Bioinformatics

How to Exploit Differential Expression?

Uwe Menzel, 2015

uwe.menzel@matstat.de

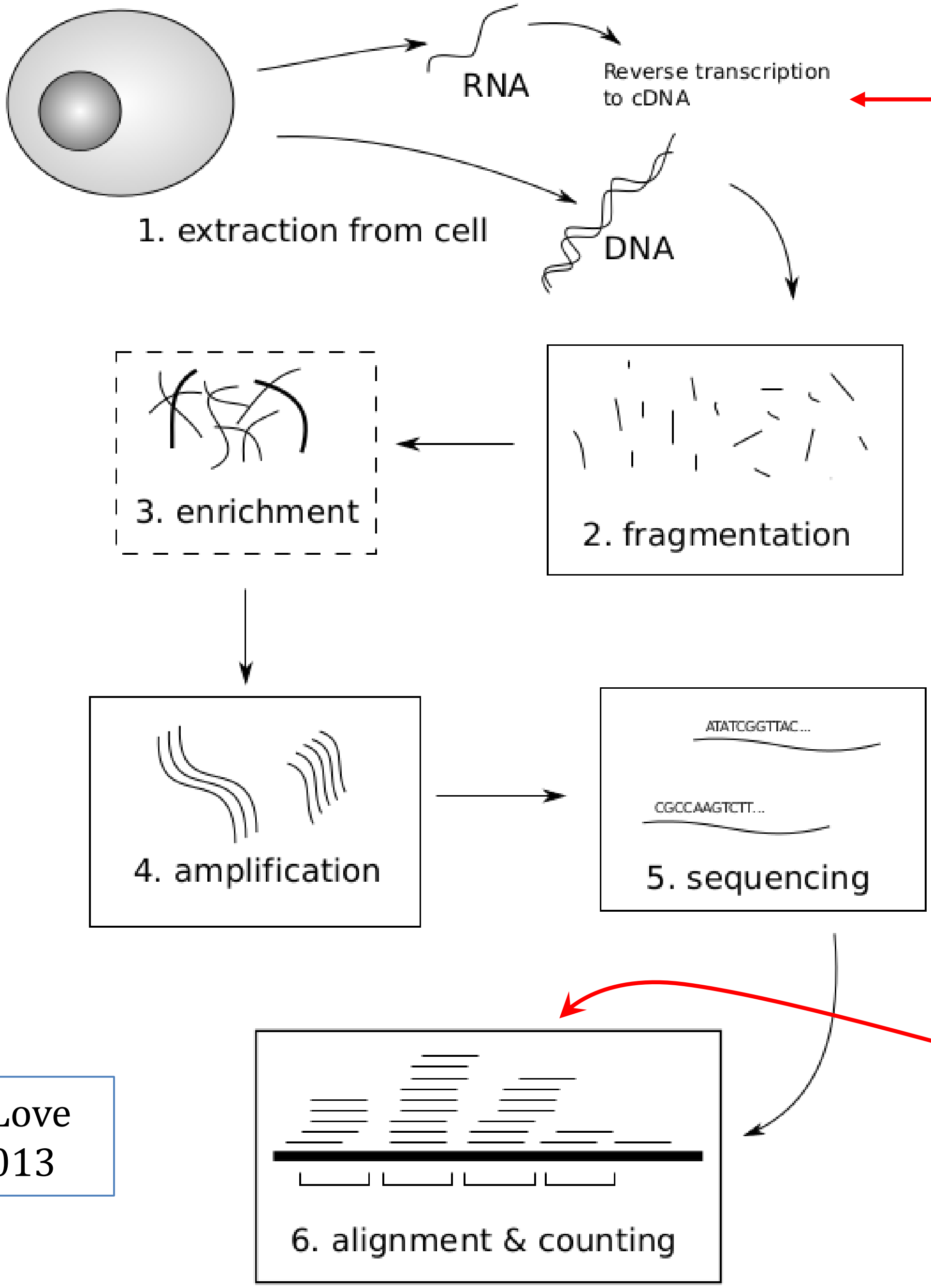
www.matstat.org

How to Exploit Differential Expression?

I've got a number of transcriptomes
under varied experimental conditions.

What comes next?

NGS

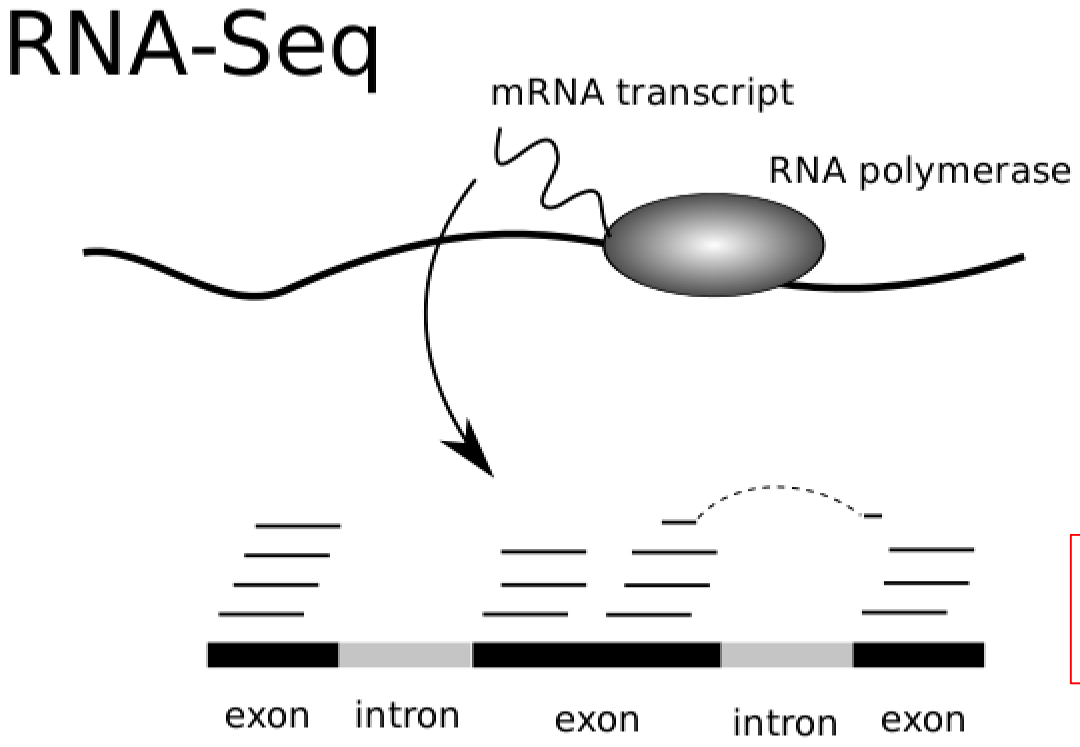


RNA-Seq

→ “counts”,
→ RPKM

Picture: Michael I. Love
Thesis FU Berlin 2013

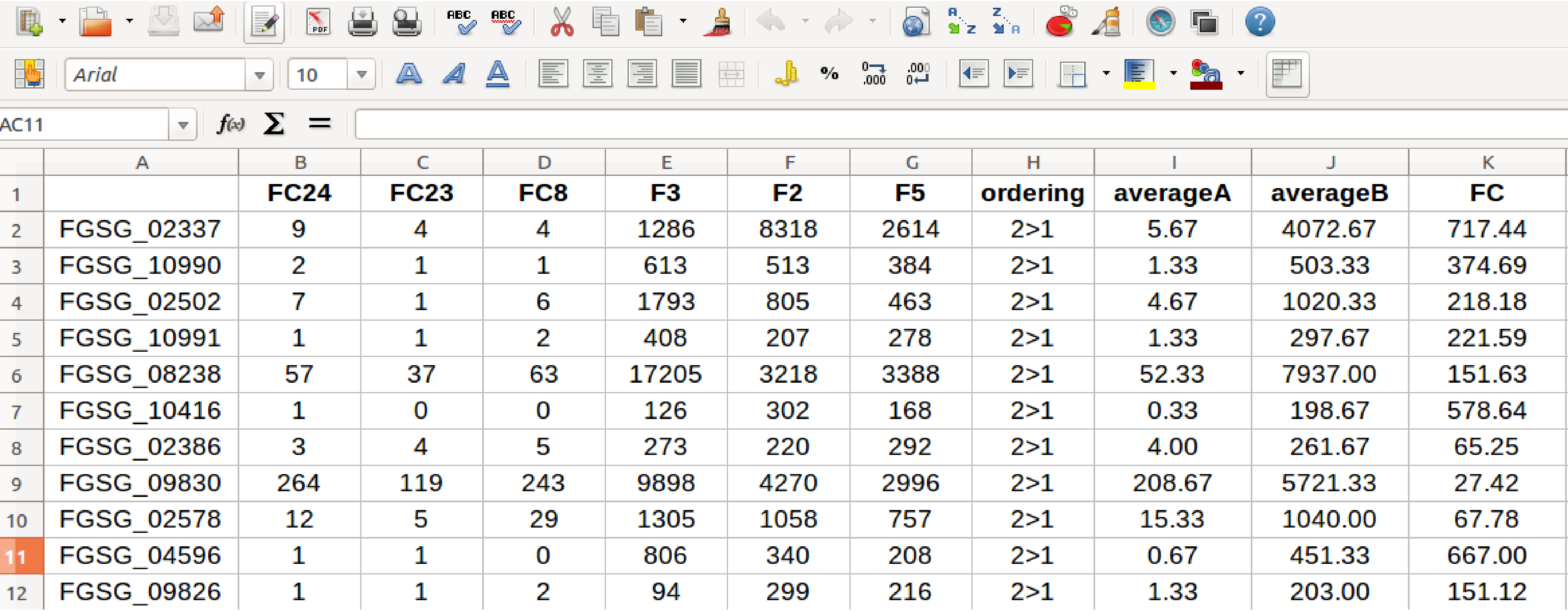
RNA-Seq



→ “counts”,
→ RPKM

Picture: Michael I. Love
Thesis FU Berlin 2013

Differences in Expression under varied Experimental Conditions



	A	B	C	D	E	F	G	H	I	J	K
1		FC24	FC23	FC8	F3	F2	F5	ordering	averageA	averageB	FC
2	FGSG_02337	9	4	4	1286	8318	2614	2>1	5.67	4072.67	717.44
3	FGSG_10990	2	1	1	613	513	384	2>1	1.33	503.33	374.69
4	FGSG_02502	7	1	6	1793	805	463	2>1	4.67	1020.33	218.18
5	FGSG_10991	1	1	2	408	207	278	2>1	1.33	297.67	221.59
6	FGSG_08238	57	37	63	17205	3218	3388	2>1	52.33	7937.00	151.63
7	FGSG_10416	1	0	0	126	302	168	2>1	0.33	198.67	578.64
8	FGSG_02386	3	4	5	273	220	292	2>1	4.00	261.67	65.25
9	FGSG_09830	264	119	243	9898	4270	2996	2>1	208.67	5721.33	27.42
10	FGSG_02578	12	5	29	1305	1058	757	2>1	15.33	1040.00	67.78
11	FGSG_04596	1	1	0	806	340	208	2>1	0.67	451.33	667.00
12	FGSG_09826	1	1	2	94	299	216	2>1	1.33	203.00	151.12

- Treatment: Expression profile of *F. avenaceum* when provoking infection
- Control: Expression profile of *F. avenaceum* when on culture medium

For the majority of the genes, the differences in counts are not as obvious as in this figure → statistical tools needed!

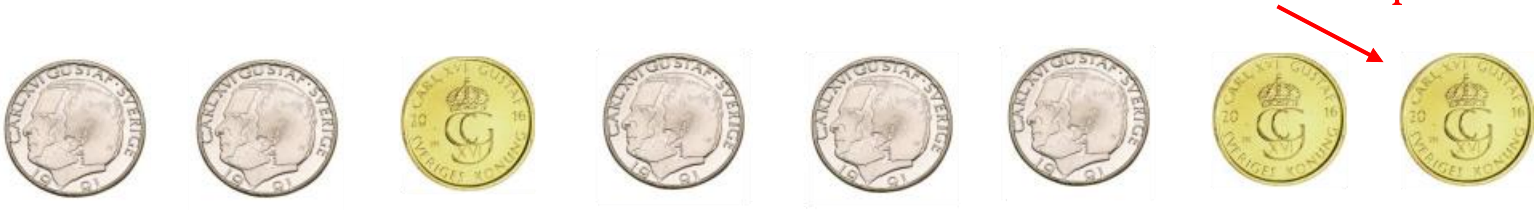
Tools to Indentify Differentially Expressed Genes (DEG's)

- Many! (available in R)
- **Question:** Is there a significantly different expression of gene “ABC” between the two (or more) conditions … or is the calculated fold change just caused by random noise?
  - edgeR
  - DESeq, DESeq2
  - DEGseq
  - baySeq
  - NOISeq
  - …
- Most of the tools rely on the Negative Binomial Distribution

Background: The Negative Binomial Distribution

- Sequence of independent Bernoulli trials, e.g. coin toss:
  - $P(succes) = P(head) = 0.6$
  - $P(failure) = P(tail) = 0.4$

$r = 3$ → stop at 3rd failure

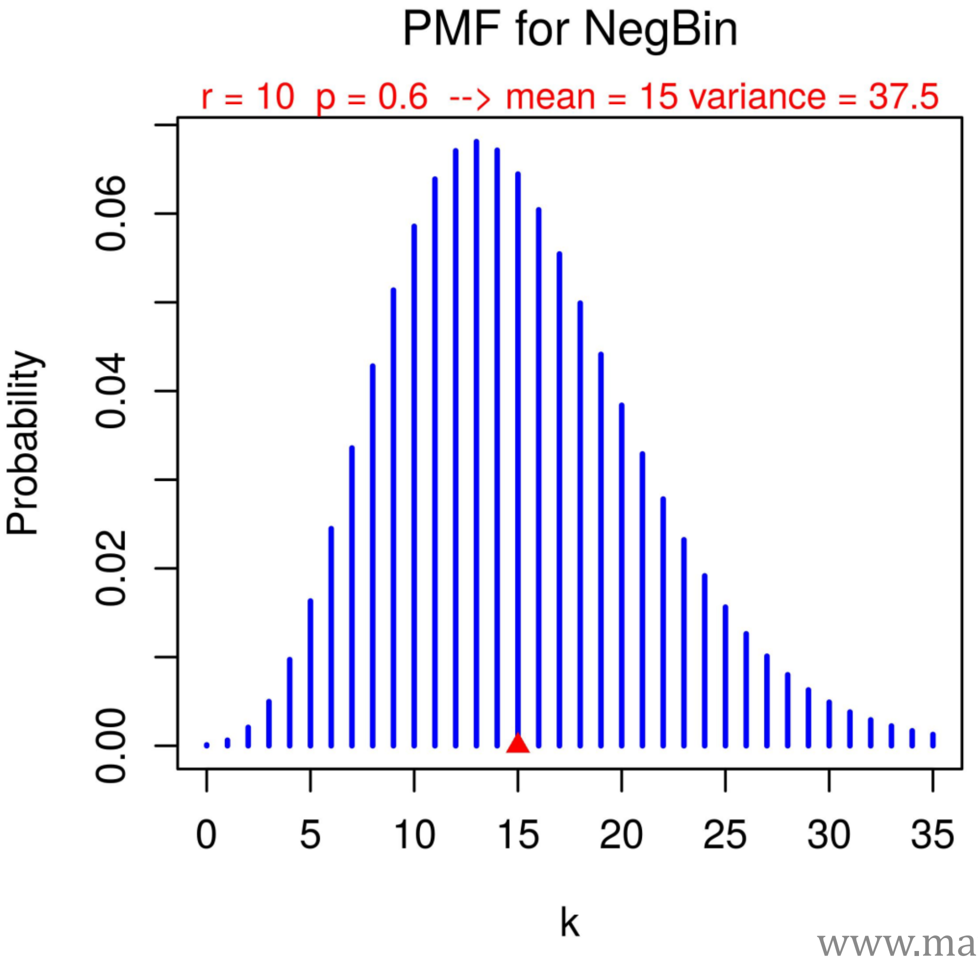


- Stop when a specified (non-random) number "r" of failures (here: tails) has occurred
- Now, count the number of successes (here: heads) in this sequence → random variable
- the Negative Binomial distribution gives the probability distribution of the number of successes (heads) in this sequence (termed k)
- $k = 0, 1, 2, \ldots, \infty$ (above: $k = 5$)

Background: Negative Binomial Distribution

$$P\left(X=k\right)=\binom{k+r-1}{k}\cdot p^{k}\cdot\left(1-p\right)^{r}$$

Probability Mass Function (PMF)



- X: number of successes (heads)
- p: probability of success
- r: number of failures (tails), parameter, not random

It is obvious that all this has nothing to do with count data!

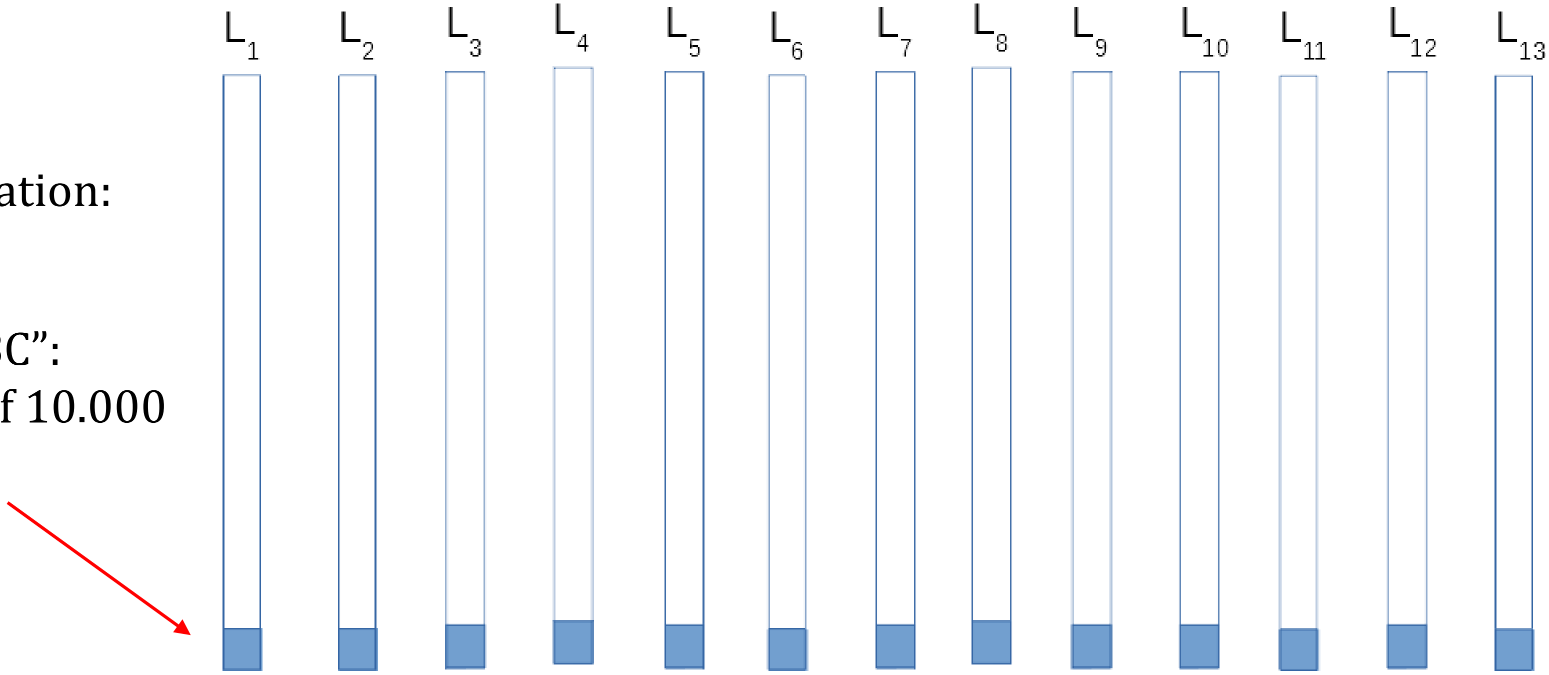
Or does it?

😕

Distribution of a particular transcript in a library

After amplification:

RNA “ABC”:
10 out of 10.000
~ 0.1%



The probability to sequence (“fish”) *k* molecules of a particular species of RNA (“ABC”) can be described by a Binomial distribution:

$$P\left(X=k\right)=\binom{n}{k}\cdot p^k\cdot\left(1-p\right)^{n-k}$$ PMF

X – number of RNA molecules “ABC” sequenced - random variable!
n - total number of reads sequenced
p - probability to obtain this particular RNA (fraction, 1/1000)

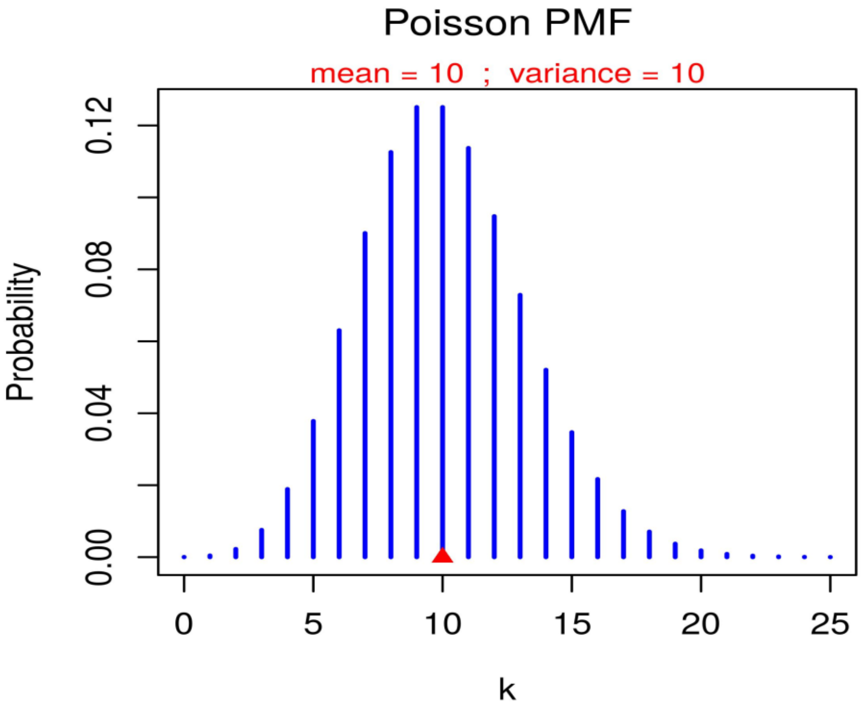
Distribution of a particular transcript in a library

If p is small and n is big, the Binomial distribution can be approximated by a Poisson distribution with the mean $n \cdot p$:

Example: “light rain”

$$P(X = k) = \frac{\mu^k}{k!} \cdot \exp(-\mu) \qquad E(X) = \mu = n \cdot p$$

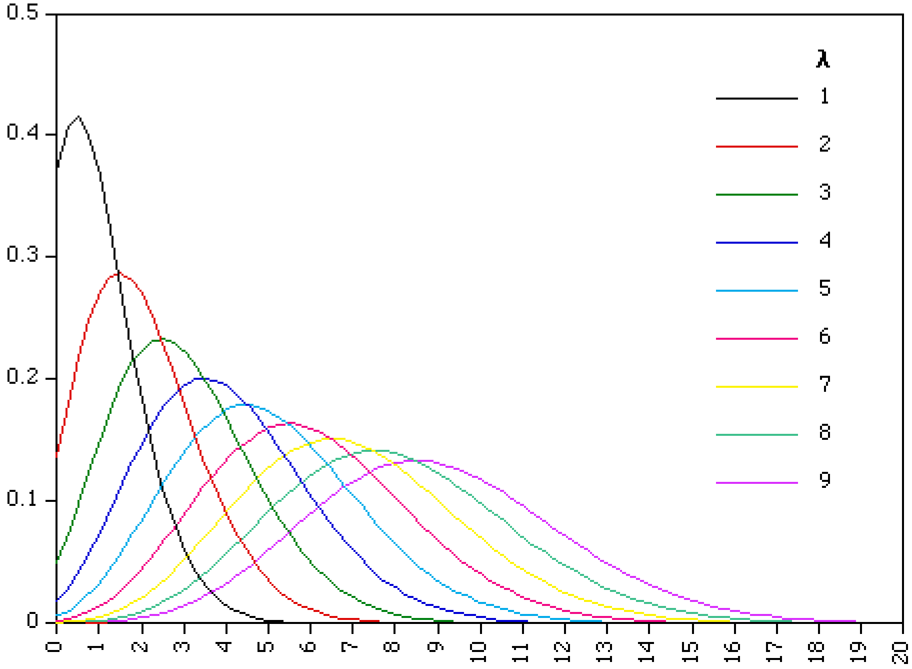
- Accordingly, the number of counts for a particular RNA species “ABC” should be Poisson-distributed.
- The distribution parameter μ (mean) is deduced from the concentration of species “ABC”, and is therefore characteristic for technical replicates.



Problem: variance = mean for Poisson distribution

Distribution of counts for a transcript when biological replicates are involved

- Different biological replicates are characterized by different concentrations of transcript “ABC”,
- … and should therefore have different distribution parameters μ (i.e. different means and variances of the Poisson).
- Consequently, a realistic distribution to describe biological replicates is a mixture (weigthed average) of Poissons with different μ’s:



Calculate a weighted average of (many) Poisson distributions! →

www.matstat.org

Mixture of Poissons

- Mixture (weighted mean) of Poisson distributions
- use the Gamma distribution as weights, ...
- ... and sum up over infinitely many Poissons → Integral

$$\begin{aligned}\alpha &= r\\ \beta &= \frac{1-p}{p}\end{aligned}$$

$$\begin{aligned}
P(X=k) &= \int_0^{+\infty} \frac{\mu^k}{k!}\cdot\exp(-\mu)\times\frac{\beta^\alpha}{\Gamma(\alpha)}\cdot\mu^{\alpha-1}\cdot e^{-\beta\mu}\,d\mu\\
&= \frac{(1-p)^r\cdot p^{-r}}{k!\cdot\Gamma(r)}\times\int_0^{+\infty}\mu^{k+r-1}\cdot e^{-\frac{\mu}{p}}\,d\mu\\
&= \frac{(1-p)^r\cdot p^{-r}}{k!\cdot\Gamma(r)}\times p^{r+k}\cdot\Gamma(r+k)\\
&= \binom{k+r-1}{k}\cdot p^k\cdot(1-p)^r
\end{aligned}$$

https://en.wikipedia.org/wiki/Negative_binomial_distribution#Overdispersed_Poisson

A weighted mixture of Poisson distributions with a Gamma mixing distribution has the same probability mass function as the Negative Binomial distribution.

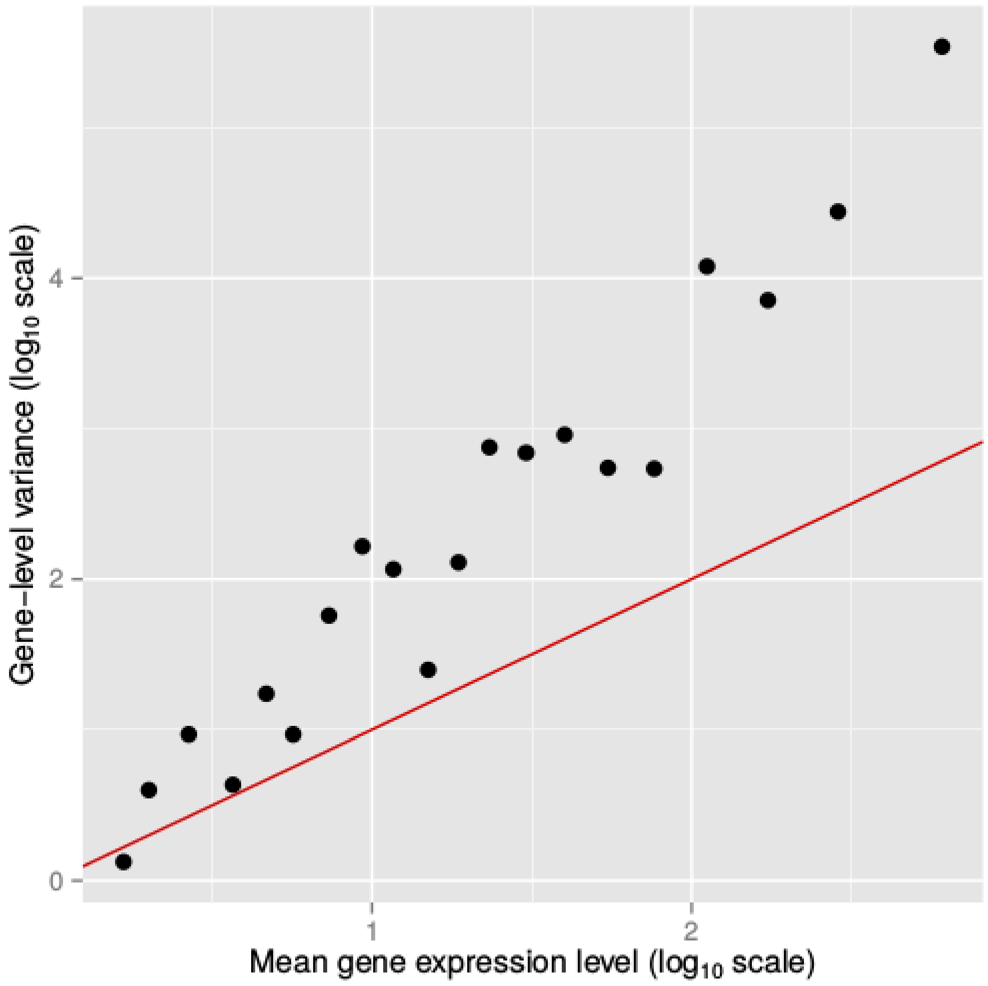
accounts for overdispersion:

$$variance = mean + \alpha\cdot mean^2$$

“overdispersed Poisson model”

Overdispersion

log(variance) versus log(mean) for a sequencing project, each point represents a transcript:



Source: Ignacio Gonzalez
"Statistical analysis of RNA-Seq data"
Tutorial

Another parametrisation for NB

r and p can be replaced by μ and σ → $NB(k, \mu, \sigma)$

$$\left.\begin{aligned} \mu &= \frac{p \cdot r}{1-p} \\ \sigma^2 &= \frac{p \cdot r}{(1-p)^2} \end{aligned}\right\}$$

solve for p and r →

$$\sigma^2 = \mu + \alpha \cdot \mu^2 \qquad \alpha = \frac{1}{r}$$

$$\left.\begin{aligned} p &= \frac{\sigma^2 - \mu}{\sigma^2} \\ r &= \frac{\mu^2}{\sigma^2 - \mu} \end{aligned}\right\}$$

$$P(k, \mu, \sigma) = \binom{k + \frac{\mu^2}{\sigma^2 - \mu} - 1}{k} \cdot \left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{\sigma^2 - \mu}} \cdot \left(\frac{\sigma^2 - \mu}{\sigma^2}\right)^k$$

PMF of the NB distribution as a function of μ and σ

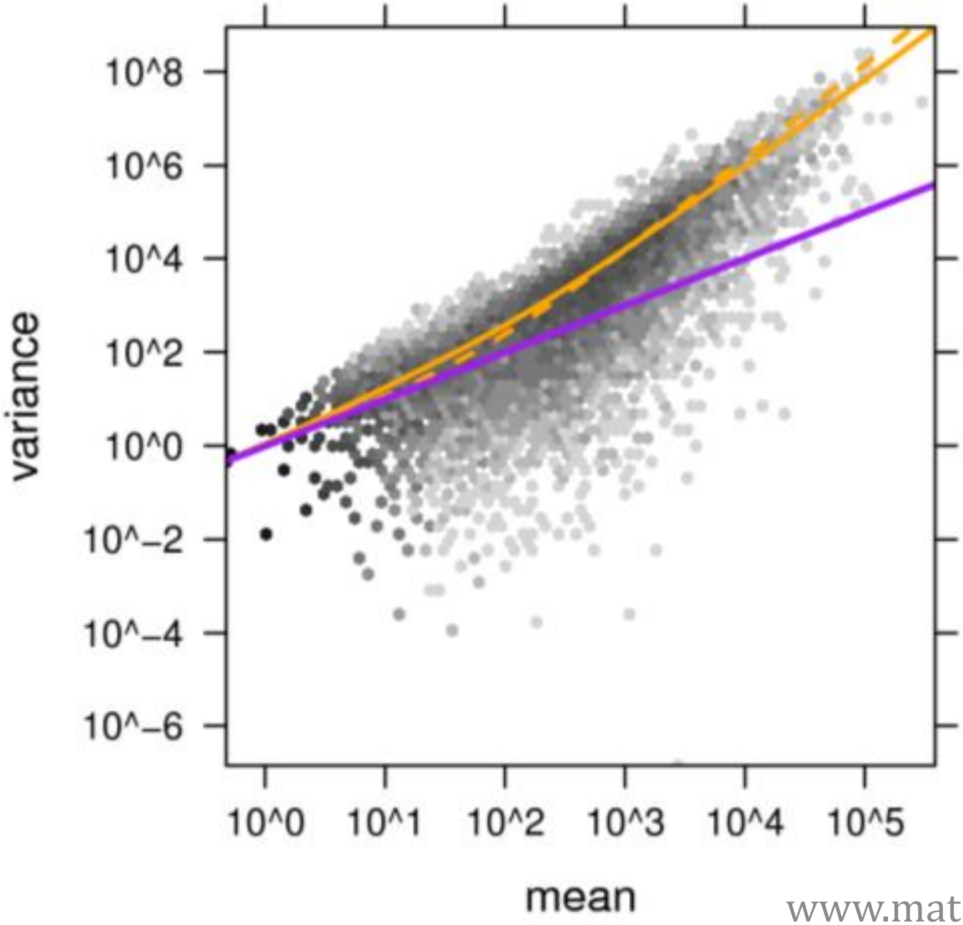
Tools: edgeR

- Robinson, McCarthy, Smyth
- R/Bioconductor package
- replicated count data (in at least one experimental condition)
- Negative Binomial distribution, overdispersed Poisson model
- empirical Bayes method: estimate overdispersion across transcripts
  - $variance = mean + \alpha \cdot mean^2$
  - "shrink dispersion towards a consensus value"
  - "borrowing information between genes"

Robinson, McCarthy, Smyth
"edgeR: a Bioconductor package for differential expression analysis of digital gene expression data"
Bioinformatics, 2010

Tools: DESeq

- Negative Binomial distribution (overdispersion)
- variance-mean relationship estimated from data (mean-dependent local regression
- " ... pool the data from genes with similar expression strength for the purpose of variance estimation" (Anders & Huber, Genome Biol., 2010)"



Purple: single Poission (variance = mean)
Orange: DESeq (piece-wise dispersion estimate)
Dashed: comparison to edgeR (single common dispersion estimate for all genes)

Anders S., Huber W.
"Differential expression analysis for sequence count data."
Genome Biol. 2010

Tools: baySeq

- Hardcastle, Kelly
- Negative Binomial Distribution (NB), accounting for overdispersion
- Empirical Bayes:
  - borrow information across the dataset
  - estimate empirical prior distributions for NB-parameters
- not restricted to pairwise comparisons, complex experimental designs possible

Hardcastle, Kelly
"baySeq: Empirical Bayesian methods for identifying
differential expression in sequence count data"
BMC Bioinformatics, 2010

Further steps

- Profiles, Protein Motifs, and Domains (after multiple alignment)
- Blast GO (http://amigo1.geneontology.org/cgi-bin/amigo/blast.cgi)
- blast2GO (commercial)
- clustering of commonly regulated genes (based of sequence similarity)

Appendix

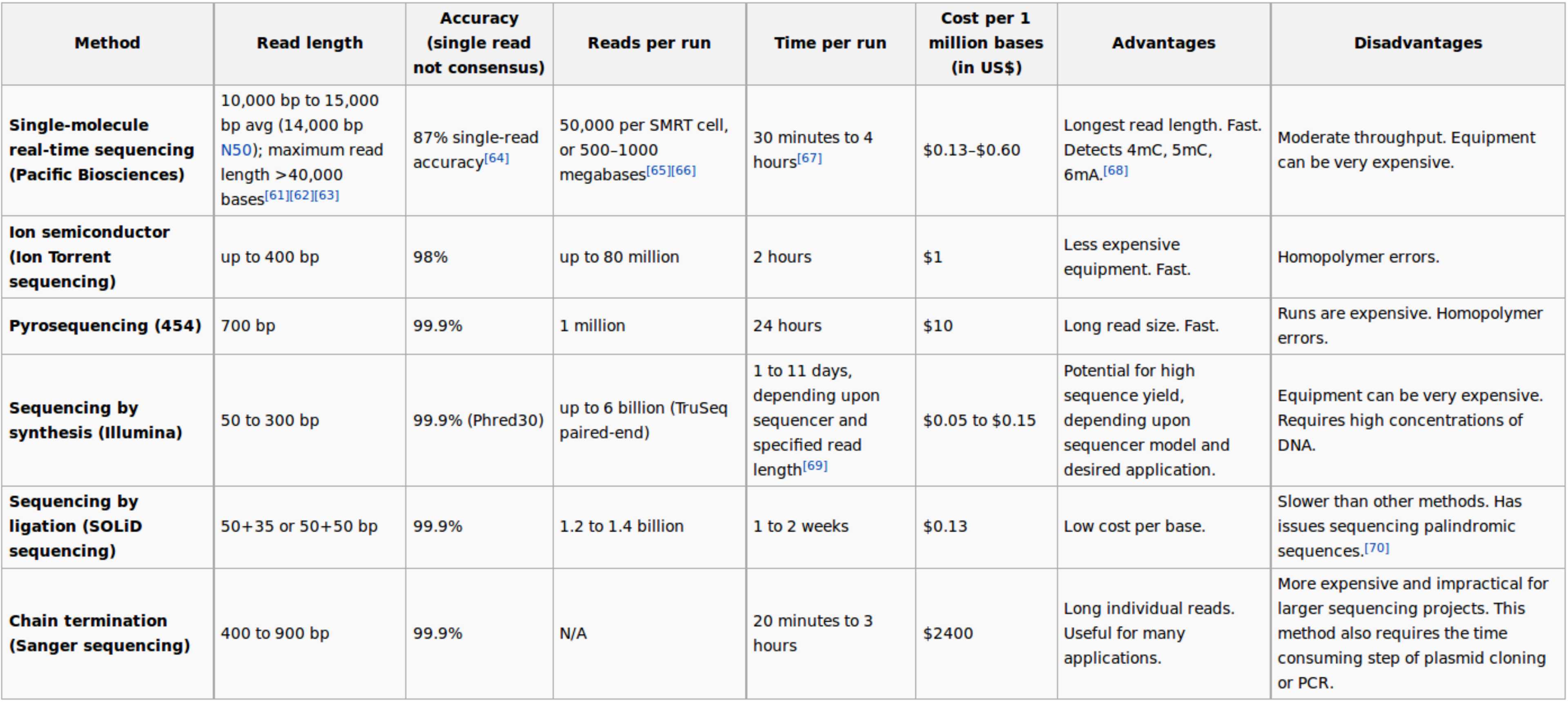
How to Exploit Differential Expression?

Uwe Menzel, 2015

uwe.menzel@matstat.de

www.matstat.org

NGS



Method	Read length	Accuracy (single read not consensus)	Reads per run	Time per run	Cost per 1 million bases (in US$)	Advantages	Disadvantages
Single-molecule real-time sequencing (Pacific Biosciences)	10,000 bp to 15,000 bp avg (14,000 bp N50); maximum read length >40,000 bases[61][62][63]	87% single-read accuracy[64]	50,000 per SMRT cell, or 500–1000 megabases[65][66]	30 minutes to 4 hours[67]	$0.13–$0.60	Longest read length. Fast. Detects 4mC, 5mC, 6mA.[68]	Moderate throughput. Equipment can be very expensive.
Ion semiconductor (Ion Torrent sequencing)	up to 400 bp	98%	up to 80 million	2 hours	$1	Less expensive equipment. Fast.	Homopolymer errors.
Pyrosequencing (454)	700 bp	99.9%	1 million	24 hours	$10	Long read size. Fast.	Runs are expensive. Homopolymer errors.
Sequencing by synthesis (Illumina)	50 to 300 bp	99.9% (Phred30)	up to 6 billion (TruSeq paired-end)	1 to 11 days, depending upon sequencer and specified read length[69]	$0.05 to $0.15	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipment can be very expensive. Requires high concentrations of DNA.
Sequencing by ligation (SOLiD sequencing)	50+35 or 50+50 bp	99.9%	1.2 to 1.4 billion	1 to 2 weeks	$0.13	Low cost per base.	Slower than other methods. Has issues sequencing palindromic sequences.[70]
Chain termination (Sanger sequencing)	400 to 900 bp	99.9%	N/A	20 minutes to 3 hours	$2400	Long individual reads. Useful for many applications.	More expensive and impractical for larger sequencing projects. This method also requires the time consuming step of plasmid cloning or PCR.

Negative Binomial Distribution

The Negative Binomial distribution arises as a continuous mixture of Poisson distributions where the mixing distribution of the Poisson mean is a Gamma distribution. That is, we can view the Negative Binomial as a Poisson(λ) distribution, where λ is itself a random variable, distributed as a Gamma distribution with shape $= r$ and scale $\theta = p/(1-p)$.
Because of this, the Negative Binomial distribution is also known as the Gamma-mixture of Poisson distributions.

(Wikipedia: Negative Binomial distribution)

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} \cdot e^{-\beta x}$$

Probability Density Function (PDF)

$\alpha > 0$ shape $x \in (0, \infty)$

$\beta > 0$ rate

Poisson Distribution

- can be applied to systems with a large number of possible events, each of which is rare
- discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time (or space) if these events occur independently and with a known average rate
- can also be used for the number of events in other specified intervals such as distance, area or volume
- **Example**: light rain

$$P\left(X=k\right)=\frac{\mu^{k}}{k!}\cdot\exp\left(-\mu\right) \qquad k=0,1,2,\ldots$$

$$\left.\begin{array}{l}\circ\ \ \text{mean} = \mu \\ \circ\ \ \text{variance} = \mu\end{array}\right\} \quad \text{cannot account for overdispersion !}$$

Overdispersed Poisson

- biological replicates do not have the same distribution parameters
- mixture of Poissons (if there is more than one biological replicate for a certain experimental condition)
- variance of the Poisson mixture can be greater than the mean → distribution (of counts) is overdispersed with respect to a Poisson distribution
- mixture is realized by letting the parameter be distributed itself
- Poisson-gamma mixture distribution: mean (μ) varies according to a Gamma distribution
- Poisson distribution with a gamma-distributed mean parameter = Negative Binomial distribution

Overdispersed Poisson

- The Negative Binomial distribution can be used as an alternative to the Poisson distribution
- It is especially useful for discrete data over an unbounded positive range whose sample variance exceeds the sample mean
- In such cases, the observations are overdispersed with respect to a Poisson distribution, for which the mean is equal to the variance, making the Poisson an unappropiate model
- Since the Negative Binomial distribution has one more parameter than the Poisson, the second parameter can be used to adjust the variance independently of the mean

The False Discovery Rate

- Define the False Discovery Proportion (FDP) to be the (unobserved) *proportion of false discoveries among total rejections.*

  As a function of threshold t (and implicitly P^m and H^m), write this as

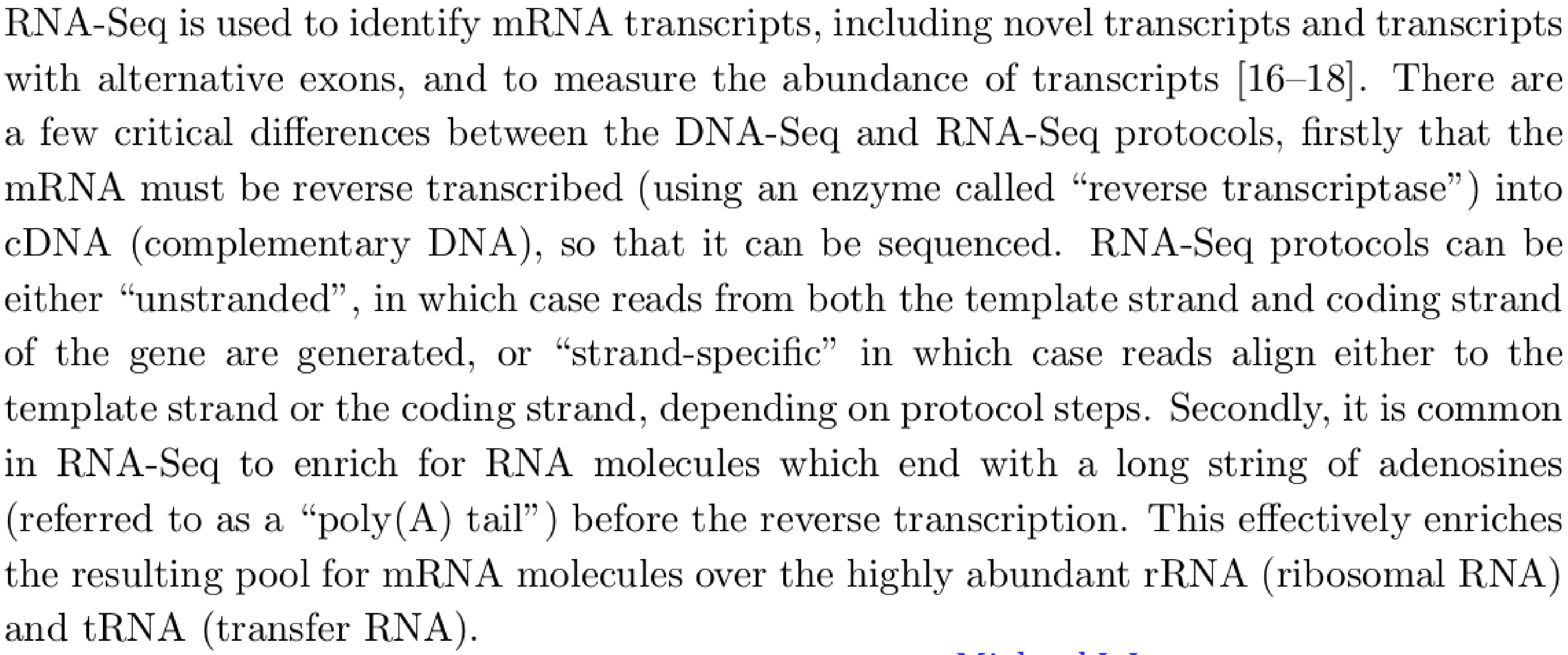
$$\text{FDP}(t) = \frac{\sum_i 1\left\{P_i \leq t\right\}(1 - H_i)}{\sum_i 1\left\{P_i \leq t\right\} + 1\left\{\text{all } P_i > t\right\}} = \frac{\#\text{False Discoveries}}{\#\text{Discoveries}}$$

- The False Discovery Rate (FDR) for a multiple testing threshold T is defined as the expected FDP using that procedure:

$$\text{FDR} = \mathbb{E}\left(\text{FDP}(T)\right).$$

Benjamini & Hochberg (1995, 2000)

RNA-Seq



RNA-Seq is used to identify mRNA transcripts, including novel transcripts and transcripts with alternative exons, and to measure the abundance of transcripts [16–18]. There are a few critical differences between the DNA-Seq and RNA-Seq protocols, firstly that the mRNA must be reverse transcribed (using an enzyme called "reverse transcriptase") into cDNA (complementary DNA), so that it can be sequenced. RNA-Seq protocols can be either "unstranded", in which case reads from both the template strand and coding strand of the gene are generated, or "strand-specific" in which case reads align either to the template strand or the coding strand, depending on protocol steps. Secondly, it is common in RNA-Seq to enrich for RNA molecules which end with a long string of adenosines (referred to as a "poly(A) tail") before the reverse transcription. This effectively enriches the resulting pool for mRNA molecules over the highly abundant rRNA (ribosomal RNA) and tRNA (transfer RNA).

Michael I. Love
Thesis FU Berlin 201

Empirical Bayes

- Empirical Bayes methods are procedures for statistical inference in which the prior distribution is estimated from the data.
- “ … shares information across all observations to improve inference.” (edgeR publication)
- This approach stands in contrast to standard Bayesian methods, for which the prior distribution is fixed before any data are observed.