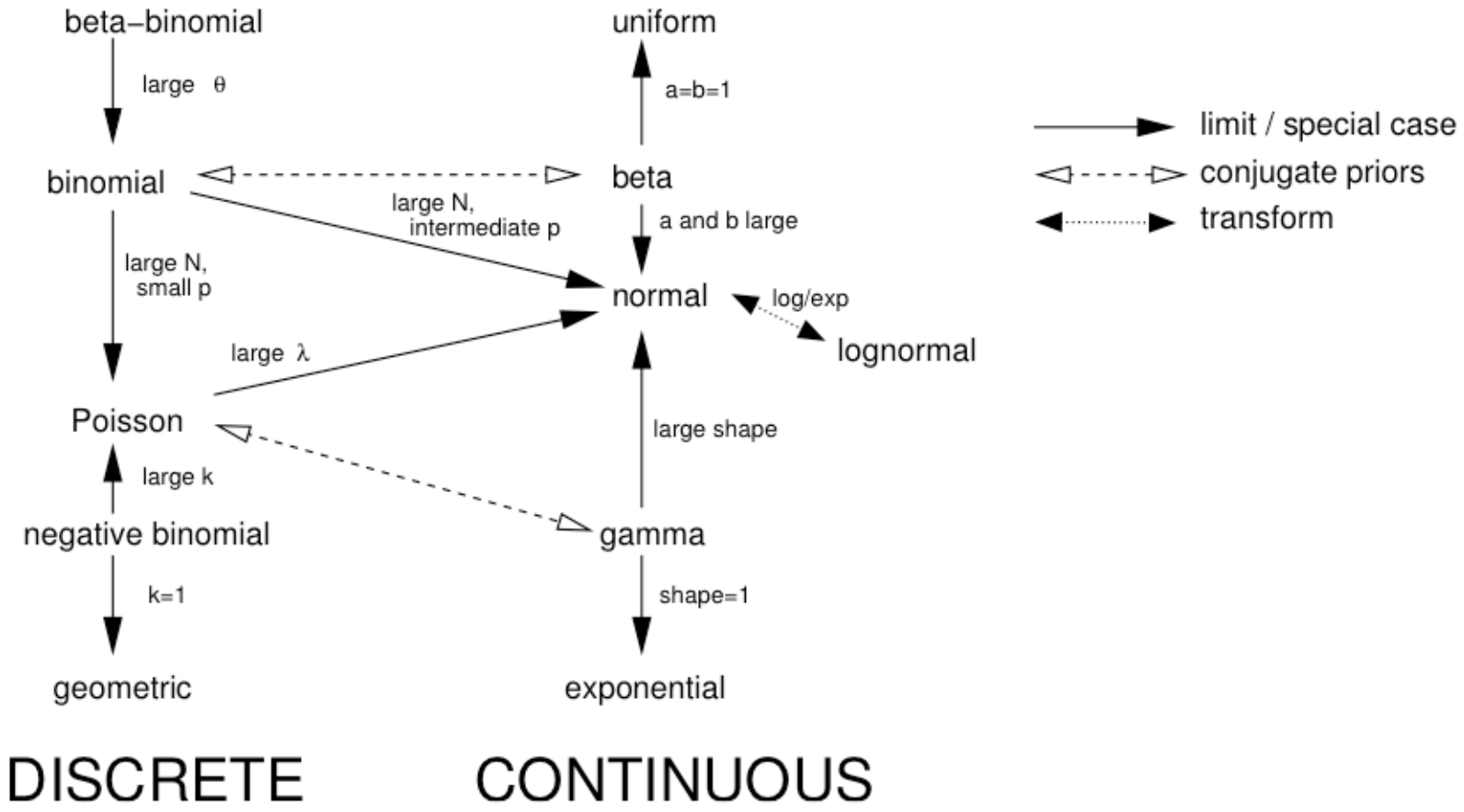


Figure 4.17 Relationships among probability distributions.

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 may want to use the 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.

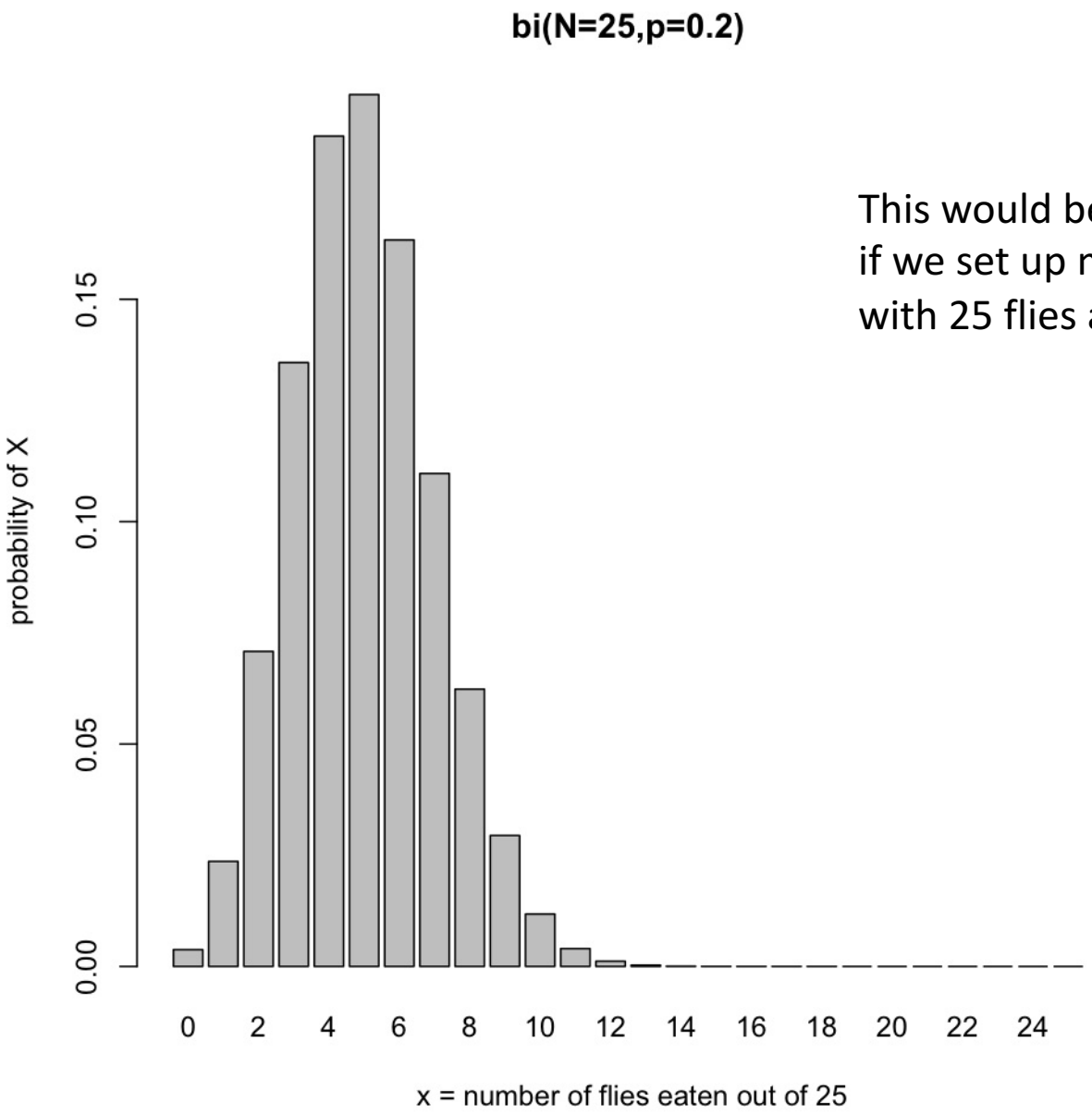
You will often see x=k and hear " N choose k"

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.

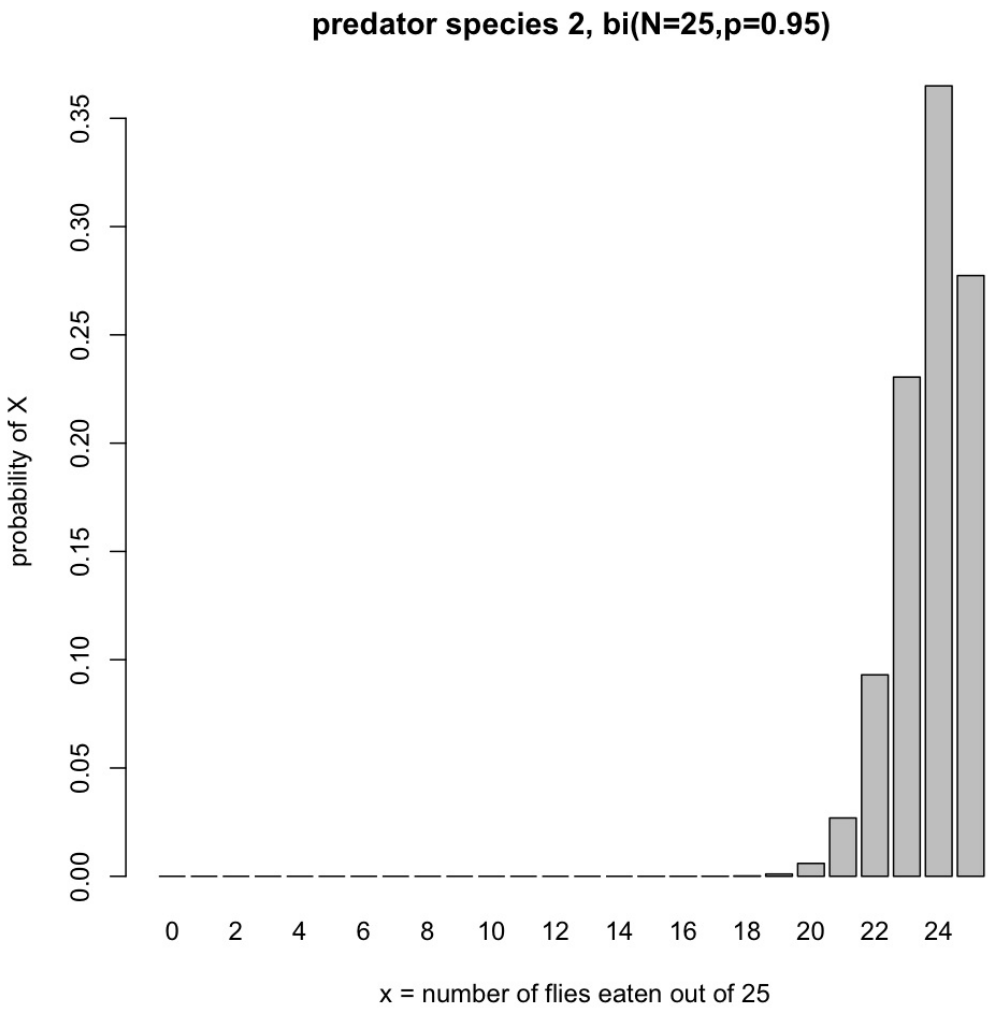
$P(x=10 \mid bi(N=25,p=0.2)) = 0.0118$

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

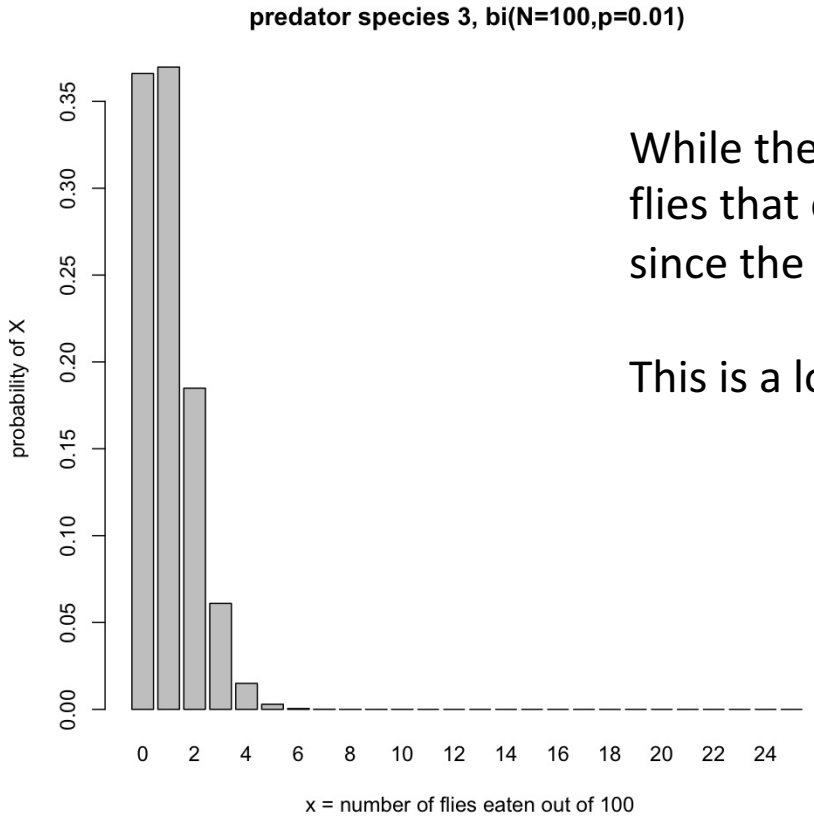


This would be the expected distribution if we set up many replicate enclosures with 25 flies and this predator.

Predator species 2 is much hungrier….



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 (well sort of)…

Poisson

- It is also seemingly useful for RNA-Seq data. (although we will see not very useful in practice).

Poisson

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

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

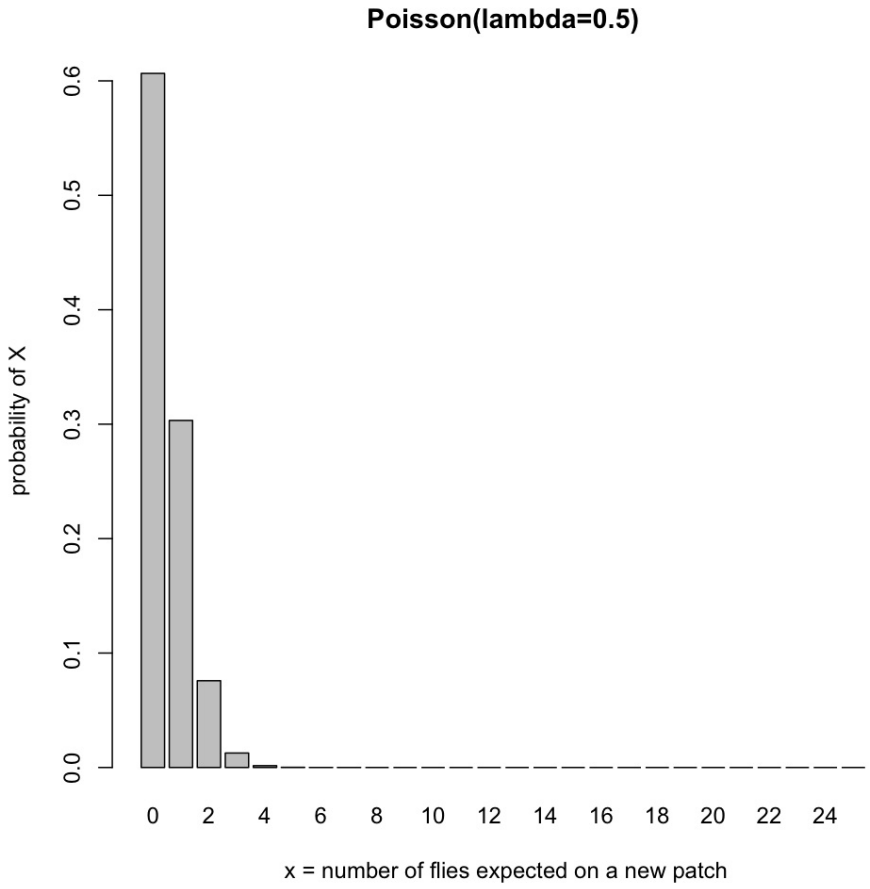
For its relation to a binomial when N is large and p is small
$\lambda = N*p$

Poisson

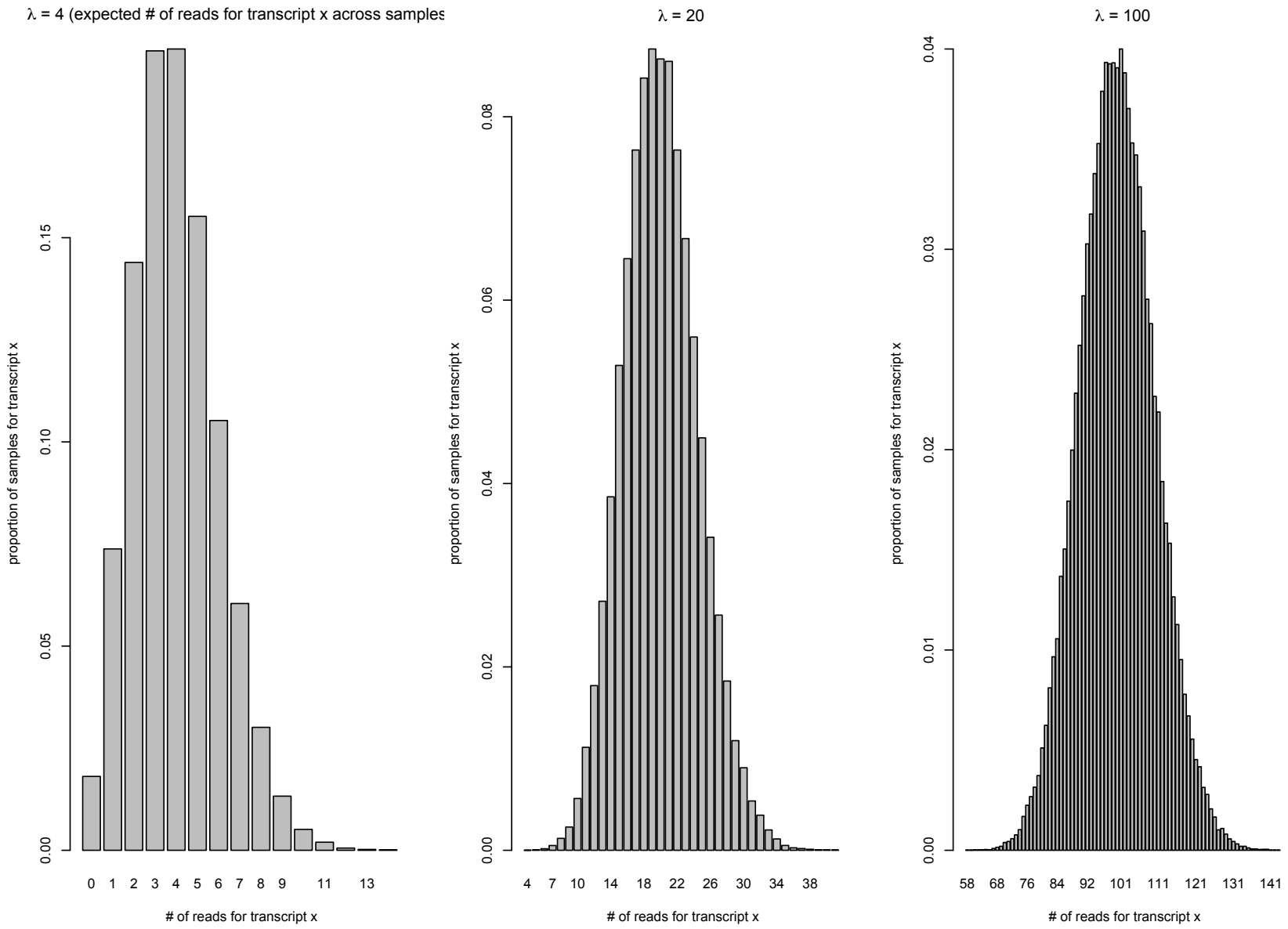
- Let's say flies disperse to colonize a new patch at a very low rate (previous estimates suggest we will observe one fly for every two new patches we examine, $\lambda=0.5$).
- What is the probability of observing 2 flies on a new patch of land?

P(x=2| poisson(λ=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).

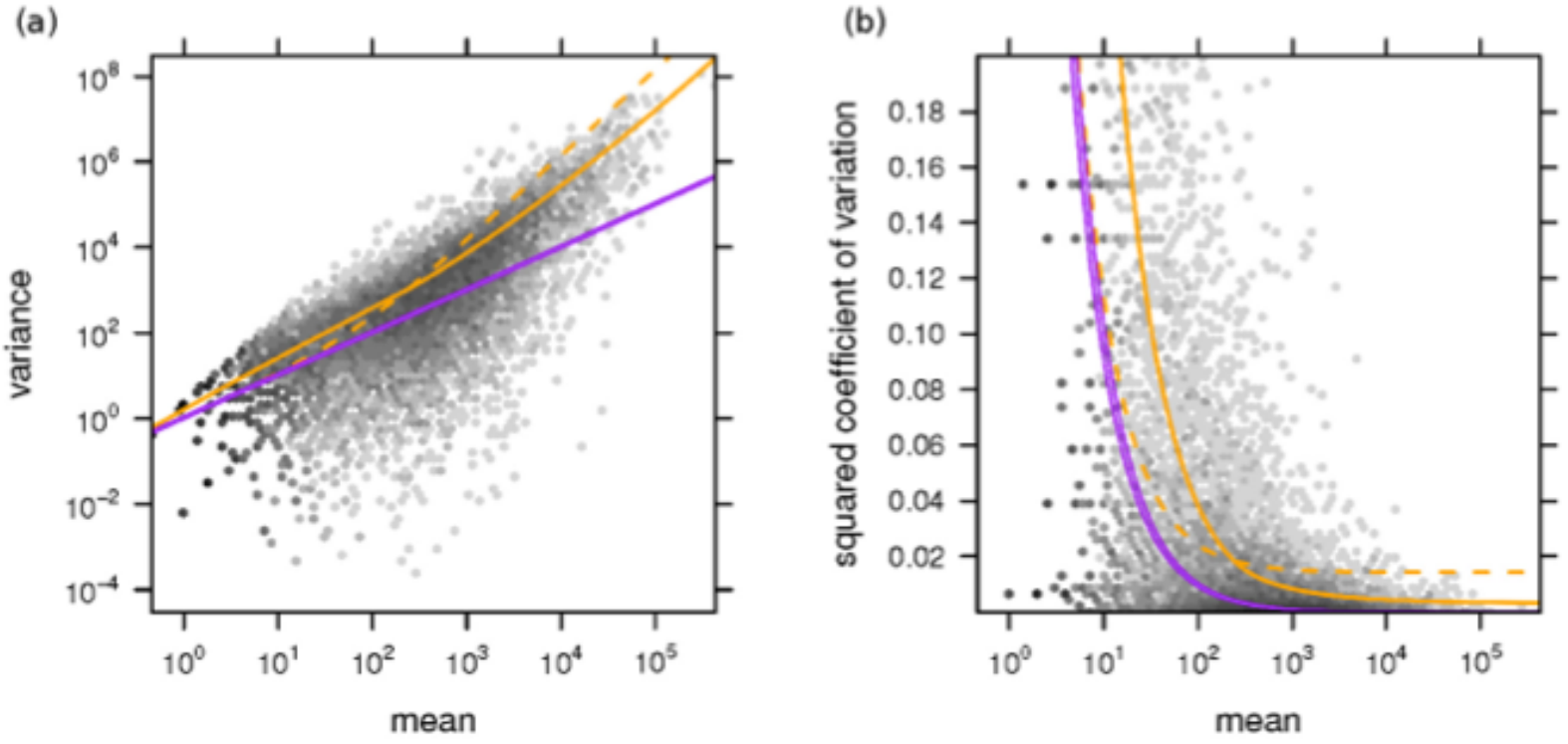


Figure 1 Dependence of the variance on the mean for condition *A* in the fly RNA-Seq data. (a) The scatter plot shows the common-scale sample variances (Equation (7)) plotted against the common-scale means (Equation (6)). The orange line is the fit $w(q)$. The purple lines show the variance implied by 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) = λ, it is

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

How about a normal distribution?

- Despite working with discrete count data, several authors use normal distributions. Several reasons.

How about a normal distribution?

- Despite working with discrete count data, several authors use normal distributions. Several reasons:

1. When the mean number of counts is far enough away from zero, often the normal distribution does a good job of fitting the data (and capturing mean & variance relationship). For low mean counts a variance stabilization can aid modeling (the approach used in limma/voom).

2. Our response variable (counts of features) are not measured without error, and therefore are not true measures. When estimating effects in our model we account for this uncertainty and assuming a normal distribution enables additional flexibility.

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.

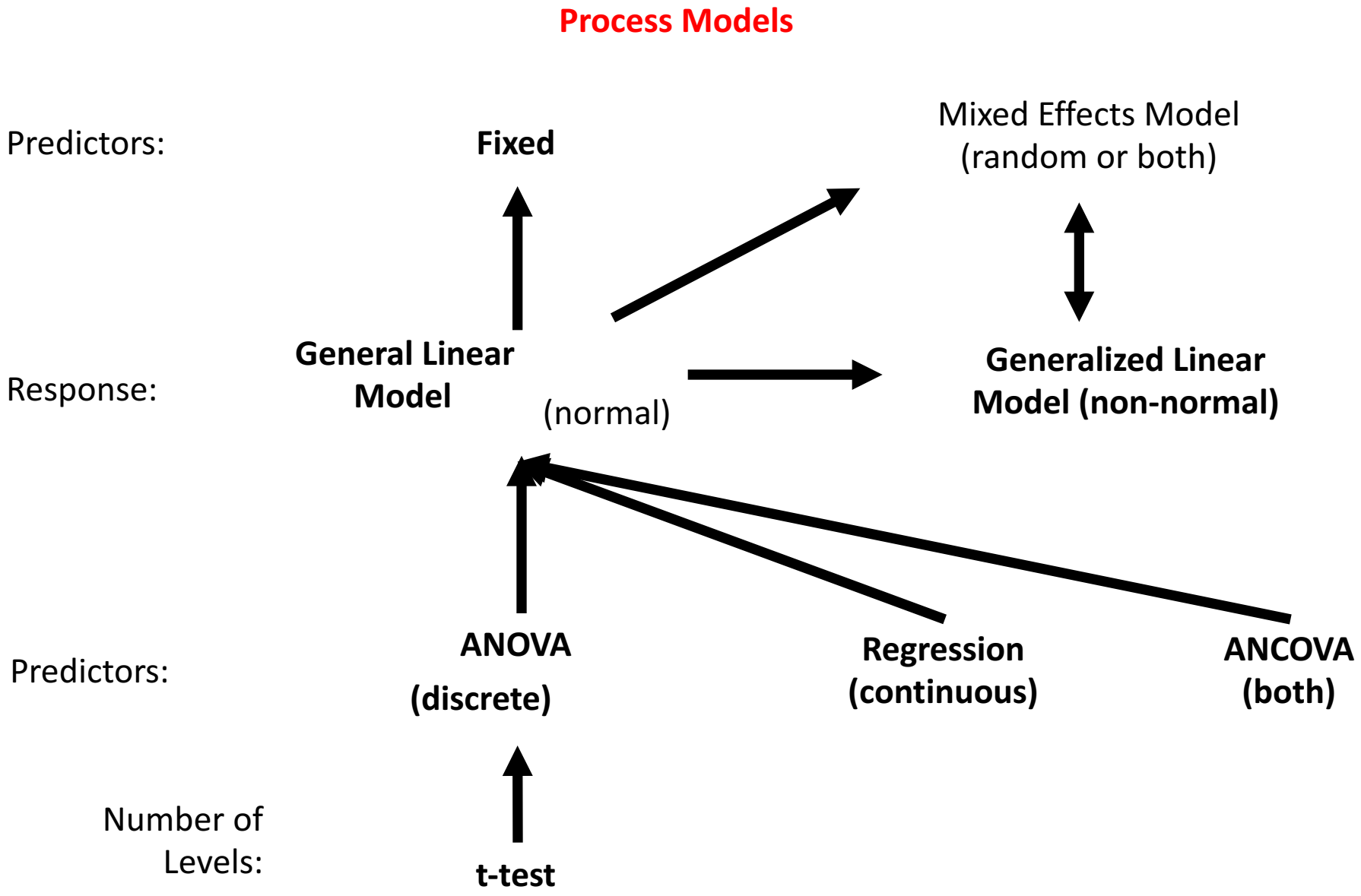
$$y \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right)$$

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

Generalized linear model

- MANY of the differential expression tools utilize a linear model framework.
- Thus it is important to get familiar with the framework.
- The class by Jonathan and Ben (B) is probably a great place to start.

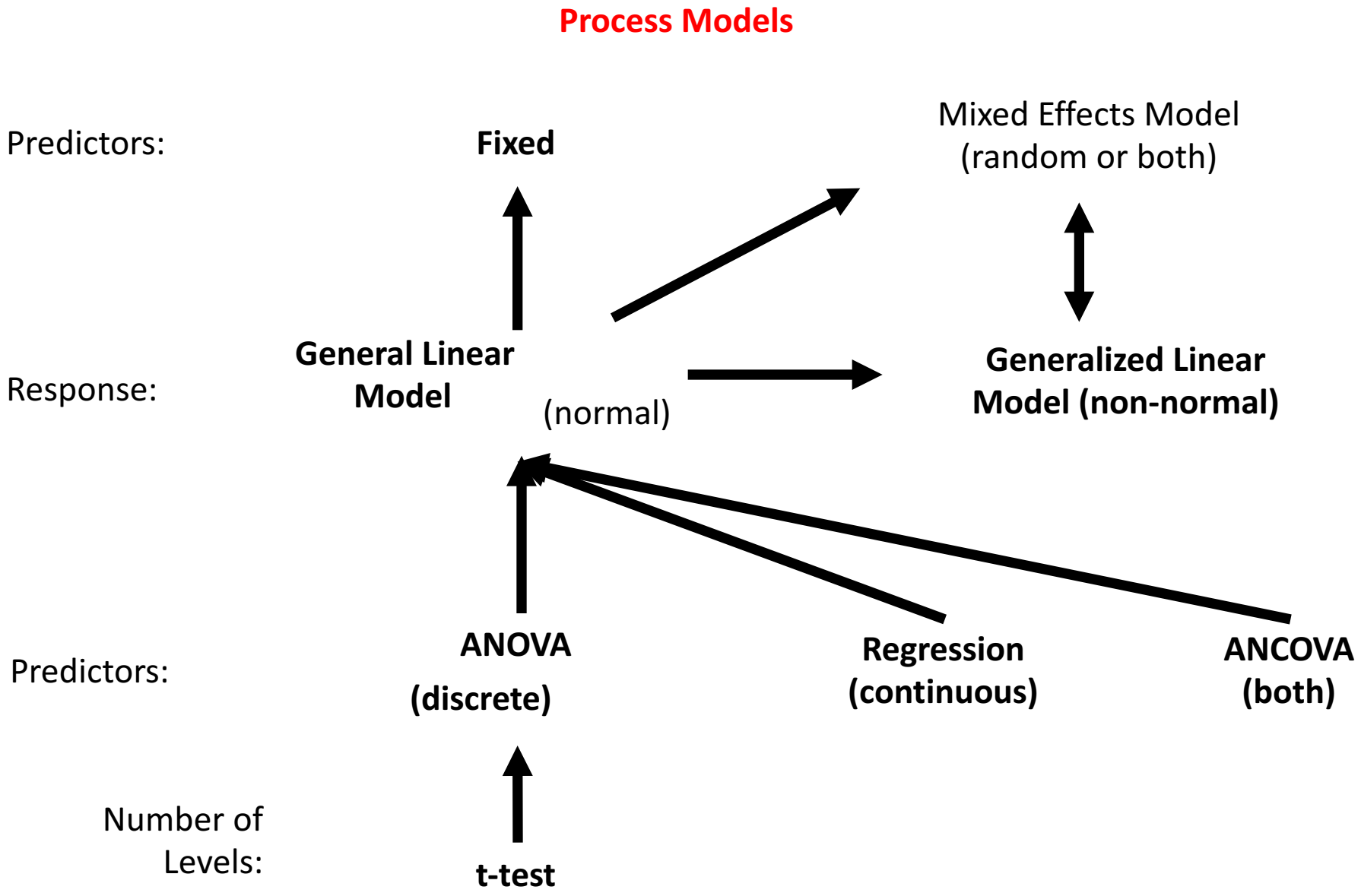
Continuity of Statistical Approaches



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

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

Poisson and negative binomial 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/DESeq2 (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 useful

- Since we can fit these as a generalized linear model, we can fit arbitrarily complex designs (if we have sufficient sample sizes to estimate all 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 uncertainty in counts, mean-variance relationships, and how best to model such effects.

- 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

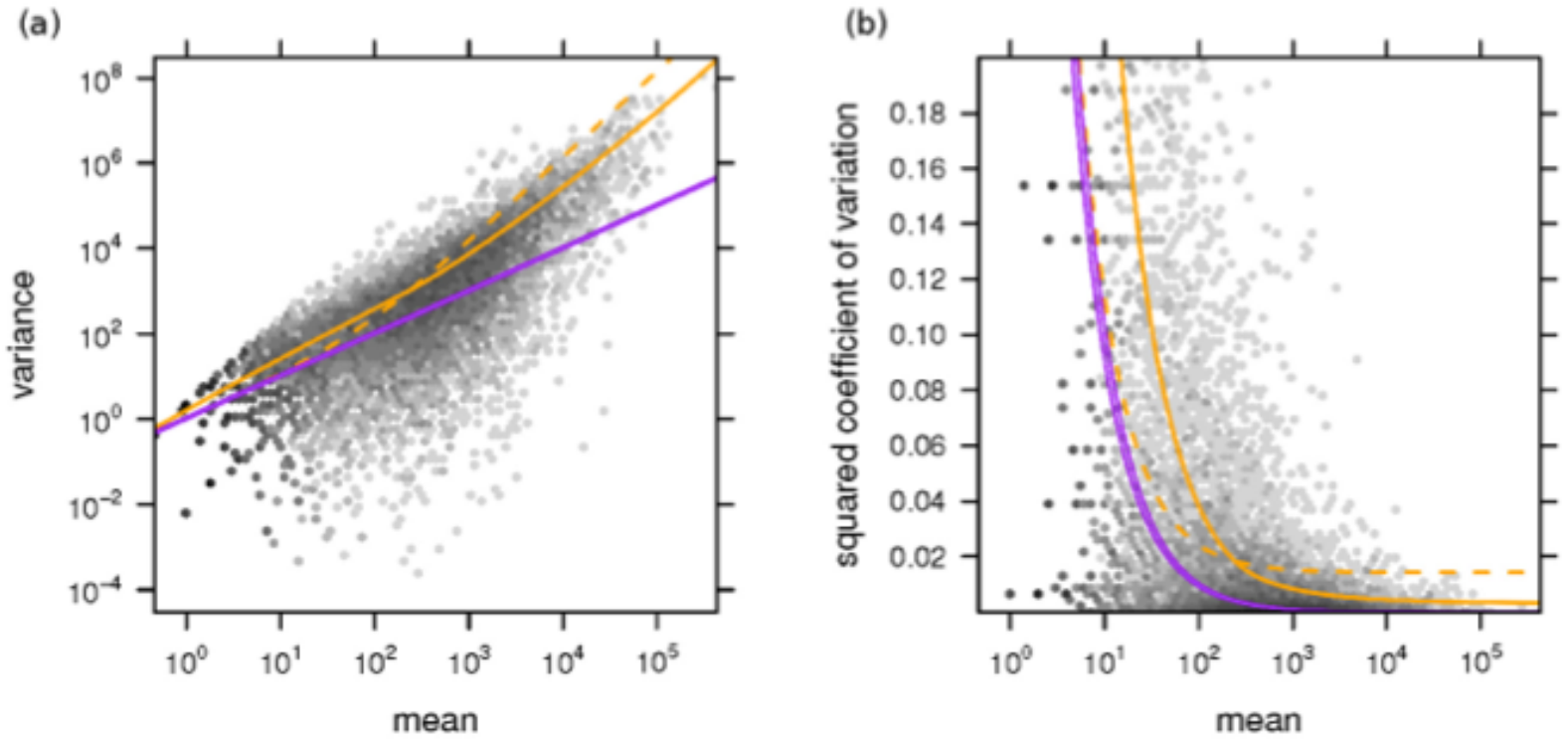


Figure 1 Dependence of the variance on the mean for condition *A* in the fly RNA-Seq data. (a) The scatter plot shows the common-scale sample variances (Equation (7)) plotted against the common-scale means (Equation (6)). The orange line is the fit $w(q)$. The purple lines show the variance implied by 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

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 ξ is a common dispersion and δ 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)	$X_{gij} \sim Binom(p_{gi}, N_{ij})$ $p_{gi} \sim Beta,\ logitE\{p_{gi}\} = Z\beta,$ $\mathrm{V}(p_{gi}) = \mathrm{E}(p_{gi})\,(1-\mathrm{E}(p_{gi}))\phi_g$	$\hat{\phi}_g^{BBSeq}$ 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). μ_{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) ϕ_{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). β 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.*

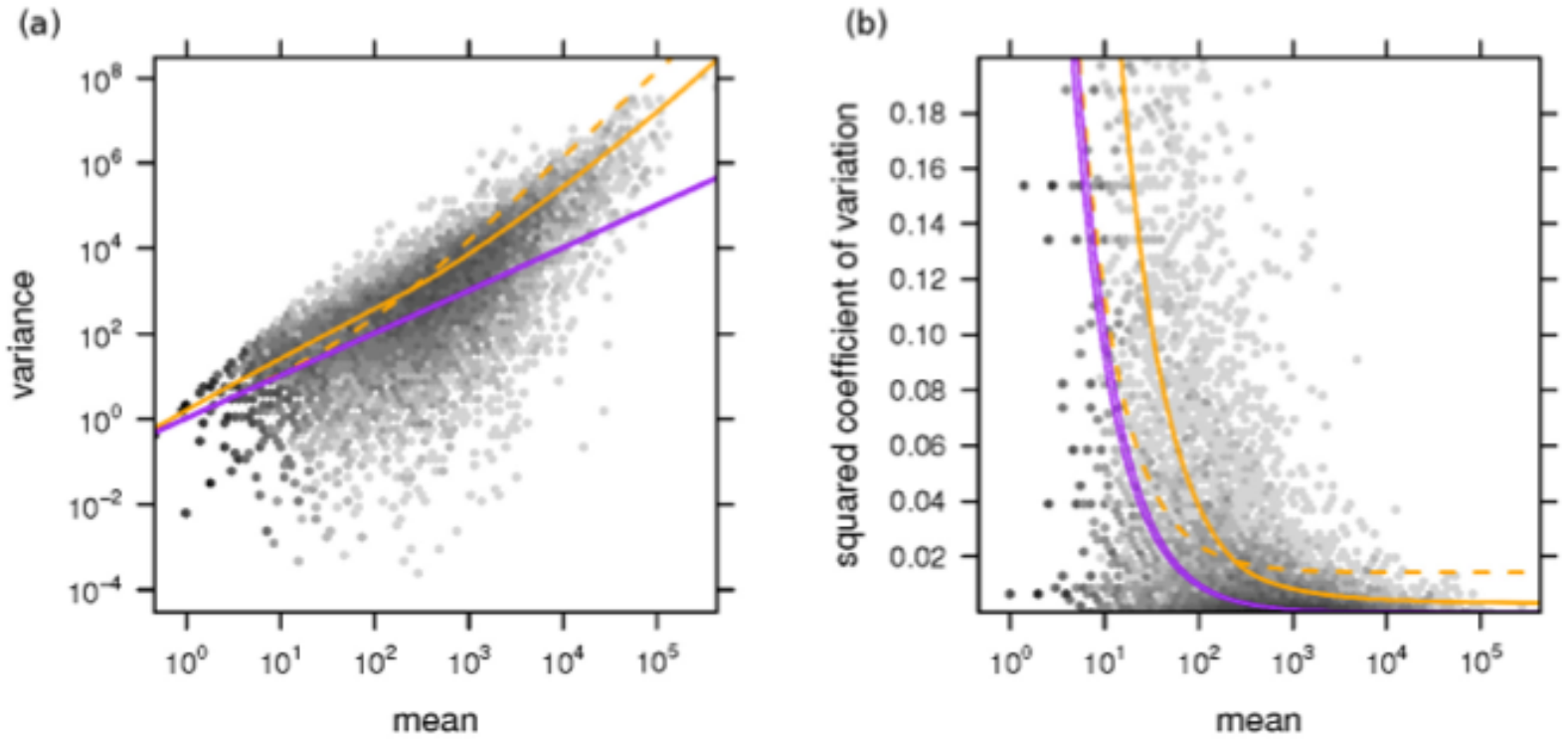
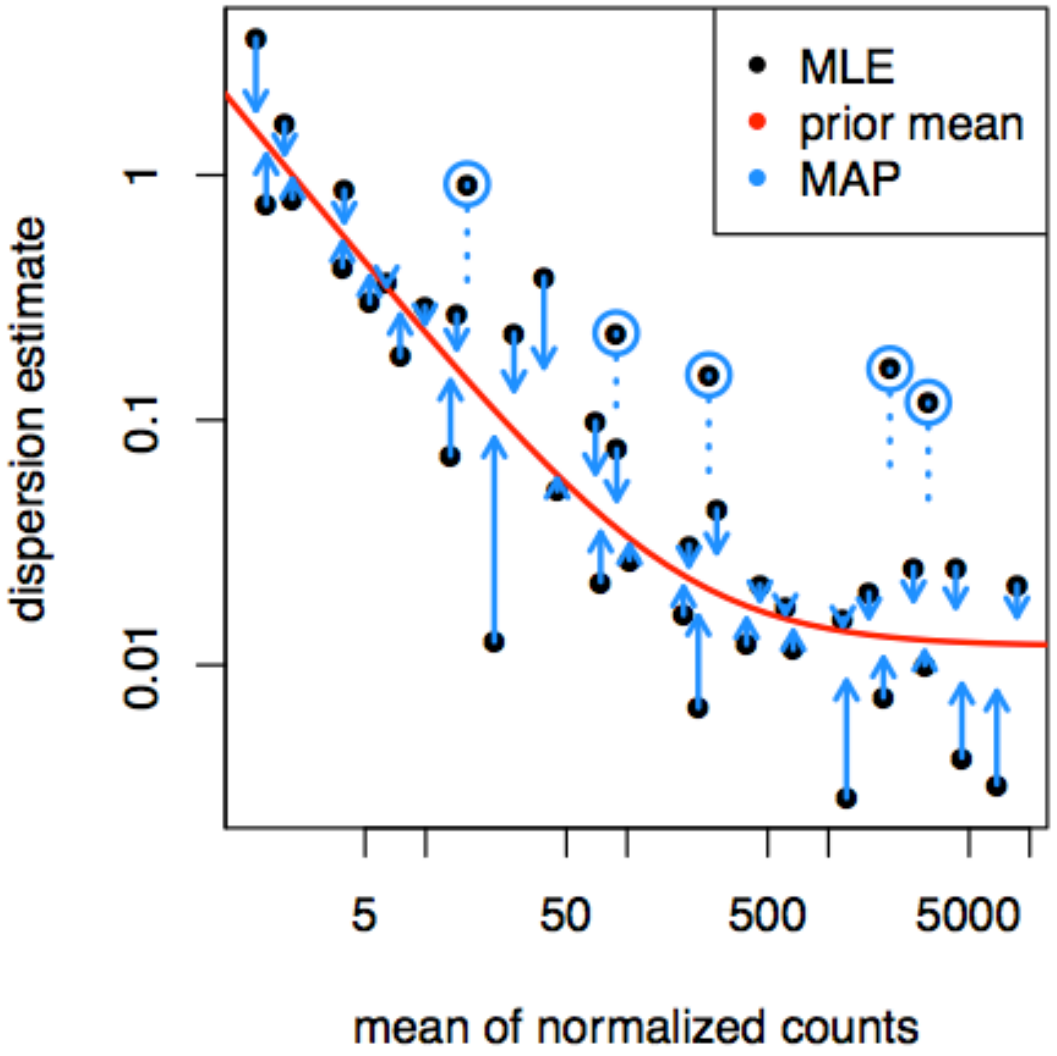


Figure 1 Dependence of the variance on the mean for condition *A* in the fly RNA-Seq data. (a) The scatter plot shows the common-scale sample variances (Equation (7)) plotted against the common-scale means (Equation (6)). The orange line is the fit $w(q)$. The purple lines show the variance implied by 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.)

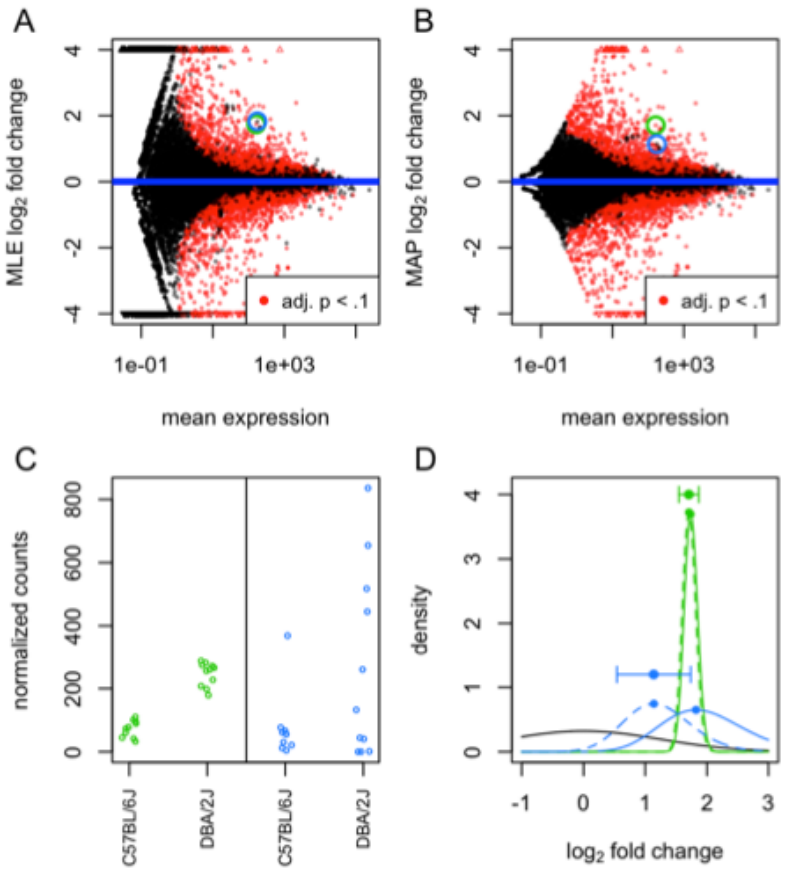
Anders and Huber 2010

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

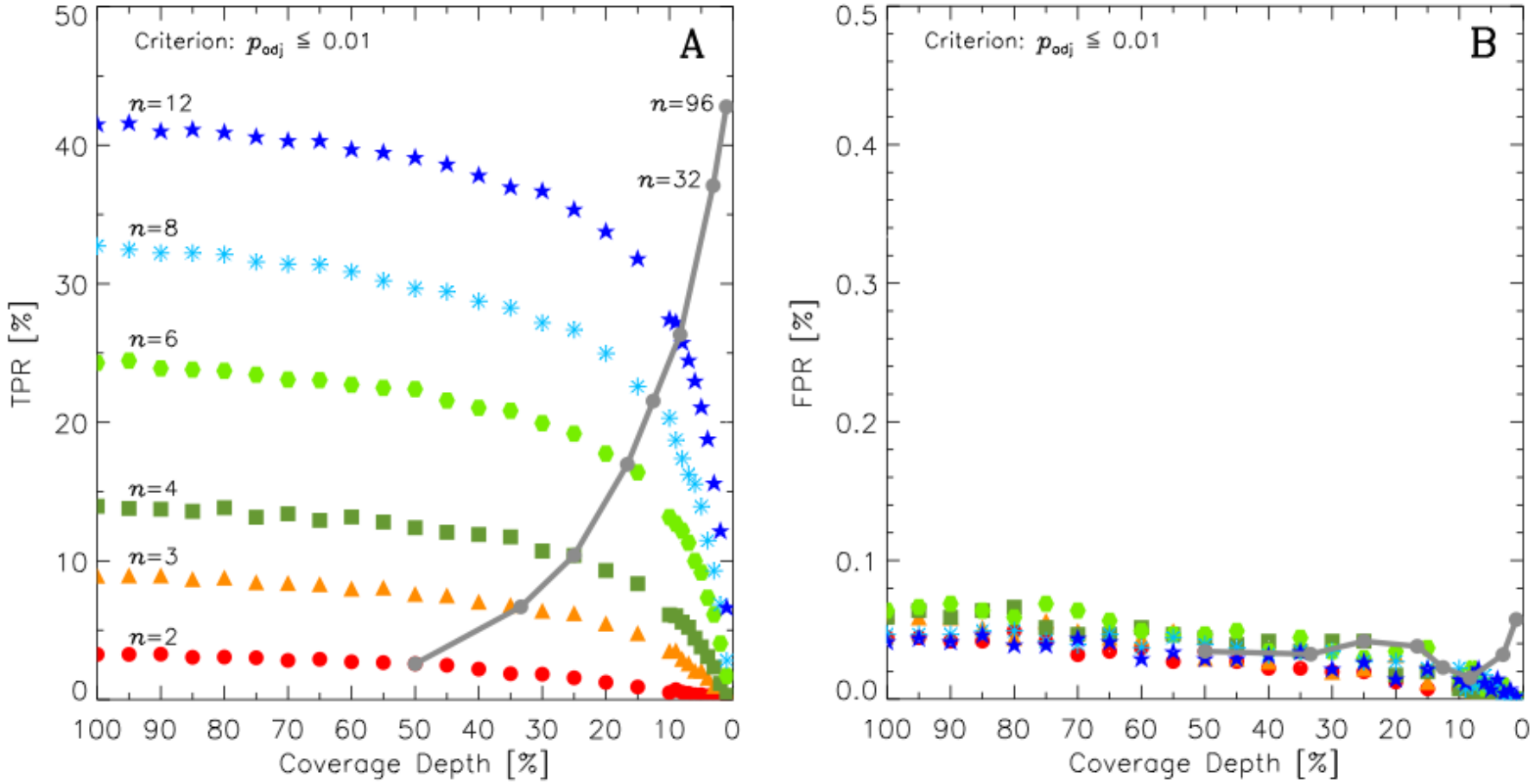
- With small sample sizes, the methods use different approaches to get gene-wise over-dispersion (based on all data).

- EdgeR is more powerful (more significant hits) than DESeq generally. But much more susceptible to false positives due to outliers.
- DESeq2 “should” be somewhere in the middle.

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

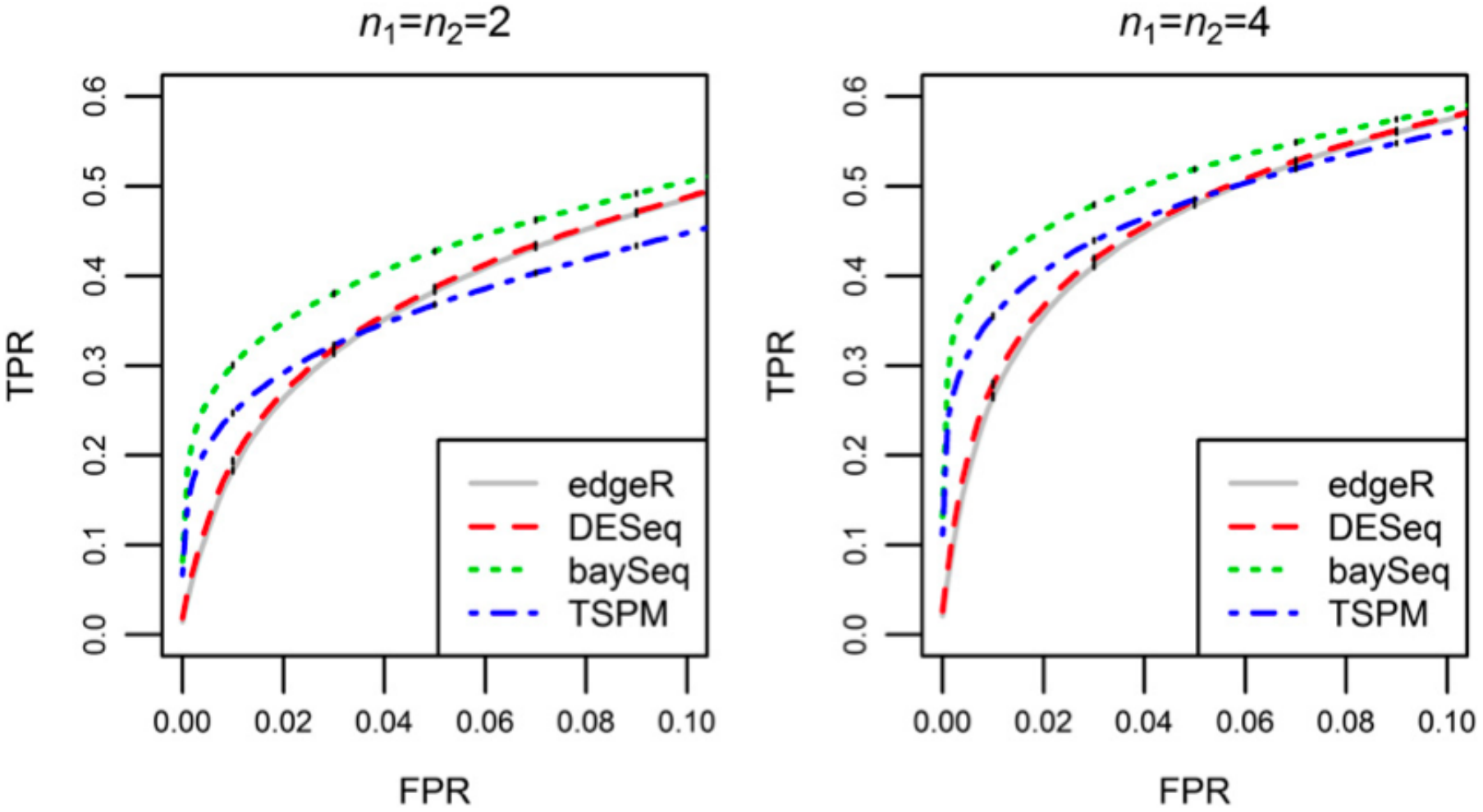
- Sequencing (and library prep) costs are still sufficiently expensive that most experiments use small numbers of biological replicates.
- Given the additional costs of library costs (~225$/sample at our facility), many folks go for increased depth instead of more samples.
- For a given level of sequencing 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!!!!



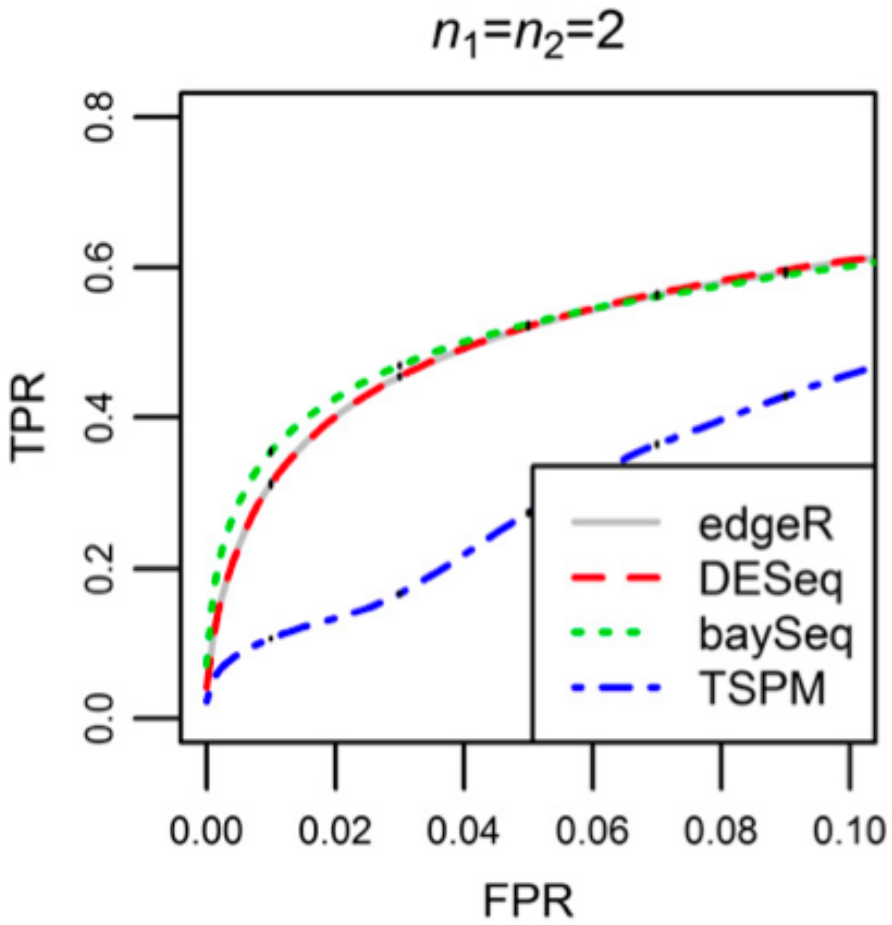
Robles et al. 2012

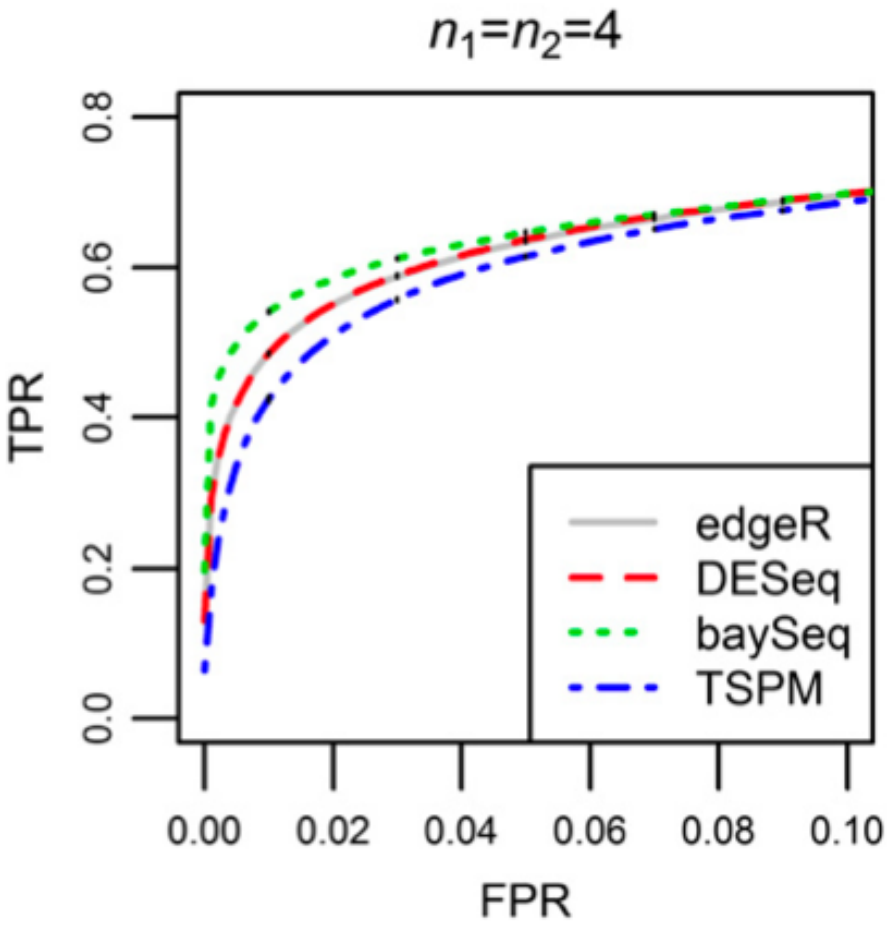
How do the methods compare in simulation?



Kvam et al. 2012

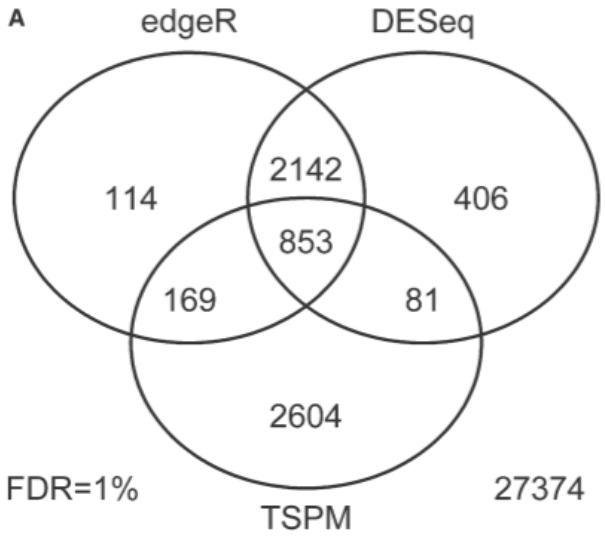
How do the methods compare in

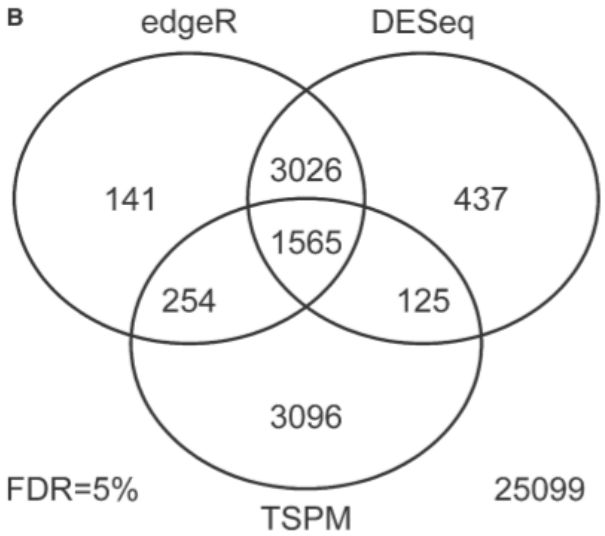




Kvam et al. 2012

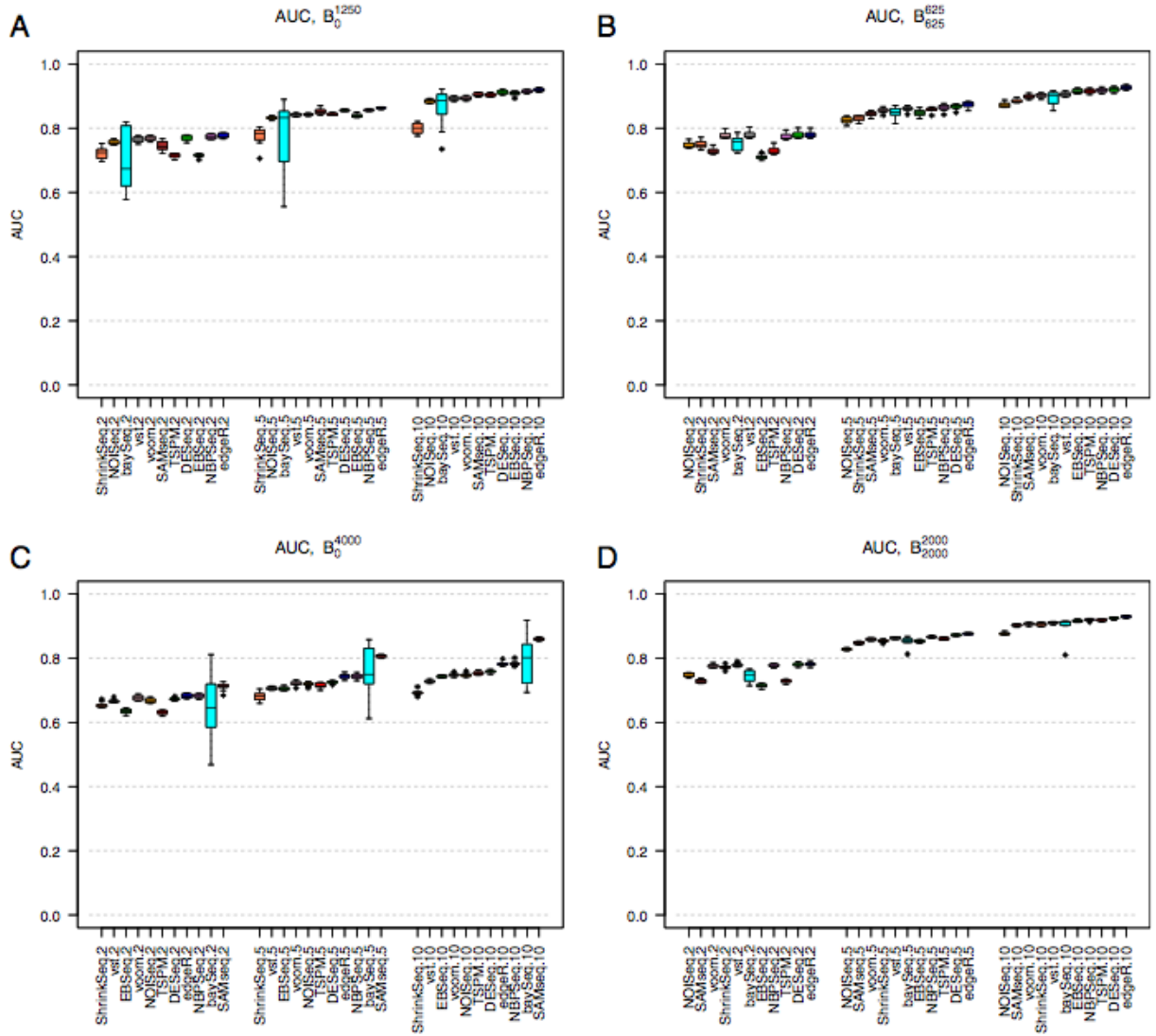
How do the methods compare for real data?





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

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

Why do we care about multiple comparisons?

How can we deal with multiple comparisons