



Leipziger Forschungszentrum für  
Zivilisationserkrankungen

UNIVERSITÄT LEIPZIG

Medizinische Fakultät

# Expression Quantification (I)

Mario Fasold, LIFE, IZBI





Leipziger Forschungszentrum für  
Zivilisationserkrankungen

## Sequencing Technology

One Illumina HiSeq 2000 run  
produces 2 times (paired-end)

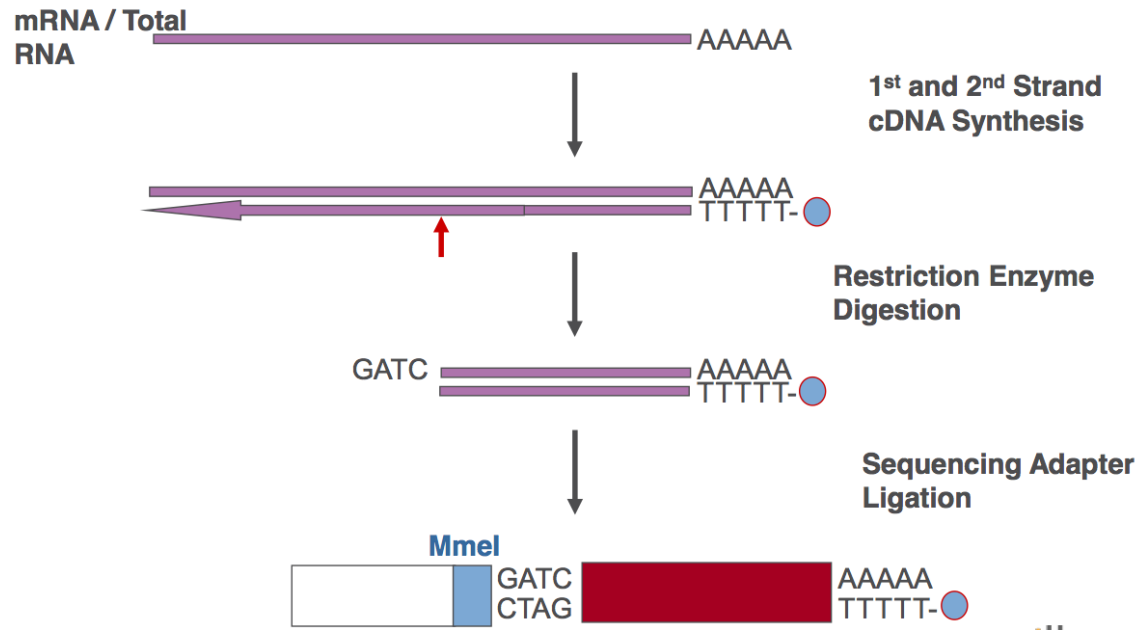
- ca. 1,2 Billion reads
- ca. 120 GB FASTQ file



©2010, Illumina Inc. All rights reserved.



## RNA-seq protocol





Leipziger Forschungszentrum für  
Zivilisationserkrankungen

UNIVERSITÄT LEIPZIG

Medizinische Fakultät

## Task

- Obtain some “value” (estimate) representing the true abundance of transcripts in vivo



Europa fördert Sachsen.  
**ESF**  
Europäischer Sozialfonds





## Overview

1. Basic Expression Measures
2. Statistical Poisson Model
  1. Single Isoform
  2. Multiple Isoform
  3. Statistical Inference
  4. Results
3. Paired-End Sequencing
4. Negative-Binomial





# Expression Quantification

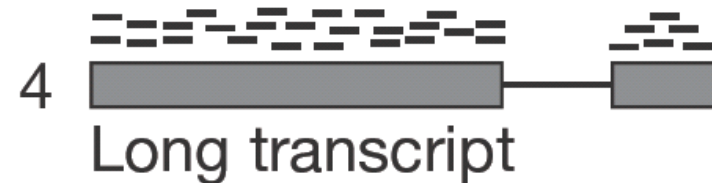
- What properties should the expression values have?
- Values should be comparable
  - Between transcripts
  - Between samples (-> differential expression)





# Problems in Expression Quantification

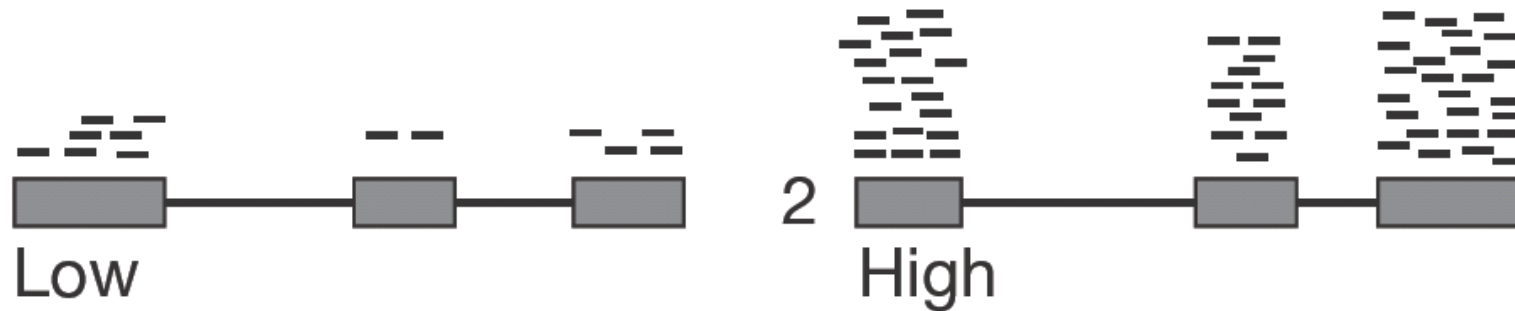
Differences in transcript length





# Problems in Expression Quantification

Fluctuations in sequencing depth



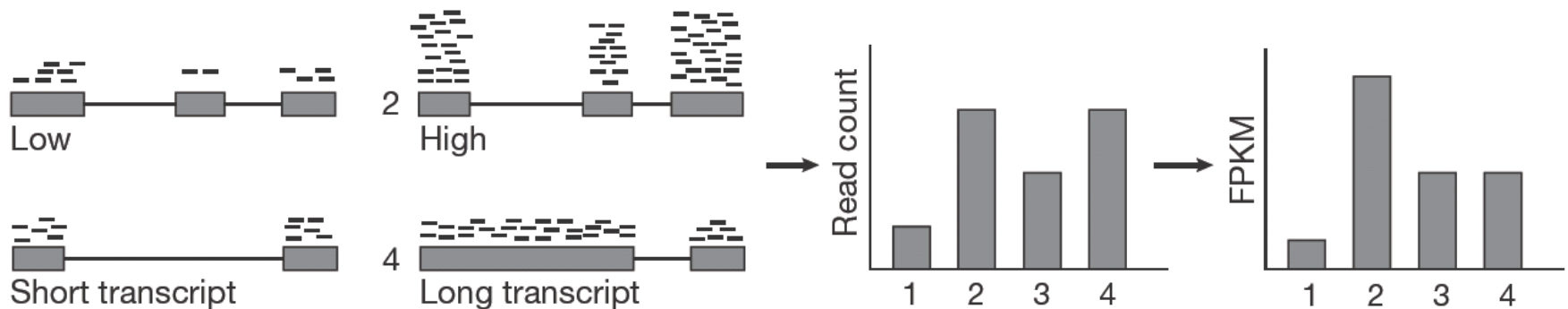


## Basic Measure: RPKM

RPKM/FPKM = Reads/Fragments per kilobase of exon model per million mapped reads

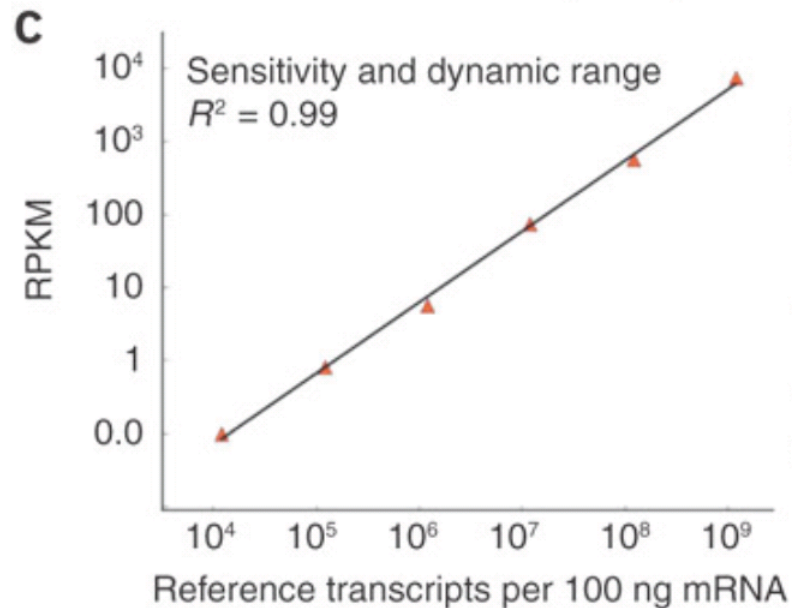
$$FPKM = 10^9 \times \frac{x}{wl}$$

$x$  = the number of reads mapped onto the gene's exons  
 $w$  = total number of reads in the experiment  
 $l$  = the sum of the exons in base pairs.



## RNA spikes:

- 300 and 1500nt (arabidopsis) and 10000nt ( $\lambda$ -phage)
- $10^4$ ,  $10^5$ , ...,  $10^9$  transcripts per 100ng mRNA



However, there are isoforms...

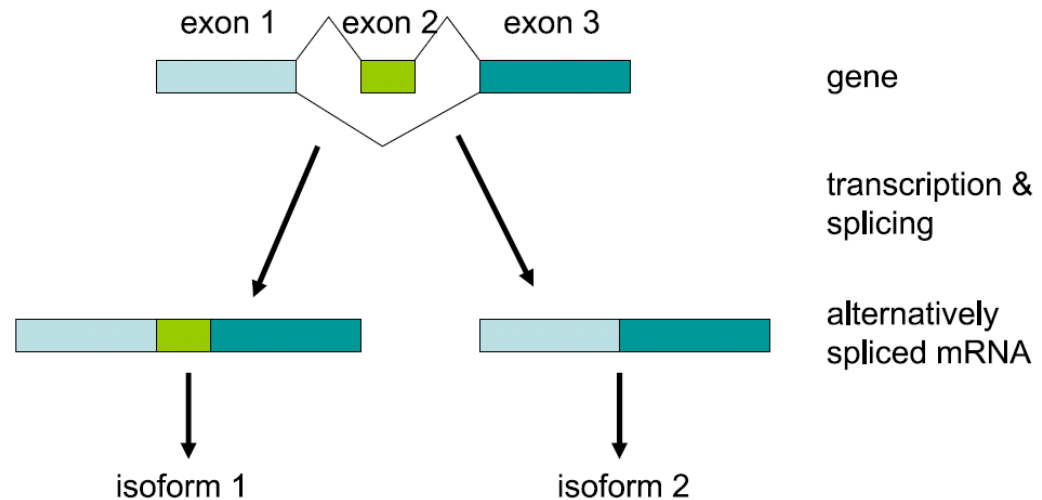
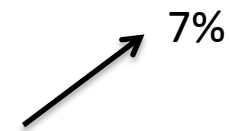


FIG. 1. A gene (DNA sequence) with three exons. During transcription, two isoforms are generated. The first isoform contains all of the gene's three exons. The second isoform contains the first and third exon, skipping the middle exon. This process is called alternative splicing and the middle exon is called an alternatively spliced exon.



# Typical number of isoforms

Among all the 19 069 RefSeq genes in the database, 1510 genes have more than one isoform. For these genes, the average number of isoforms is 2.39 and the maximum number of isoforms is 12. 7%



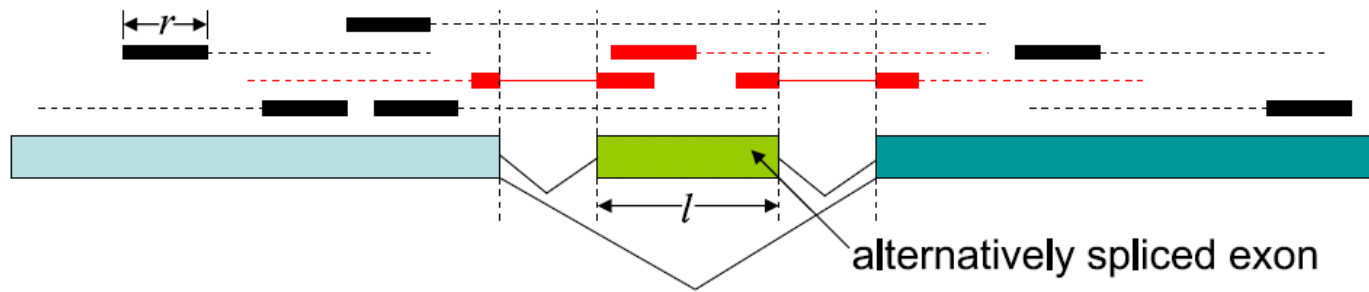
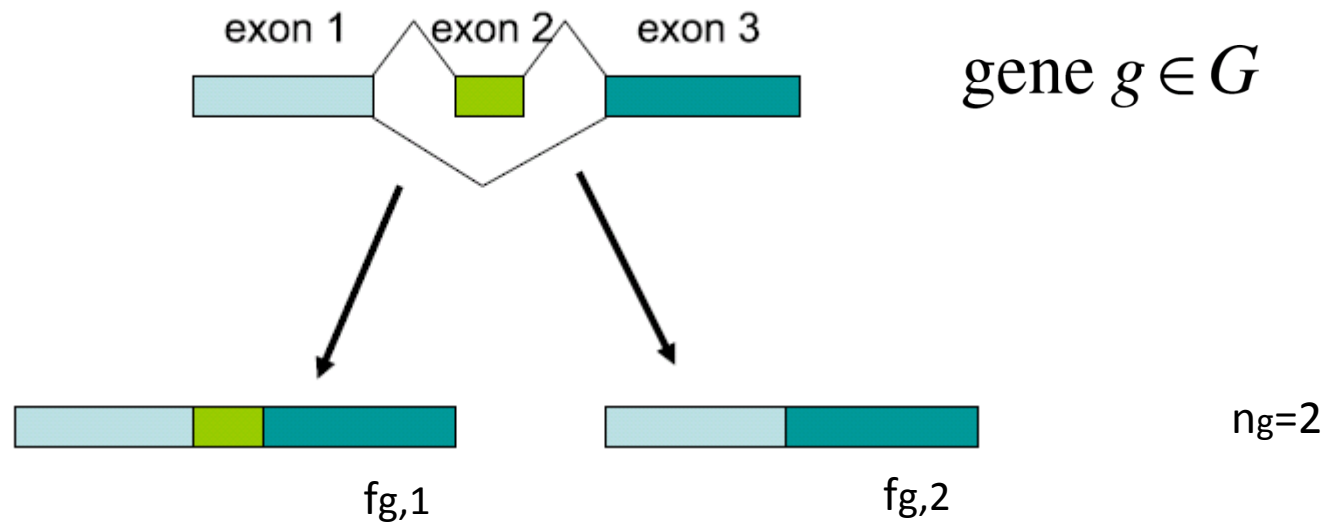
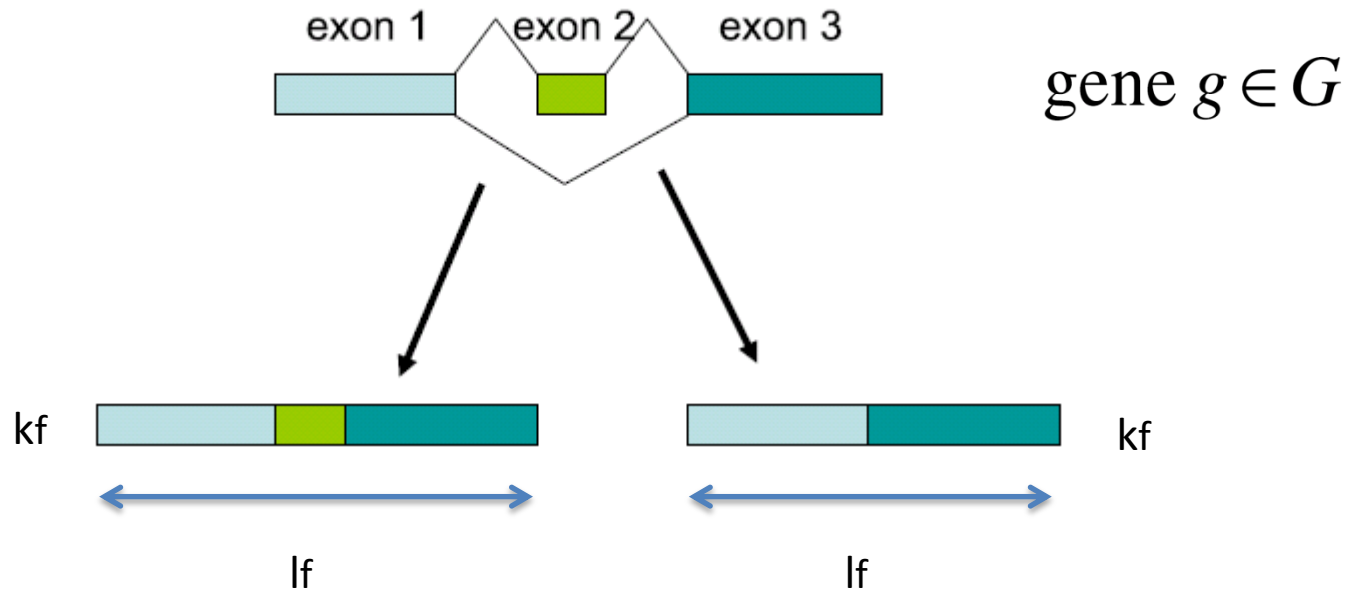


FIG. 2. *Single end sequencing. A gene of three exons is shown with the middle exon of length  $l$  being alternatively spliced. Reads that come from this gene are shown above the gene in solid bars and the parts that are not sequenced are shown in broken lines. Reads that span an exon–exon junction are shown in solid bars connected by thin lines. Reads that are related to the AS exon are shown in red color. In this case only the reads in red are isoform informative.*

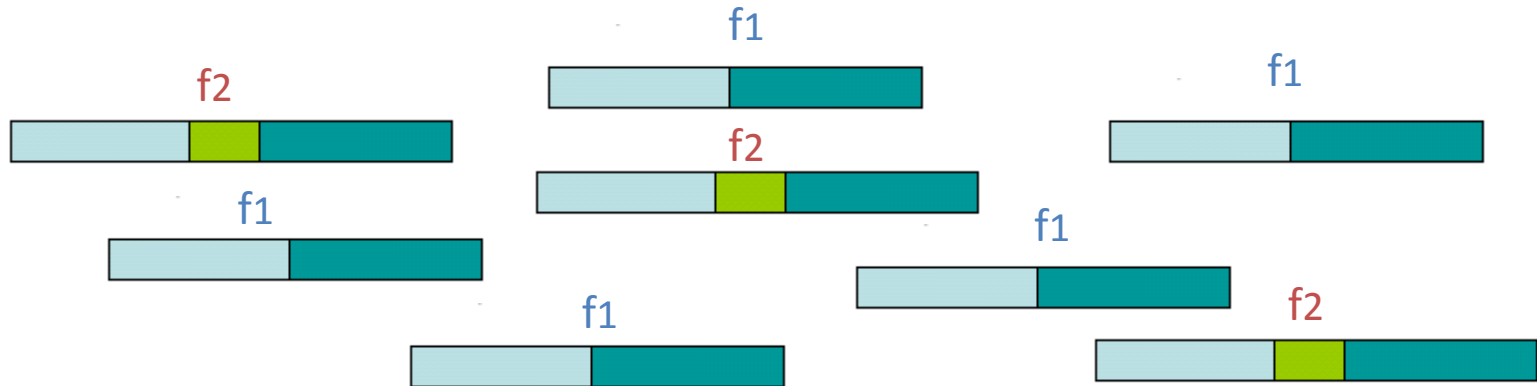
## Notation



let  $F_g = \{f_{g,i} | i \in [1, n_g]\}$  be the set of its isoforms



any isoform  $f \in F$ , let  $l_f$  be its length, and let  $k_f$  be the number of copies of transcripts in the form of isoform  $f$  in the sample.



Set of all isoforms in sample

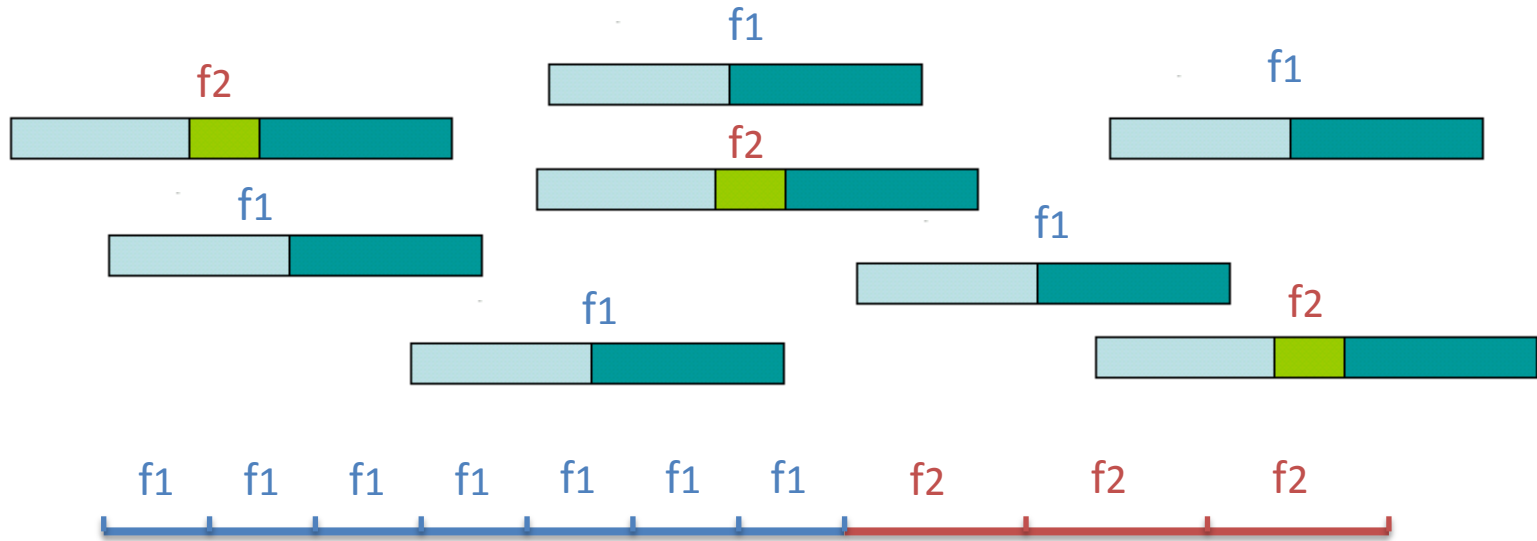
$$F = \{f_{g,i} \mid g \in G, i \in [1, n_g]\}$$

Total length of isoforms

$$\sum_{f \in F} k_f l_f$$







Expression value

$$\theta_f = k_f / \sum_{f \in F} k_f l_f$$

Probability of a read coming  
from isoform f

$$p_f = k_f l_f / \sum_{f \in F} k_f l_f$$

$$p_f = \theta_f l_f$$



# Model Assumptions

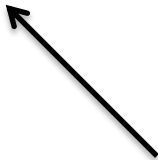
- Reads sampled independently (-> unique reads)
- Read sampling probability uniform over all sequences (uniform coverage)
- Each transcript processed independently and then sequenced



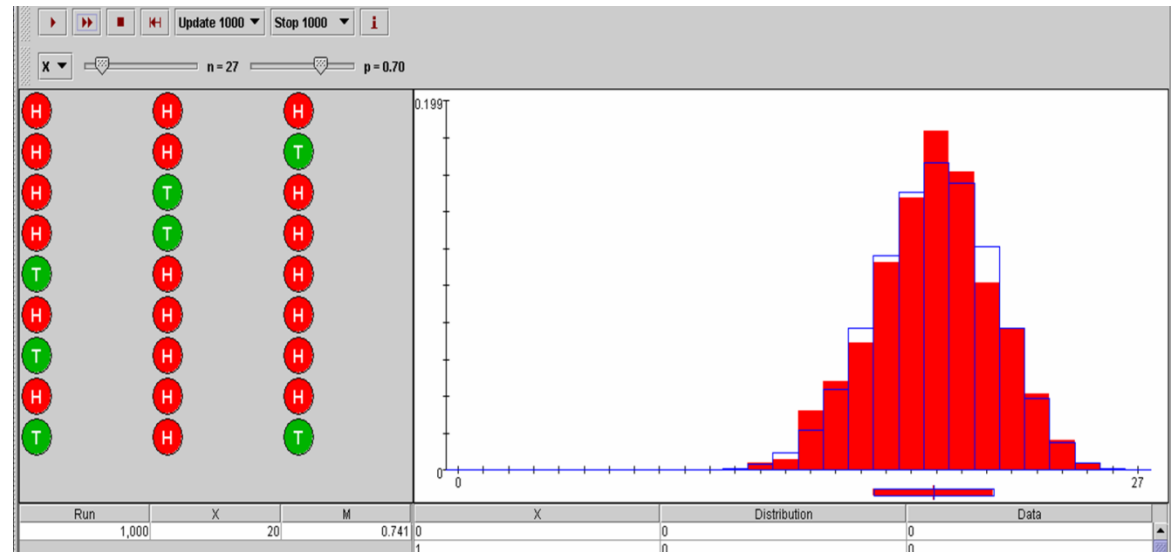


## Model (1)

- Let  $w$  be total number of reads
- Given isoform  $f$  and region of length  $l$  in  $f$
- Let  $X$  be a random variable representing the number of reads falling in that region
- Then  $X \sim \text{bin}(w, \theta_f l)$

$$p_f = \theta_f l_f$$


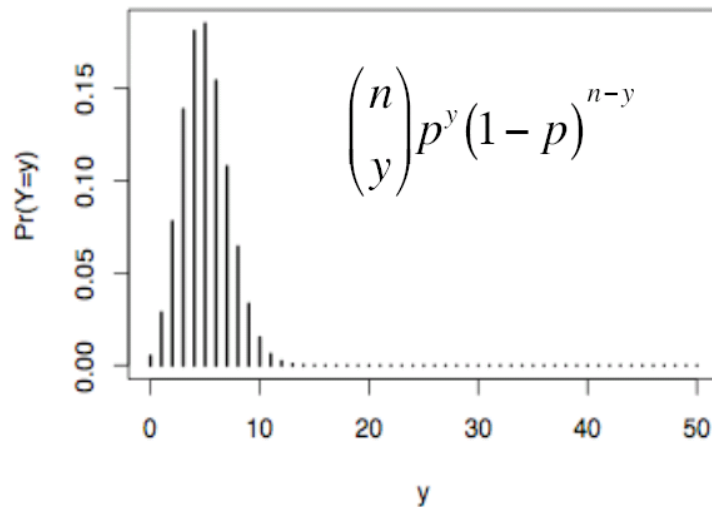
## Coin Toss gives Binomial Dist'n



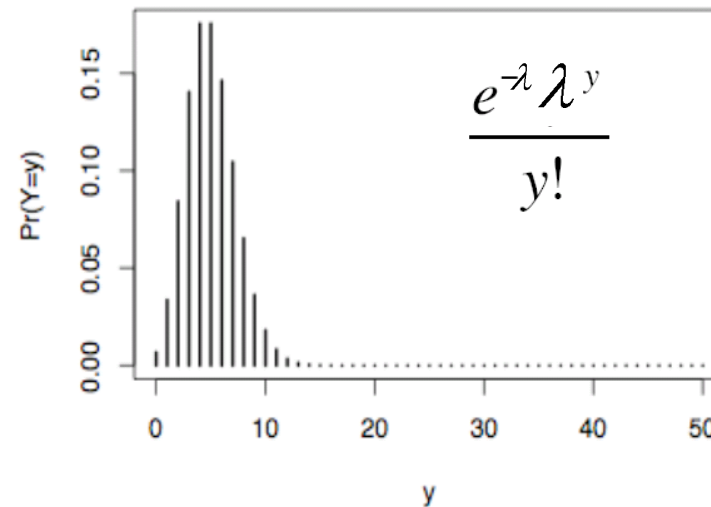
Recall Poisson is the limit of binomial as the number of 'trials' gets big but the probability of 'success' gets small:

$$\text{bin}(n, p) \rightarrow \text{Pois}(\lambda) \quad \text{as } n \rightarrow \infty, p \rightarrow 0, np = \lambda$$

bin(50,0.1)



Pois(5)





# Poisson Model

- Each gene  $g$  handled separately
- Exons only shared as a whole – isoforms either share an exon, or they don't
- Let  $X \sim \text{Poisson}(\lambda)$  be a random variable representing the number of reads falling in some region of interest in  $g$





# Poisson Model

- Each gene  $g$  handled separately
- Exons only shared as a whole – isoforms either share an exon, or they don't
- Let  $X \sim \text{Poisson}(\lambda)$  be a random variable representing the number of reads falling in some region of interest in  $g$
- Then for some exon  $j$ :  $\lambda = l_j w \sum_{i=1}^n c_{ij} \theta_i$
- $c_{ij}$  is 1 if isoform  $i$  contains exon  $j$ , 0 otherwise

$a_i$  is the “sampling rate”

In general,  $\lambda$  is a linear function of  $\theta_1, \theta_2, \dots, \theta_n$ , i.e.  $\lambda = \sum_{i=1}^n a_i \theta_i$ .



## Estimation of expression

From the probability mass function of the Poisson distribution, we have the likelihood function

$$\mathcal{L}(\Theta|x) = P(X=x|\Theta) = \frac{e^{-\lambda} \lambda^x}{x!}. \quad (1)$$

For the whole set of observations  $\mathbf{X} = \{X_s | s \in \mathcal{S}\}$ , if the corresponding regions do not overlap then  $X_s$ 's are independent and we can write the joint log-likelihood function as

$$\log(\mathcal{L}(\Theta|x_s, s \in \mathcal{S})) = \sum_{s \in \mathcal{S}} \log(\mathcal{L}(\Theta|x_s)) \quad (2)$$

The MLE is obtained by

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \log(\mathcal{L}(\Theta|x_s, s \in \mathcal{S}))$$



## Single isoform case

Taking logarithm on (1), we have

$$\begin{aligned}\log(\mathcal{L}(\Theta|x)) &= -\lambda + x \log \lambda - \log(x!) \\ &= -\sum_{i=1}^n a_i \theta_i + x \log \left( \sum_{i=1}^n a_i \theta_i \right) - \log(x!)\end{aligned}\quad (3)$$

Taking derivatives, we get

$$\frac{\partial \log(\mathcal{L}(\Theta|x))}{\partial \theta_j} = -a_j + \frac{x a_j}{\sum_{i=1}^n a_i \theta_i}\quad (4)$$

When  $n=1$ , so  $\Theta = \theta_1$  is a real number; it is well known that  $\hat{\Theta} = x/a$ . When  $x$  is the number of reads falling into some region of length  $l$ , we have  $a = wl$  and therefore  $\hat{\Theta} = x/wl$ , which is equivalent to the RPKM defined in Mortazavi *et al.* (2008).



## Multiple isoform case

- No closed-form solution available for ML if  $n > 1$
- Likelihood function is concave – any local maximum is also a global one
- Use numerical methods to find maximum
  - Here: coordinate-wise hill climbing



## Fisher information matrix

- We now have a point estimate for the expression index according to some Poisson model
- How reliable is the estimate?
- We need some confidence interval to test for DE
- The distribution of the estimate can be approximated with a normal distribution with mean (true parameter) and covariance (inverse Fisher information matrix)

$$\mathcal{J}_{jk}(\hat{\Theta}) = - \left. \frac{\partial^2 \log(\mathcal{L}(\Theta|x))}{\partial \theta_j \partial \theta_k} \right|_{\Theta = \hat{\Theta}} = \frac{x a_j a_k}{\left( \sum_{i=1}^n a_i \hat{\theta}_i \right)^2}$$

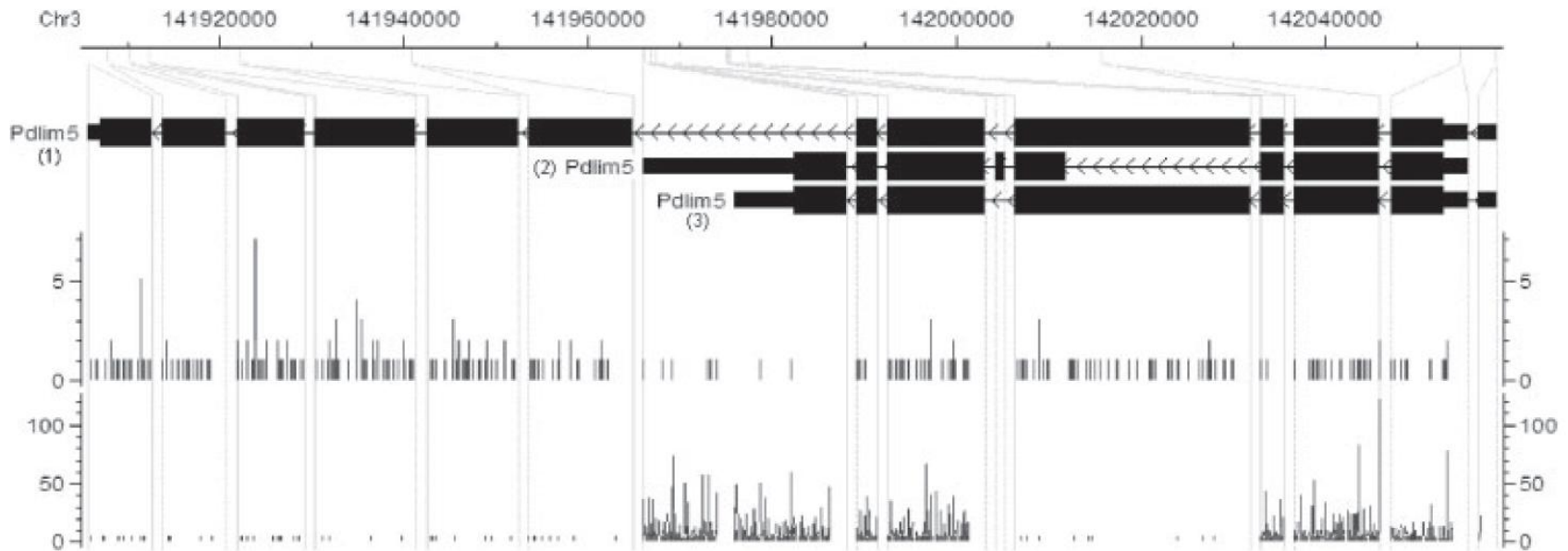


## Results

- Use RNA-seq dataset from Mortavazi et al.(2008): three mouse tissue samples with 2 replicates
- 60-80 M reads each
- Mapped with SeqMap tool from the same authors to mm9
- mRNA annotations from mm9 RefSeq mRNA database

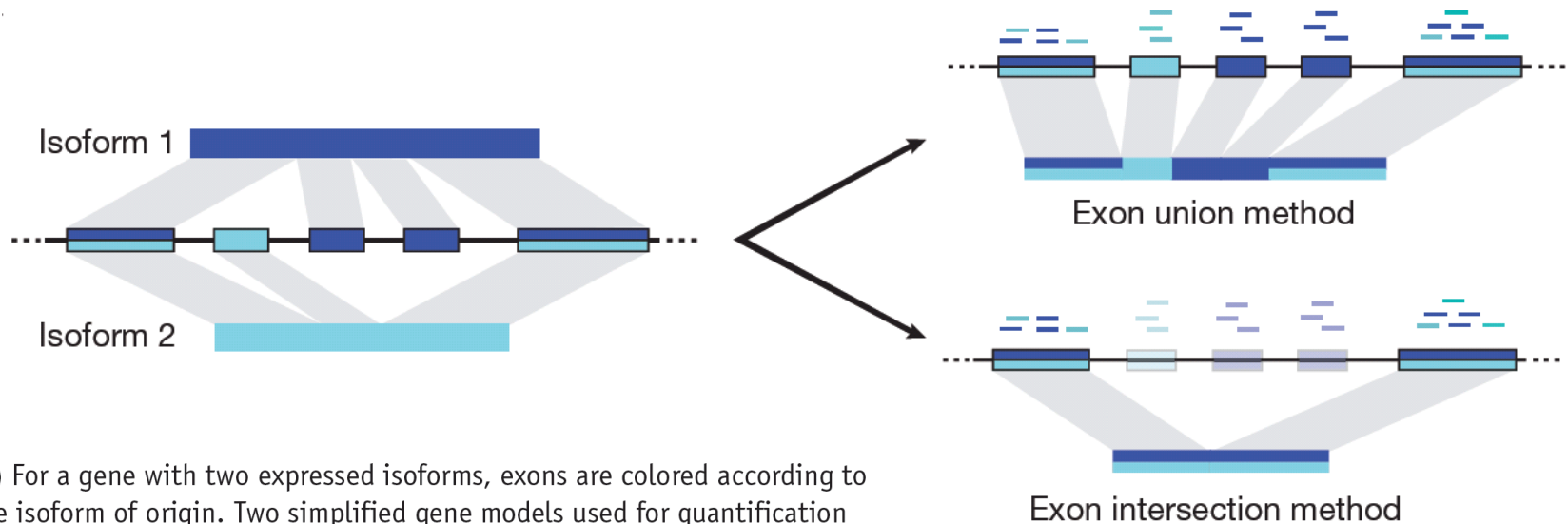


## Differentially expressed isoforms



A gene (Pdlim5) whose isoforms are differentially expressed is shown in Figure 2a. In brain samples, the estimated expressions for the three isoforms (from top to bottom) are 5.05, 0.42 and 0, respectively. In muscle samples they are expressed at 1.91, 238.67 and 14.89, respectively. As we can see, the first isoform is actually downregulated in muscle, although in terms of gene-level expression it shows upregulation.

## From isoforms to gene expression: which reads to count?



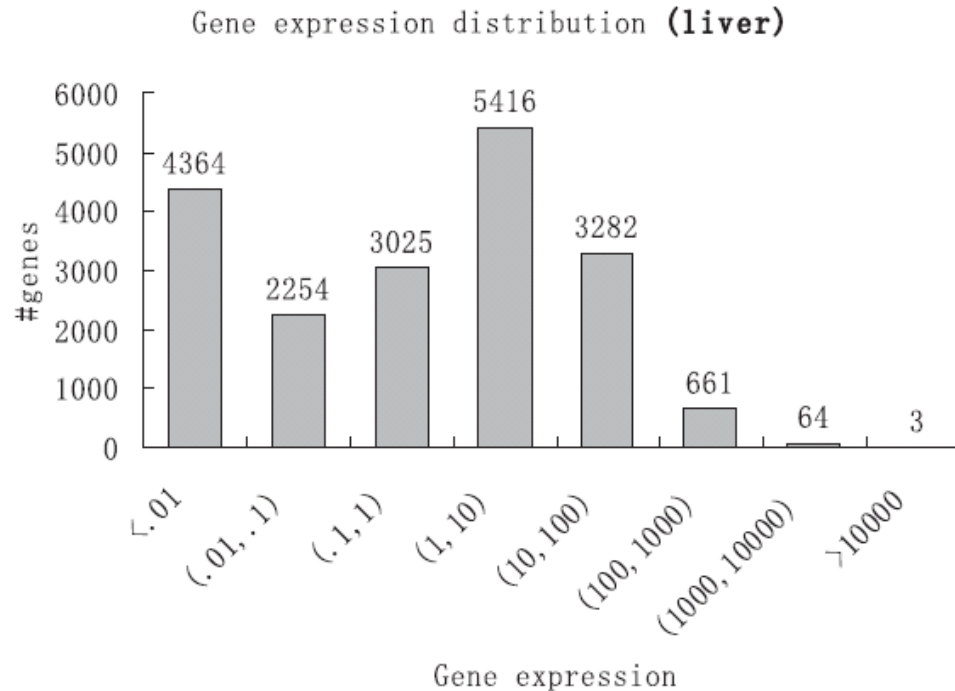
(c) For a gene with two expressed isoforms, exons are colored according to the isoform of origin. Two simplified gene models used for quantification purposes, spliced transcripts from each model and their associated lengths, are shown to the right. The 'exon union model' (top) uses exons from all isoforms. The 'exon intersection model' (bottom) uses only exons common to all gene isoforms. (d) Comparison of true versus estimated FPKM values in



## From isoforms to gene expression

- Exon intersection similar to microarrays, but they can reduce statistical power as reads are not considered
- Exon union underestimated expression for alternatively spliced genes
- This paper used “sum of all reads” = exon union





**Fig. 1.** Histogram of gene expressions in liver samples in the unit of RPKM. Genes are grouped into eight log-scaled bins according to their expressions. Genes are considered to be lowly (or highly) expressed if their RPKMs are below 1 (or above 100). Genes that have RPKMs between 1 and 100 are considered to be moderately expressed.

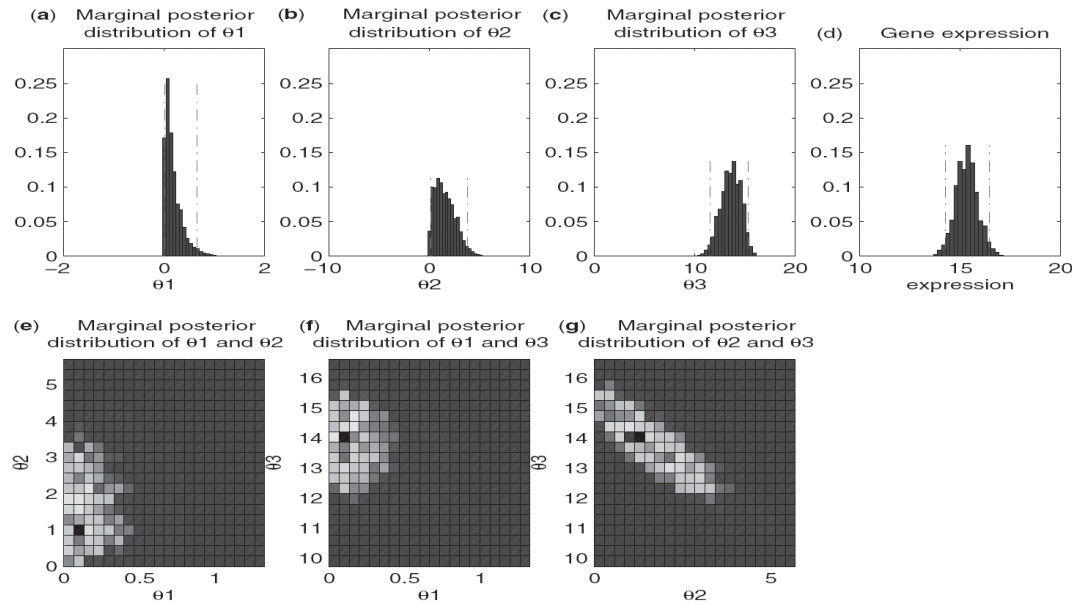
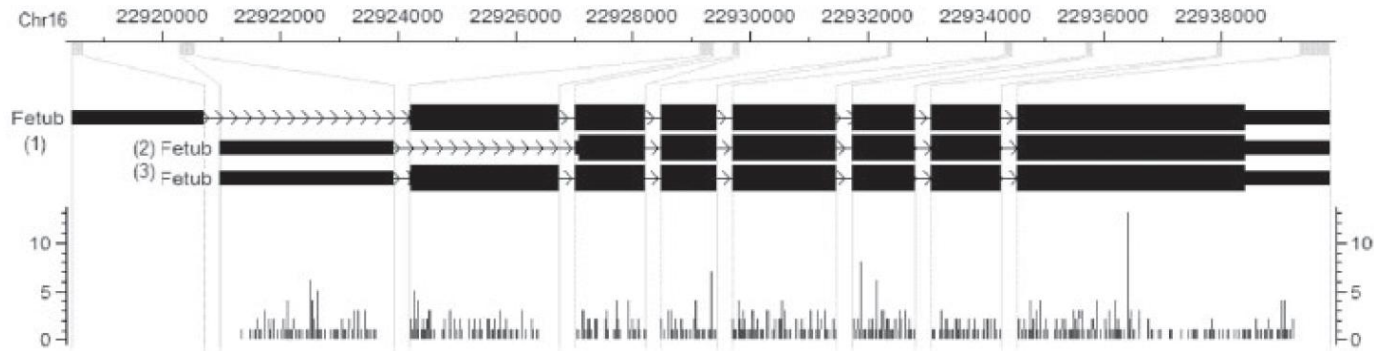




## Statistical inference

- Problem: Fisher information matrix is degenerated especially for isoforms with low expression
- Need to regularize covariance matrix
- Use Importance Sampling method to simulate from posterior distribution







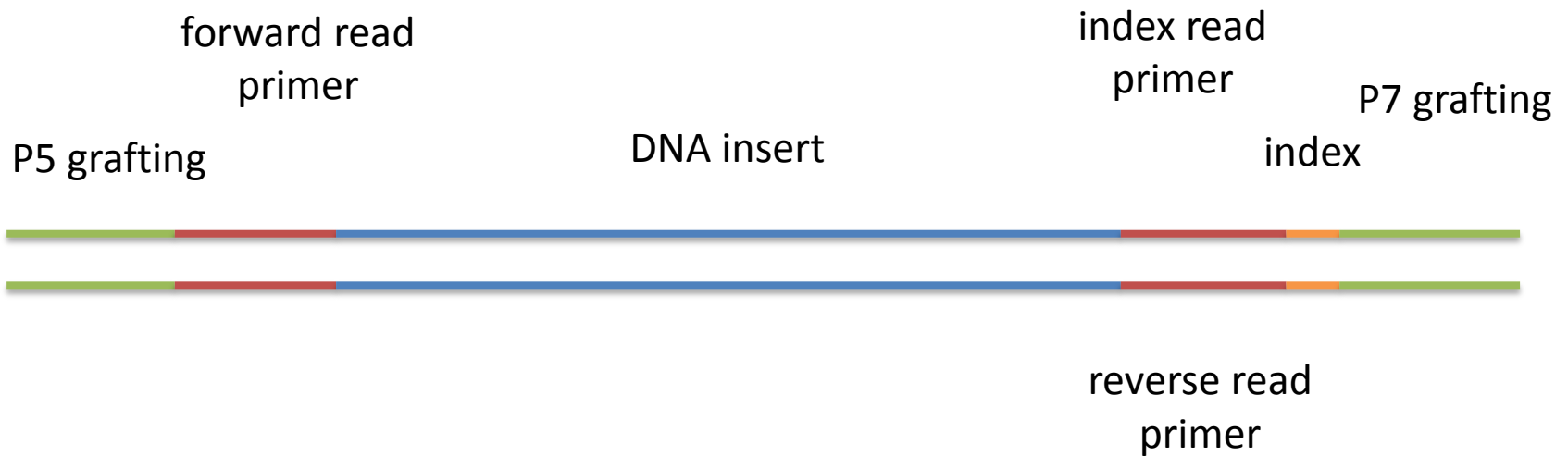
## Summary: isoform expression

- Isoform expression inference in RNA-seq possible using Poisson model
- “Inferences agree with detailed inspection”
- Exon junction reads reduce confidence interval

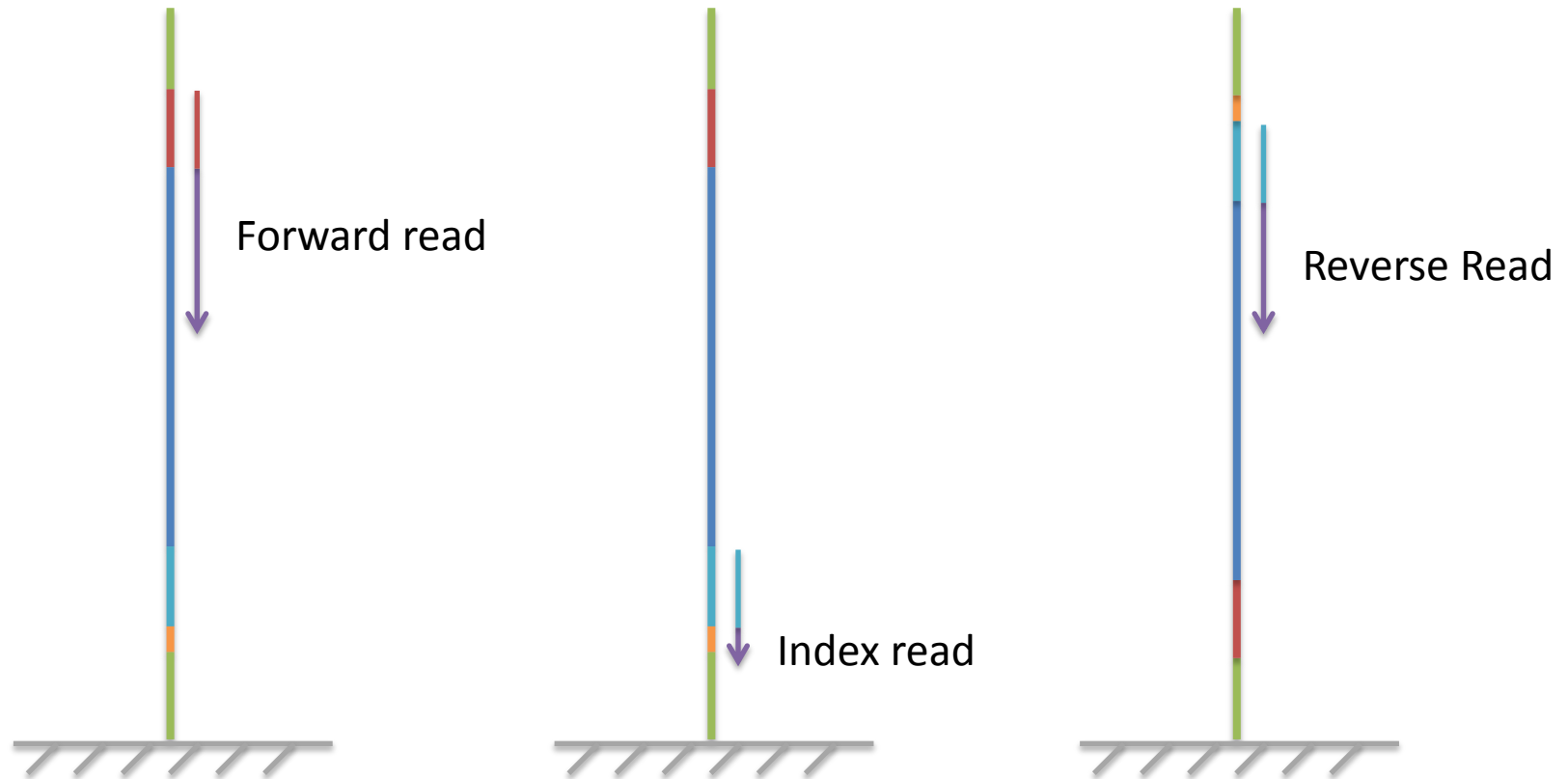




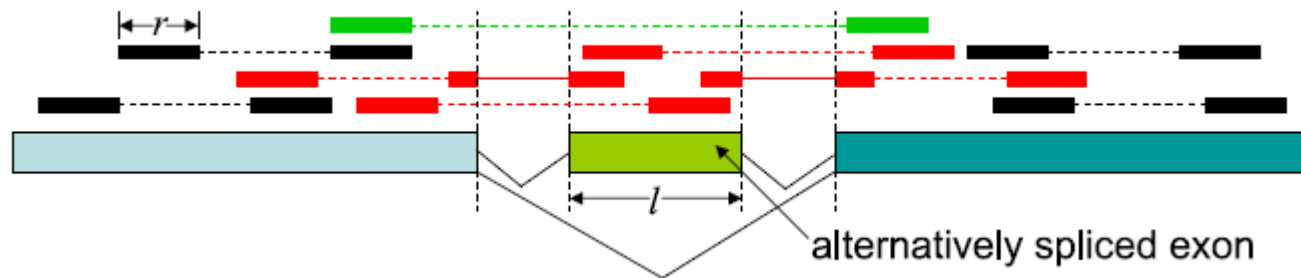
## Paired-end sequencing



## Paired-end sequencing



## Informative Reads



# Incorporation in the Poisson Model

- Model insert size using sampling rate  $a_i$

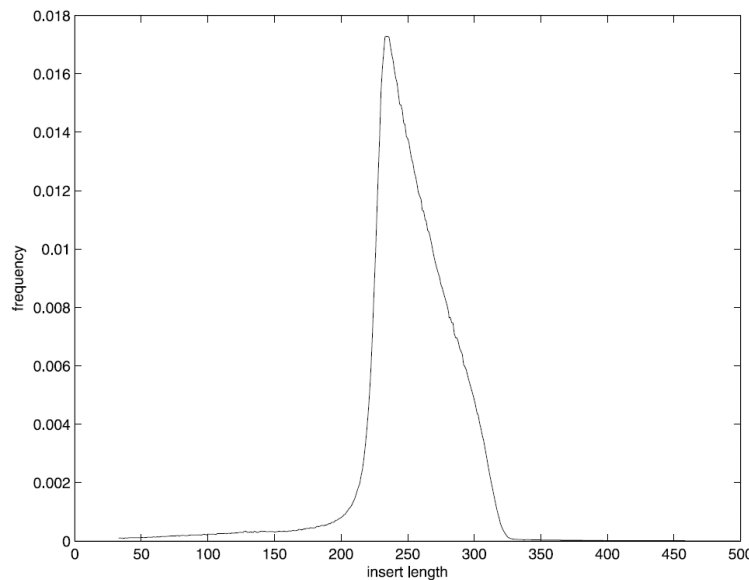


FIG. 5. A typical empirical mass function of the insert length.

$\lambda$  is a linear function of  $\theta_1, \theta_2, \dots, \theta_n$ , i.e.  $\lambda = \sum_{i=1}^n a_i \theta_i$ .



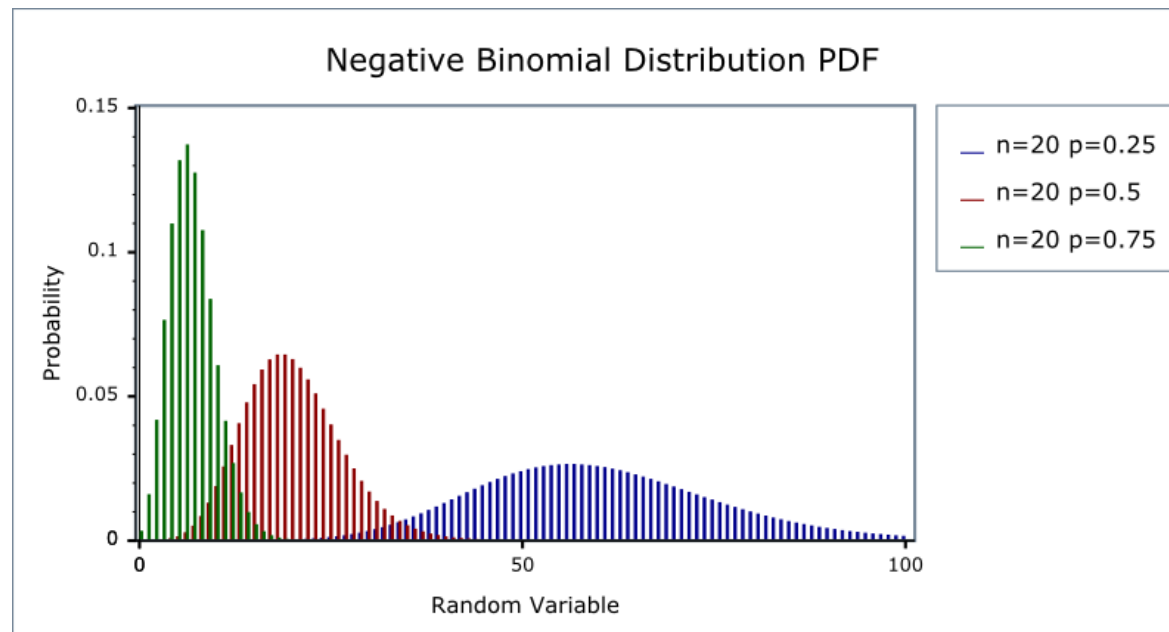
## Overdispersion and Negative Binomial

- Poisson distribution (one parameter) cannot account for (high) biological variability across samples
- Most RNA-seq studies have not enough replicates to estimate variability using a permutation-derived approach
- Model variance using e.g. negative binomial distribution



$$f(k; n, p) = \Pr(K = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$f(k) \equiv \Pr(X = k) = \binom{k + r - 1}{k} (1 - p)^r p^k \quad \text{for } k = 0, 1, 2, \dots$$





## Summary and Challenges

- Straightforward statistical model for isoform expression available
- However, reads are not truly random and uniform
  - High peaks and 3' bias in read distributions
  - Positive correlations in read distributions between tissues
- Isoform-level annotations not complete
- Most studies comprise few biological replicates

