

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

# RNA-seq

## Quantification and Differential Expression

Peter N. Robinson

Institut für Medizinische Genetik und Humangenetik  
Charité Universitätsmedizin Berlin

Genomics: Lecture #12

# Today

RNA-seq (2)

**Peter N.  
Robinson**

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- Gene Expression per RNA-seq
- Sources of bias, normalization, and problems

# What is Differential Expression?

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

## Differential Expression

A gene is declared differentially expressed if an observed difference or change in read counts between two experimental conditions is statistically significant, i.e. if the difference is greater than what would be expected just due to random variation.

- Statistical tools for microarrays were based on numerical intensity values
- Statistical tools for RNA-seq instead need to analyze read-count distributions

# Outline

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- 1 RNA-seq
- 2 RPKM and Length Bias
- 3 Fisher's exact test
- 4 Poisson
- 5 Likelihood Ratio Test
- 6 Negative Binomial

# RNA-seq: From Counts to Expression

## RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

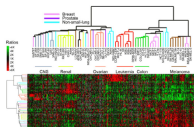
Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- For many applications, we are interested in measuring the absolute or relative expression of each mRNA in the cell
- Microarrays produced a numerical estimate of the relative expression of (nearly) all genes across the genome (although it was usually difficult to distinguish between the various isoforms of a gene)
- How can we do this with RNA-seq? Do read counts correspond directly to gene expression?



# RNA-seq: Workflow

RNA-seq (2)

Peter N. Robinson

RNA-seq

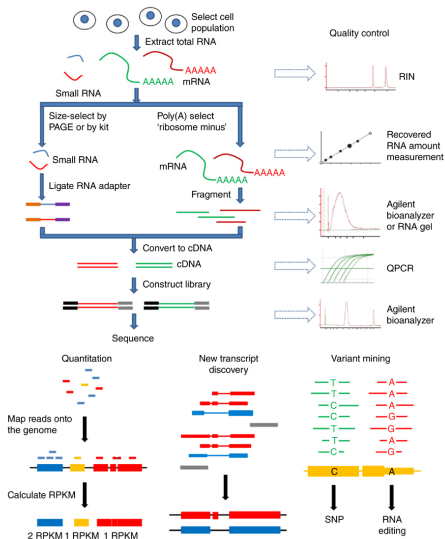
RPKM

Fisher's exact test

Poisson

LRT

Negative Binomial



# Outline

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- 1 RNA-seq
- 2 RPKM and Length Bias**
- 3 Fisher's exact test
- 4 Poisson
- 5 Likelihood Ratio Test
- 6 Negative Binomial

# RNA-seq: From Counts to Expression

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Current RNA-seq protocols use an mRNA fragmentation approach prior to sequencing to gain sequence coverage of the whole transcript. Thus, the total number of reads for a given transcript is proportional to the expression level of the transcript multiplied by the length of the transcript.

- In other words a long transcript will have more reads mapping to it compared to a short gene of similar expression.
- Since the power of an experiment is proportional to the sampling size, there is more power to detect differential expression for longer genes.

Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 4:14.



# RNA-seq: Length Bias

## RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

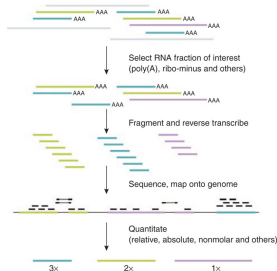
Poisson

LRT

Negative  
Binomial

- Let  $X$  be the measured number of reads in a library mapping to a specific transcript.
- The expected value of  $X$  is proportional to the total number of transcripts  $N$  times the length of the gene  $L$

$$\mathbb{E}[X] \propto N \cdot L$$



# Length Normalization

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

For this reason, most RNA-seq analysis involves some sort of length normalization. The most commonly used is RPKM.

- RPKM: **Reads per kilobase transcript per million reads**

$$RPKM(X) = \frac{10^9 \cdot C}{N \cdot L} \quad (1)$$

- C is the number of mappable reads that fell onto the gene's exons
- N is the total number of mappable reads in the experiment
- L is the total length of the exons in base pairs

Example: 1kb transcript with 2000 alignments in a sample of 10 million reads (out of which 8 million reads

can be mapped) will have  $RPKM = \frac{10^9 \cdot 2000}{8 \times 10^6 \cdot 1000} = \frac{2 \times 10^{12}}{8 \times 10^9} = 250$

# Length Normalization

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- RPKM: **Reads per KB per million reads**

$$RPKM(X) = \frac{10^9 \cdot C}{N \cdot L}$$

- Note that this formula can also be written as

$$\begin{aligned} RPKM(X) &= \frac{\text{Reads mapped to transcript}}{\frac{\text{total reads}}{1,000,000} \cdot \text{transcript length in kb}} \\ &= \frac{\text{Reads mapped to transcript}}{\frac{\text{total reads}}{1,000,000} \cdot \frac{\text{transcript length in bp}}{1000}} \\ &= \frac{10^9 \times \text{Reads mapped to transcript}}{\text{total reads} \cdot \text{transcript length in bp}} \end{aligned}$$

# RPKM: Question

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

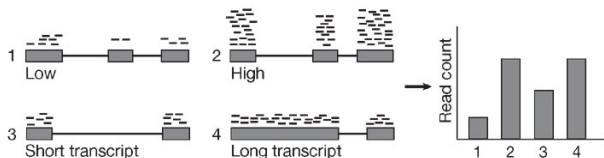
Poisson

LRT

Negative  
Binomial

$$RPKM(X) = \frac{10^9 \cdot C}{N \cdot L}$$

- Question: What are the RPKM-corrected expression values and why?



# RPKM: Answer

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

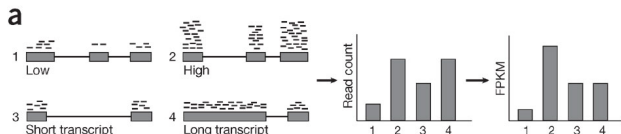
Poisson

LRT

Negative  
Binomial

$$RPKM(X) = \frac{10^9 \cdot C}{N \cdot L}$$

- Note especially normalization for fragment length (transcripts 3 and 4)



Graphic credit: Garber et al. (2011) *Nature Methods* 8:469–477. Note that the authors here use the related term FPKM, **Fragments per KB per million reads**, which is suitable for paired-end reads (we will not cover the details here).

# RPKM: Another question

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

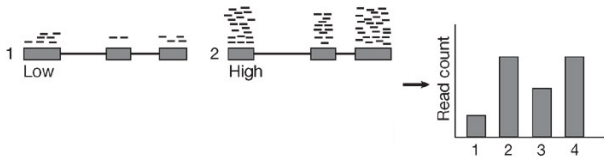
Poisson

LRT

Negative  
Binomial

$$RPKM(X) = \frac{10^9 \cdot C}{N \cdot L}$$

- What if now assume that the same gene is sequenced in two libraries, and the total read count in library 1 was  $\frac{1}{10}$  of that in library 2? In which library is the gene more highly expressed?



# Length Normalization

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Unfortunately, this kind of length normalization does not solve all of our problems.

- In essence, RPKM and related length normalization procedures produce an unbiased estimate of the mean of the gene's expression
- However, they do not compensate for the effects of the length bias on the variance of our estimate of the gene's expression
- It is instructive to examine the reasons for this<sup>1</sup>

---

<sup>1</sup>This was first noted by Oshlack A (2009) *Biol Direct* 4:14, from whom the following slides are adapted.

# RNA-seq: Length Bias

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

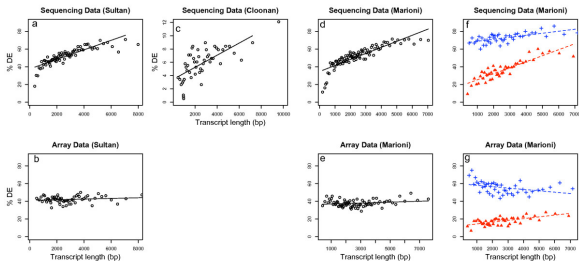
Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Ability to detect DE is strongly associated with transcript length for RNA-seq. In contrast, no such trend is observed for the microarray data



- data is binned according to transcript length
- percentage of transcripts called differentially expressed using a statistical cut-off is plotted (points)
- Oshlack 2009



# RNA-seq: Length Bias

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- As noted above, the expected value of  $X$  is proportional to the total number of transcripts  $N$  times the length of the gene  $L$

$$\mu = \mathbb{E}[X] = cN \cdot L$$

- $c$  is the proportionality constant.
- Assuming the data is distributed as a Poisson random variable, the variance is equal to the mean.

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mu$$

# RNA-seq: Length Bias

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- Under these assumptions, it is reasonable to check if the difference in counts from a particular gene between two samples of the same library size is significantly different from zero using a  $t$ -test:

$$T = \frac{D}{SE(D)} = \frac{cN_1L - cN_2L}{\sqrt{cN_1L + cN_2L}} \quad (2)$$

In the  $t$  test,  $D$  is the difference in the sample means, and  $SE(D)$  is the standard error of  $D$ .

Recall that with the  $t$  test, the null hypothesis is rejected if  $|T| > t_{1-\alpha/2, \nu}$  where  $t_{1-\alpha/2, \nu}$  is the critical value of the  $t$  distribution with  $\nu$  degrees of freedom

# RNA-seq: Length Bias

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- It can be shown that the power of the  $t$  test depends on  $\frac{E(D)}{SE(D)} = \delta$

$$\delta = \frac{\mathbb{E}[D]}{SE(D)} = \frac{\mathbb{E}[cN_1L - cN_2L]}{\sqrt{cN_1L + cN_2L}} \propto \sqrt{L} \quad (3)$$

- Thus, the power of the test is proportional to the square root of  $L$ .
- Therefore for a given expression level the test becomes more significant for longer transcript lengths!
- It is simple to show that dividing by gene length (which is essentially what RPKM does) does not correct this bias<sup>2</sup>

# RNA-seq: Differential Expression

## RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson


LRT

Negative  
Binomial

Today, we will examine some of the methods that have been used to assess differential expression in RNAseq data.

- Simple(st) case: two-sample comparison without replicates<sup>3</sup>
- Modeling read counts with a Poisson distribution
- Overdispersion and the negative binomial distribution

---

<sup>3</sup>For so called didactic purposes only, do not do this at home! 

# Outline

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- 1 RNA-seq
- 2 RPKM and Length Bias
- 3 Fisher's exact test**
- 4 Poisson
- 5 Likelihood Ratio Test
- 6 Negative Binomial

# Fisher's exact test

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Let us get warm by examining an observational study with no biological replication. For instance, one sample each is processed and sequenced from the brain and the liver. What can we say about differential expression?

- The Fisher's exact test can be used for RNA-seq data without replicates, proceeding on a gene-by-gene basis and organizing the data in a  $2 \times 2$  contingency table

## $2 \times 2$ contingency table

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

	condition 1	condition 2	Total
Gene $x$	$n_{11}$	$n_{12}$	$n_{11} + n_{12}$
Remaining genes	$n_{21}$	$n_{22}$	$n_{21} + n_{22}$
Total	$n_{11} + n_{21}$	$n_{12} + n_{22}$	$N$

- The cell counts  $n_{ki}$  represent the observed read count for gene  $x$  ( $k = 1$ ) or the remaining genes ( $k = 2$ ) for condition  $i$  (e.g.,  $i = 1$  for brain and  $i = 2$  for liver)
- The  $k^{\text{th}}$  marginal row total is then  $n_{k1} + n_{k2}$
- $n_{1i} + n_{2i}$  is the marginal total for condition  $i$
- $N$  is the grand total

# Fisher's exact test

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Fisher's exact test for RNAseq counts tests the null hypothesis that the conditions (columns) that the proportion of counts for some gene  $x$  amongst two samples is the same as that of the remaining genes, i.e., the null hypothesis can be interpreted as  $\frac{\pi_{11}}{\pi_{12}} = \frac{\pi_{21}}{\pi_{22}}$ , where  $\pi_{ki}$  is the true but unknown proportion of counts in cell  $ki$ .

- Let us explain how the Fisher's exact test works. We will need to examine the binomial and the hypergeometric distributions



# Fisher's exact test

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

The binomial coefficient provides a general way of calculating the number of ways  $k$  objects can be chosen from a set of  $n$  objects. Recall that  $n! = n \times (n-1) \times (n-2) \times \dots \times 2 \times 1$  is the number of ways of arranging  $n$  objects in a series.

In order to calculate the number of ways of observing  $k$  “heads” in  $n$  coin tosses, we first examine the sequence of tosses consisting of  $k$  “heads” followed by  $n - k$  “tails”:

$H$	$H$	$H$	$\dots$	$H$	$H$	$T$	$\dots$	$T$	$T$	$T$
1	2	3	$\dots$	$k-1$	$k$	$k+1$	$\dots$	$n-2$	$n-1$	$n$

# Fisher's exact test

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Each of the  $n!$  rearrangements of the numbers  $1, 2, \dots, n$  defines a different rearrangement of the letters  $HHH \dots HHTT \dots TT$ . However, not all of the rearrangements change the order of the  $H$ 's and the  $T$ 's. For instance, exchanging the first two positions leaves the order  $HHH \dots HHTT \dots TT$  unchanged.

Therefore, to calculate the number of rearrangements that lead to different orderings of the  $H$ 's and the  $T$ 's (e.g.,  $HTH \dots HHTT \dots TH$ ), we need to correct for the reorderings that merely change the  $H$ 's or the  $T$ 's among themselves.

# Fisher's exact test

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Noting that there are  $k!$  ways of reordering the “heads” and  $(n - k)!$  ways of reordering the “tails,”

- it follows that there are  $\binom{n}{k}$  different ways of rearranging  $k$  “heads” and  $n - k$  “tails.” This quantity<sup>4</sup> is known as the binomial coefficient.

$$\binom{n}{k} = \frac{n!}{(n - k)!k!} \quad (4)$$

---

<sup>4</sup>  $\binom{n}{k}$  should be read as “ $n$  choose  $k$ ”.

# Fisher's exact test

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Getting back to our RNAseq data, Fisher showed that the probability of getting a certain set of values in a  $2 \times 2$  contingency table is given by the hypergeometric distribution

	condition 1	condition 2	Total
Gene x	$n_{11}$	$n_{12}$	$n_{11} + n_{12}$
Remaining genes	$n_{21}$	$n_{22}$	$n_{21} + n_{22}$
Total	$n_{11} + n_{21}$	$n_{12} + n_{22}$	$N$

$$p = \frac{\binom{n_{11} + n_{12}}{n_{11}} \binom{n_{21} + n_{22}}{n_{21}}}{\binom{N}{n_{11} + n_{21}}} \quad (5)$$

# Fisher's exact test

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

$$p = \frac{\binom{n_{11}+n_{12}}{n_{11}} \binom{n_{21}+n_{22}}{n_{21}}}{\binom{n}{n_{11}+n_{21}}}$$

- This expression can be interpreted based on the total number of ways of choosing items to obtain the observed distribution of counts. If there are lots of different ways of obtaining a given count distribution, it is not that surprising (not that statistically significant) and *vice versa*
- Thus,  $\binom{n_{11}+n_{12}}{n_{11}}$  is the number of ways of choosing  $n_{11}$  reads for gene  $x$  in condition 1 from the total number of reads for that gene in conditions 1 and 2.
- $\binom{n_{21}+n_{22}}{n_{21}}$  is the number of ways of choosing  $n_{21}$  reads for the remaining genes in condition 1 from the total number of reads for the remaining genes in conditions 1 and 2.
- $\binom{n}{n_{11}+n_{21}}$  is the total number of ways of choosing all the reads for condition 1 from the reads in both conditions.

# Fisher's exact test

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

To make a hypothesis test out of this, we need to calculate the probability of observing some number  $k$  or more reads  $n_{11}$  in order to have a statistical test. In this case, the sum over the tail of the hypergeometric distribution is known as the *Exact Fisher Test*:

$$p(\text{read count} \geq n_{11}) = \sum_{k=n_{11}}^{n_{11}+n_{12}} \frac{\binom{k+n_{12}}{k} \binom{n_{21}+n_{22}}{n_{21}}}{\binom{n}{k+n_{21}}} \quad (6)$$

- We actually need to calculate a two-sided Fisher exact test unless we are testing explicitly for overexpression in one of the two conditions
- We would thus add the probability for the other upper tail<sup>5</sup>

<sup>5</sup>There are other methods that we will not mention here.

# Problems

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

The fundamental problem with generalizing results gathered from unreplicated data is a complete lack of knowledge about biological variation. Without an estimate of variability within the groups, there is no sound statistical basis for inference of differences between the groups.

© Original Artist  
Reproduction rights obtainable from  
[www.CartoonStock.com](http://www.CartoonStock.com)



search ID: jhjn35

# Outline

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- 1 RNA-seq
- 2 RPKM and Length Bias
- 3 Fisher's exact test
- 4 Poisson**
- 5 Likelihood Ratio Test
- 6 Negative Binomial



# Poisson Model

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

In this lecture we will begin to explore some of the issues surrounding more realistic models of differential expression in RNAseq data. We will now examine how to perform an analysis for differential expression on the basis of a Poisson model

- Imagine we have count data for some list of genes  $g_1, g_2, \dots$  with technical and biological replicates corresponding to two conditions we want to compare
- We will let  $X \sim \text{Poisson}(\lambda)$  be a random variable representing the number of reads falling in  $g$

# Question...

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Why might it be appropriate to model read counts as a Poisson process?



# Justification of Poisson for RNA seq

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

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

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

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



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

# Justification of Poisson for RNA seq

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Imagine we don't know the number  $n$  of trials that will happen. Instead, we only know the average number of successes per interval.

- Define a number  $\lambda = np$  as the average number of successes per interval.
- Thus  $p = \frac{\lambda}{n}$
- Note that in contrast to a binomial situation, we also do not know how many times success did not happen (how many trials there were)

# Justification of Poisson for RNA seq

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Now let's substitute  $p = \frac{\lambda}{n}$  into the binomial distribution, and take the limit as  $n$  goes to infinity

$$\begin{aligned}\lim_{n \rightarrow \infty} p(X = k) &= \lim_{n \rightarrow \infty} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \lim_{n \rightarrow \infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \left(\frac{\lambda^k}{k!}\right) \lim_{n \rightarrow \infty} \frac{n!}{(n-k)!} \frac{1}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}\end{aligned}$$

# Justification of Poisson for RNA seq

RNA-seq (2)

Peter N.  
Robinson

Let's look closer at the limit, term by term

$$\begin{aligned}\lim_{n \rightarrow \infty} \frac{n!}{(n-k)!} \frac{1}{n^k} &= \lim_{n \rightarrow \infty} \frac{n(n-1)\dots(n-k)(n-k-1)\dots(2)(1)}{(n-k)(n-k-1)\dots(2)(1)} \frac{1}{n^k} \\ &= \lim_{n \rightarrow \infty} \frac{n(n-1)\dots(n-k+1)}{n^k} \\ &= \lim_{n \rightarrow \infty} \frac{(n)}{n} \frac{(n-1)}{n} \dots \frac{(n-k+1)}{n} \\ &= 1\end{aligned}$$

- The final step follows from the fact that

$$\lim_{n \rightarrow \infty} \frac{n-j}{n} = 1 \text{ for any fixed value of } j.$$

# Justification of Poisson for RNA seq

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Continuing with the middle term  $\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n$

- Recalling that  $e^x$  can also be defined as  $e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n$ , we see that the above limit is equal to  $e^{-\lambda}$

# Justification of Poisson for RNA seq

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Continuing with the final term  $\left(1 - \frac{\lambda}{n}\right)^{-k}$

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^{-k} = (1)^{-k} = 1$$

- Putting everything together, we have

$$\lim_{n \rightarrow \infty} \frac{n!}{(n-k)!} \frac{1}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} = 1 \cdot e^{-\lambda} \cdot 1 = e^{-\lambda}$$



# Justification of Poisson for RNA seq

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- We can now see the familiar Poisson distribution

$$\begin{aligned} p(X = k) &= \left(\frac{\lambda^k}{k!}\right) \lim_{n \rightarrow \infty} \frac{n!}{(n-k)!} \frac{1}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &= \left(\frac{\lambda^k}{k!}\right) e^{-\lambda} \\ &= \frac{\lambda^k e^{-\lambda}}{k!} \end{aligned}$$

# Poisson (mean = variance)

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

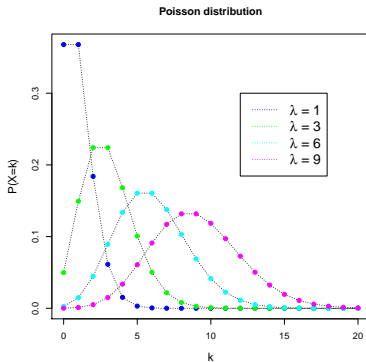
RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial



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

# Outline

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- 1 RNA-seq
- 2 RPKM and Length Bias
- 3 Fisher's exact test
- 4 Poisson
- 5 Likelihood Ratio Test**
- 6 Negative Binomial

# Likelihood ratio test

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

The likelihood ratio test is a statistical test that is used by many RNAseq algorithms to assess differential expression. It compares the likelihood of the data assuming no differential expression (null model) against the likelihood of the data assuming differential expression (alternative model).

$$D = -2 \log \frac{\text{likelihood of null model}}{\text{likelihood of alternative model}} \quad (7)$$

- It can be shown that  $D$  follows a  $\chi^2$  distribution, and this can be used to calculate a  $p$  value
- We will explain the LRT using an example from football and then show how it can be applied to RNAseq data

# Likelihood ratio test

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

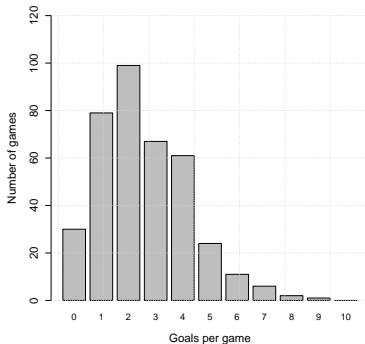
Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Let's say we are interested in the average number of goals per game in World Cup football matches. Our null hypothesis is that there are three goals per game.



# Goals per Game: MLE

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

We first decide to model goals per game as a Poisson distribution and to calculate the Maximum Likelihood Estimate (MLE) of this quantity

Goals	Frequency
0	30
1	79
2	99
3	67
4	61
5	24
6	11
7	6
8	2
9	1
10+	0
Total	380

**Likelihood:** View a probability distribution as a function of the parameters given a set of observed data

$$\mathcal{L}(\Theta|X) = \prod_{i=1}^N \text{Poisson}(x_i, \lambda)$$

Goal of MLE: find the value of  $\lambda$  that maximizes this expression for the data we have observed

# Goals per Game: MLE

RNA-seq (2)

Peter N.  
Robinson

$$\begin{aligned}\mathcal{L}(\Theta|X) &= \prod_{i=1}^N \text{Poisson}(x_i, \lambda) \\ &= \prod_{i=1}^N \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \\ &= \frac{e^{-N\lambda} \lambda^{\sum_{i=1}^N x_i}}{\prod_{i=1}^N x_i!}\end{aligned}$$

- Note we generally maximize the log likelihood, because it is usually easier to calculate and identifies the same maximum because of the monotonicity of the logarithm.

$$\log \mathcal{L}(\Theta|X) = -N\lambda + \sum_{i=1}^N x_i \log \lambda - \sum_{i=1}^N \log x_i!$$

# Goals per Game: MLE

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

To find the max, we take the first derivative with respect to  $\lambda$  and set equal to zero.

$$\frac{d}{d\lambda} \log \mathcal{L}(\Theta|X) = -N + \frac{\sum_{i=1}^N x_i}{\lambda} - 0$$

This reassuringly leads to

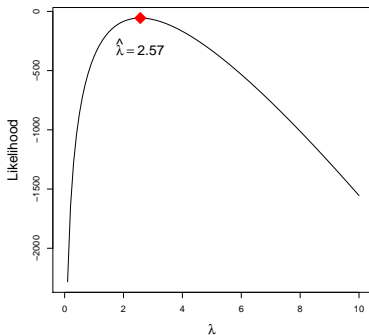
$$\hat{\lambda} = \frac{\sum_{i=1}^N x_i}{N} = \bar{x} \quad (8)$$



# Goals per Game: MLE

Our MLE for the number of goals per game is then simply

$$\hat{\lambda} = \frac{\sum_{i=1}^N x_i}{N} = \frac{\sum_{i=1}^{380} x_i}{380} = \frac{975}{380} = 2.57$$



The maximum likelihood estimate maximizes the likelihood of the data

# Likelihood Ratio Test

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- Evaluate the log-Likelihood under  $H_0$
- Evaluate the maximum log-Likelihood under  $H_a$
- Any terms not involving the parameter (here:  $\lambda$ ) can be ignored
- Under null hypothesis (and large samples), the following statistic is approximately  $\chi^2$  with 1 degree of freedom (number of constraints under  $H_0$ )

$$LRT = -2 \left[ \log \mathcal{L}(\theta_0, x) - \log \mathcal{L}(\hat{\theta}, x) \right] \quad (9)$$

# LRT: Goals per game

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Let's say that our null hypothesis is that the average number of goals per game is 3, i.e.,  $\lambda_0 = 3$ , and the executives of a private network will only get their performance bonus from the advertisers if this is true during the cup, because games with less or more goals are considered boring by many viewers.

Under the null, we have:

$$H_0 : \log \mathcal{L}(\lambda_0|X) = -380\lambda_0 + \sum_{i=1}^{380} x_i \log \lambda_0 - \sum_{i=1}^{380} x_i! \quad (10)$$

The alternative:

$$H_a : \log \mathcal{L}(\hat{\lambda}|X) = -380\hat{\lambda} + \sum_{i=1}^{380} x_i \log \hat{\lambda} - \sum_{i=1}^{380} x_i! \quad (11)$$

# LRT: Goals per game

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- To calculate the LRT, note that we can ignore the term  $\sum_{i=1}^{380} x_i!$
- Recall  $\lambda_0 = 3$  and  $\hat{\lambda} = 2.57$

$$\log \mathcal{L}(\lambda_0 | X) = -380 \times 3 + 975 \times \log 3 = -68.85$$

$$\log \mathcal{L}(\hat{\lambda} | X) = -380 \times 2.57 + 975 \times \log 2.57 = -56.29$$

Our test statistic is thus

$$LRT = -2 [-68.85 - (-56.29)] = 25.12$$

# LRT: Goals per game

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Finally, we compare the result from the LRT with the critical value for the  $\chi^2$  distribution with one degree of freedom

$$25.12 \gg \chi_{0.05,1}^2 = 3.84$$

- Thus, the result of the LRT is clearly significant at  $\alpha = 0.05$
- We can reject the null hypothesis that the number of goals per game is 3
- No bonus this year...

# LRT: RNAseq

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

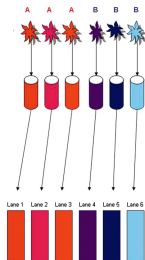
Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Marioni et al. use the LRT to investigate RNAseq samples for differential expression between two conditions A and B



# LRT: RNAseq

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- $x_{ijk}$ : number of reads mapped to gene  $j$  for the  $k^{\text{th}}$  lane of data from sample  $i$
- Then we can assume that  $x \sim \text{Poisson}(\lambda_{ijk})$
- $\lambda_{ijk} = c_{ik}\nu_{ijk}$  represents the (unknown) mean of the Poisson distribution, where  $c_{ik}$  represents the total rate at which lane  $k$  of sample  $i$  produces reads and  $\nu_{ijk}$  represents the rate at which reads map to gene  $j$  (in lane  $k$  of sample  $i$ ) relative to other genes.
- Note that  $\sum_{j'} \nu_{ij'k} = 1$ .

# LRT: RNAseq

## RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- The null hypothesis of no differential expression corresponds to  $\nu_{ijk} = \nu_j$  for gene  $j$  in all samples
- The alternative hypothesis corresponds to  $\nu_{ijk} = \nu_j^A$  for samples in group  $A$ , and  $\nu_j^B$  for samples in group  $B$  with  $\nu_j^A \neq \nu_j^B$ .



# LRT: RNAseq

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

Under the null, we have:

$$H_0 : \log \mathcal{L}(\lambda_0|X) = -N\lambda_0 + \sum_{i=1}^n x_i \log \lambda_0 - \sum_{i=1}^n x_i! \quad (12)$$

The alternative:

$$H_a : \log \mathcal{L}(\hat{\lambda}|X) = -N_A \hat{\lambda}_A - N_B \hat{\lambda}_B + \sum_{i=1}^{n_a} x_i \log \hat{\lambda}_A + \sum_{i=1}^{n_b} x_i \log \hat{\lambda}_B - \sum_{i=1}^n x_i! \quad (13)$$

- Where the total count for gene  $i$  in sample  $A$  is  $N_A$  and  $N_A + N_B = N$ , and the total number of samples in  $A$  and  $B$  is given by  $n_a$  and  $n_b$ , with the total number of samples  $n = n_a + n_b$ .

# LRT: RNAseq

## RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- The authors then used a LRT and calculated  $p$  values for each gene based on a  $\chi^2$  distribution with one degree of freedom, quite analogous to the football example
- By comparing five lanes each of liver-versus-kidney samples. At an FDR of 0.1%, they identified 11,493 genes as differentially expressed between the samples (94% of these had an estimated absolute  $\log_2$ -fold change  $> 0.5$ ; 71%  $> 1$ ).

Marioni JC et al. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**:1509-17.

# LRT: RNAseq

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

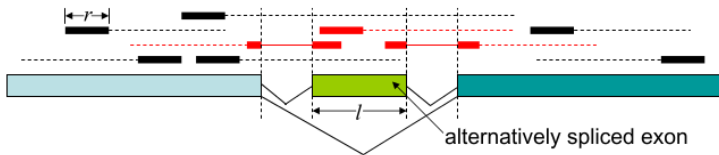
Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- Newer methods have adapted the LRT or variants thereof to examine the differential expression of the individual isoforms of a gene



# Outline

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

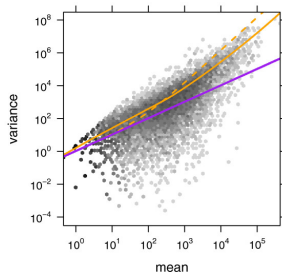
- 1 RNA-seq
- 2 RPKM and Length Bias
- 3 Fisher's exact test
- 4 Poisson
- 5 Likelihood Ratio Test
- 6 Negative Binomial**

# Problems with Poisson

RNA-seq (2)

Peter N.  
Robinson

Many studies have shown that the variance grows faster than the mean in RNAseq data. This is known as **overdispersion**.



- Mean count vs variance of RNA seq data. Orange line: the fitted observed curve. Purple: the variance implied by the Poisson distribution.

# Negative Binomial

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

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.

- The negative binomial has two parameters, the mean  $p \in ]0, 1[$  and  $r \in \mathbb{Z}$ , where  $p$  is the probability of a single success and  $r - 1$  is the total number of successes and  $k$  the total number of failures in  $x + r - 1$  trials.

$$NB(K = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \quad (14)$$

# Negative Binomial

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

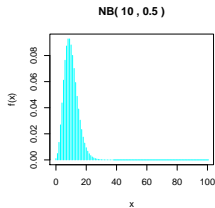
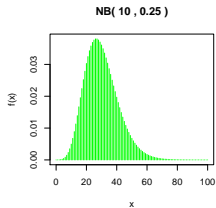
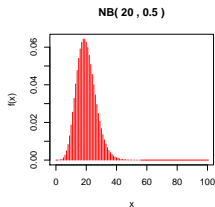
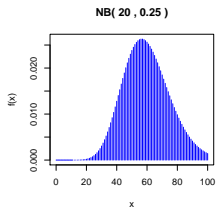
Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

The negative binomial distribution  $\text{NB}(r,p)$



# What happens if our estimate of the variance is too low?

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

For simplicity's sake, let us consider this question using the  $t$  distribution

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (15)$$

- Here,  $\bar{x}$  is the sample mean,  $\mu_0$  represents the null hypothesis that the population mean is equal to a specified value  $\mu_0$ ,  $s$  is the sample standard deviation, and  $n$  is the sample size





# What happens if our estimate of the variance is too low?

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

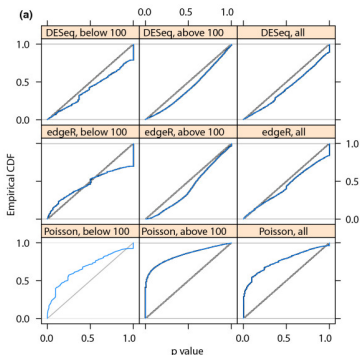
RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial



- empirical cumulative distribution functions (ECDFs) for P values from a comparison of two technical replicates
- No genes are truly differentially expressed, and the ECDF curves (blue) should remain below the diagonal (gray).
- Top row: DESeq (Negative binomial plus flexible data-driven relationships between mean and variance); middle row edgeR (Negative binomial); bottom row: Poisson-based  $\chi^2$  test

# Summary

## RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- Many issues are to be taken into account to determine expression levels and differential expression for RNA-seq data
- There are major bias issues related to transcript length and other factors<sup>6</sup>
- Many methods have been developed to assess differential expression in RNAseq data. Note that many of the assumptions that have been applied successfully previously for microarray data to not work well with RNA-seq data

---

<sup>6</sup>library size is important but was not covered here.

# Finally

RNA-seq (2)

Peter N.  
Robinson

RNA-seq

RPKM

Fisher's exact  
test

Poisson

LRT

Negative  
Binomial

- Email: peter.robinson@charite.de
- Office hours by appointment

## Further reading

- Marioni JC et al (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**:1509–17.
- Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol.* **11**:R106.
- Auer PL, Doerge RW (2010) Statistical design and analysis of RNA sequencing data. *Genetics* **185**:405-416.
- Z. Wang et al. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**:57-63.