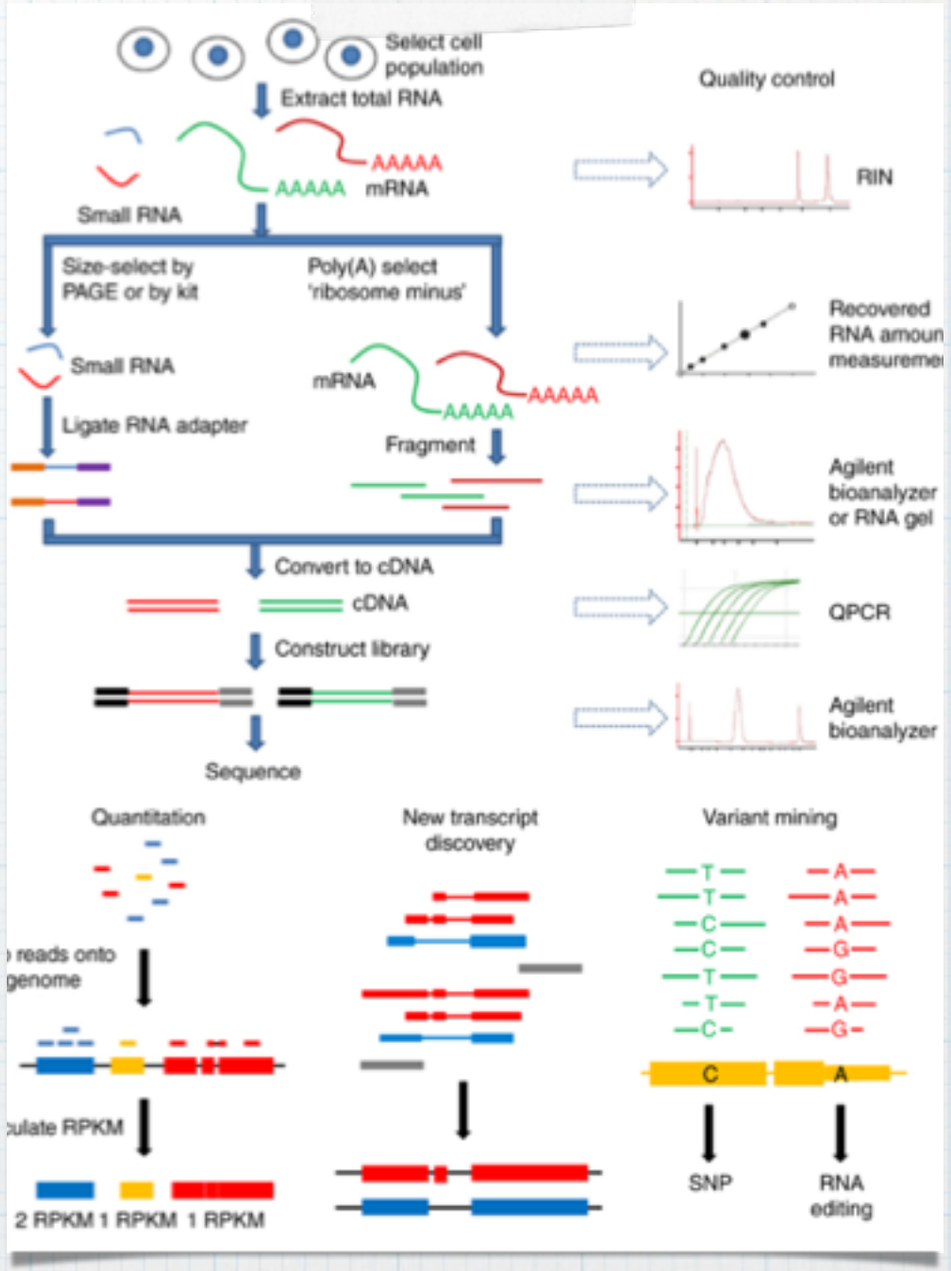
# RNA-Seq diff expression methods evolution

Journal Club. JetBrains BioLabs.
2014

Roman Cherniatchik

# RNA-Seq workflow

* Replicates
  * Biological (library)
  * Technical
* No Life cycle synch



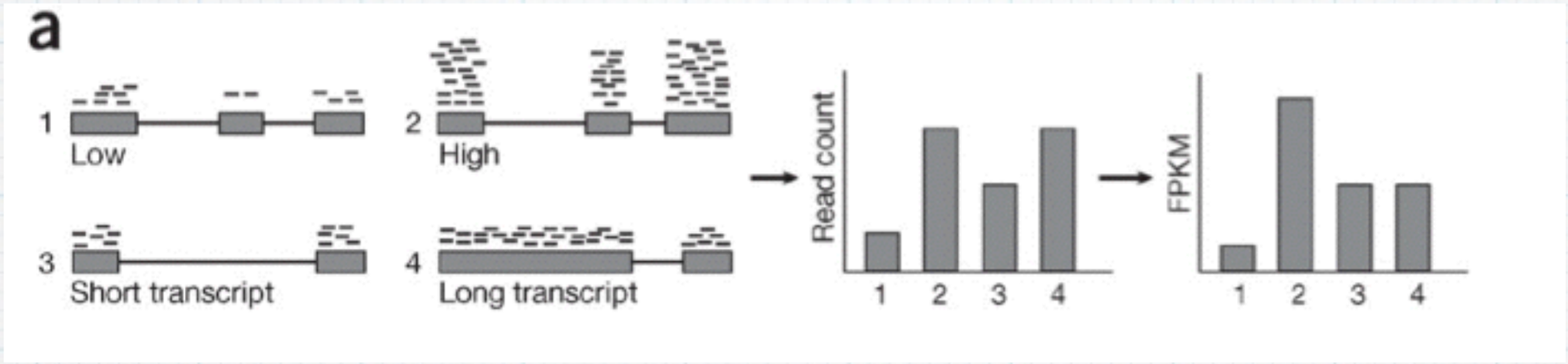
# Key moments

1. Gene / Isoform expression
2. Biological replicates variability

# Methods

* Length Normalization
  * TC (total count), UQ(upper quartile), Med (Median), Q(quantile normalization)
* RPKM/FPKM (reads per KB per 10^6 reads)

$$RPKM(X) = \frac{\text{Reads per transcript}}{\frac{\text{total reads}}{1{,}000{,}000} \cdot \frac{\text{transcript length}}{1000}}$$




* High expressed genes -> coverage impact
  * Quick fix: ignore high expressed genes, use housekeeping genes,...
* Low expressed and short genes bias

# Poisson

Each gene X $\sim$ Pois()

DE check:

* No replicates
    * t-test (expression diff = 0)
    * fisher exact 2x2 test  (most genes not DE, hyper geometric dist)

| | 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 | $n_{11} + n_{21}$ | $n_{12} + n_{22}$ | $N$ |

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

# Replicates

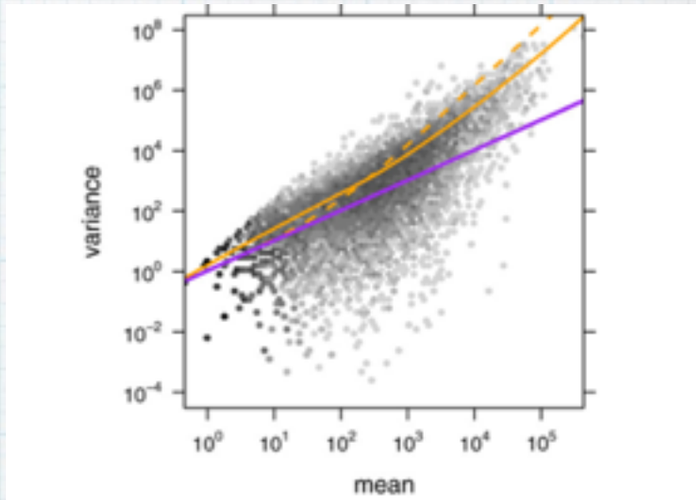
* LH ratio test (chi-sq dist)

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

* Xijk - #reads j-th gene, k-th lane, i-th sample
* X ~ Pois(lijk), lijk = Cik * V ijk
* Ho: Vijk = Vj (all samples)

# Poisson Model Problems

* Over dispersion
  * tech. replicates:0,5% genes show DE, RNA concentration bias
* Count uncertainty wrong
  * 50% reads - not unique alignment
  * pseudo genes, paralogs, alternative splicing



# NB model

* Over dispersion - OK
* A few replicates

EdgeR, DE-Seq, CuffDiff

P.S: Can be applied to ChipSeq data (e.g. DE-Seq)

# Edge R

* i-th gene, j-th sample
* Var(Kij) = E(Kij) + q * (E(Kij))^2
* q - global over dispersion in experiment
* E(Kij) = Lij * mj (Lij - proportion of gene j, mj - library size)
* Estimate q for all genes using few replicates

# DE-Seq

We assume that the number of reads in sample $j$ that are assigned to gene $i$ can be modeled by a negative binomial (NB) distribution,

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2), \tag{1}$$

which has two parameters, the mean $\mu_{ij}$ and the variance $\sigma_{ij}^2$. The read counts $K_{ij}$ are non-negative

First, the mean parameter $\mu_{ij}$, that is, the expectation value of the observed counts for gene $i$ in sample $j$, is the product of a condition-dependent per-gene value $q_{i,\rho(j)}$ (where $\rho(j)$ is the experimental condition of sample $j$) and a size factor $s_j$,

$$\mu_{ij} = q_{i,\rho(j)} s_j. \tag{2}$$

$q_{i,\rho(j)}$ is proportional to the expectation value of the true (but unknown) concentration of fragments from gene $i$ under condition $\rho(j)$. The size factor $s_j$ represents the coverage, or sampling depth, of library $j$, and we will use the term *common scale* for quantities, such as $q_{i,\rho(j)}$, that are adjusted for coverage by dividing by $s_j$.

Second, the variance $\sigma_{ij}^2$ is the sum of a *shot noise term* and a *raw variance term*,

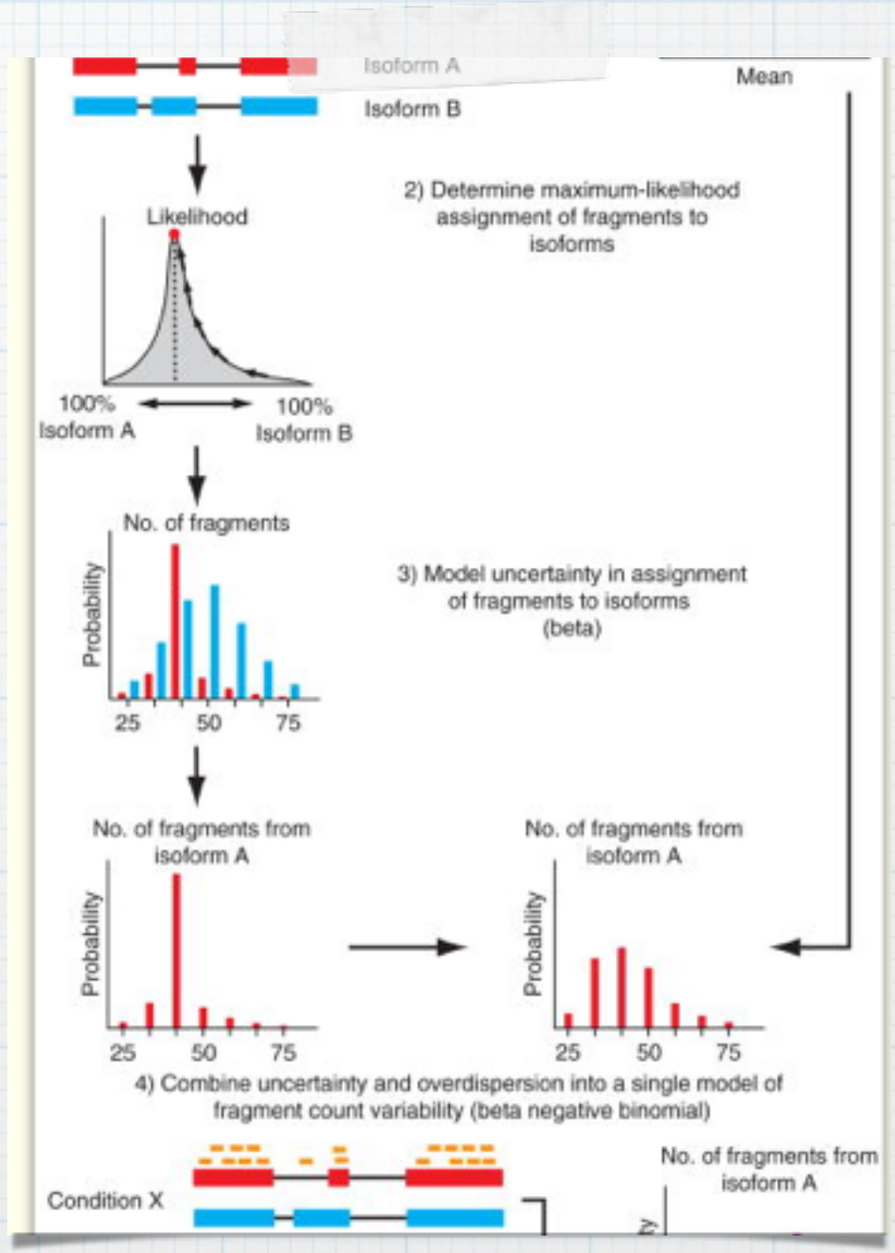
$$\sigma_{ij}^2 = \underbrace{\mu_{ij}}_{\text{shot noise}} + \underbrace{s_j^2 v_{i,\rho(j)}}_{\text{raw variance}}. \tag{3}$$

Third, we assume that the per-gene raw variance parameter $v_{i,\rho}$ is a smooth function of $q_i, \rho$,

$$v_{i,\rho(j)} = v_\rho(q_{i,\rho(j)}). \tag{4}$$

# Cuff links

* Beta (fragment to isoform) + NB (variance across replicates)
* 99% precision (genes)
* 95-99% (isoforms)



# Resources


* Robinson P. "RNA-seq Quantification and Differential Expression"
* Trapnell C., Hendrickson D., et all. "Differential analysis of gene regulation at transcript resolution with RNA-seq"
* Marioni JC, et all. "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays."
* Anders S., Humer W. "Differential expression analysis for sequence count data"
* Trapnel C. et all. "Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms"
* Kal AJ, "Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources."
* Robinson P. et all. "Small-sample estimation of negative binomial dispersion, with applications to SAGE data."