# Differential expression analysis using RNA-seq data

Joe Song

Department of Computer Science
New Mexico State University

Visiting
Department of Computer Science
Czech Technical University

April 15, 2019

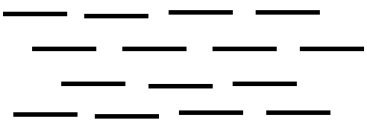


# RNA-sequencing

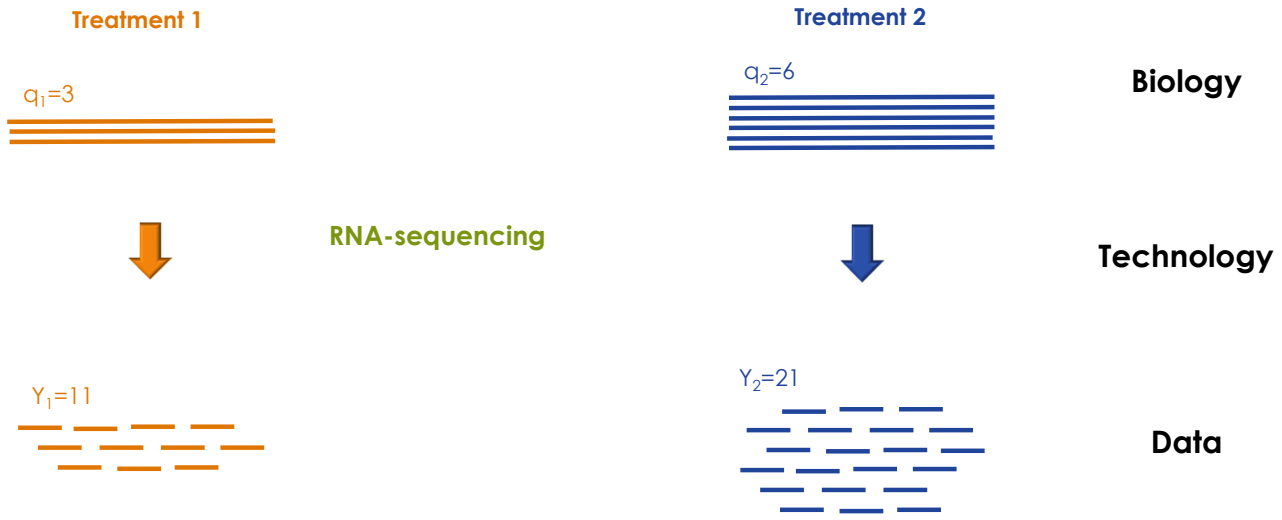
**q copies of a transcript in a sample**

**RNA-sequencing pipeline:**
**1 Library preparation from sample**
**2 Sequencing transcripts to get reads**
**3 Mapping reads to a reference genome**
**4 Quantify gene expression by genome annotation**

**Y reads**



# Differential analysis of read counts






# The differential expression question

- Are numbers of a transcript/gene different between two treatments?
- **Input**: only observed $Y_1$, $Y_2$ (but we did not observe $q_1$, $q_2$)
- **Output**: Is the transcript/gene differentially expressed $q_1 \neq q_2$?

How do we relate $Y_1$, $Y_2$ to $q_1$, $q_2$, all non-negative integers?

# Example: arrival of email

- The average number of e-mail arriving at Nancy's inbox in a given day is λ

  For example an average of 10 per day: λ = 10

- What is the probability of Nancy receiving *k* email on a given day?

- If email arrives independently of each other, *k* follows a Poisson distribution!



# Poisson distribution

**Definition:**

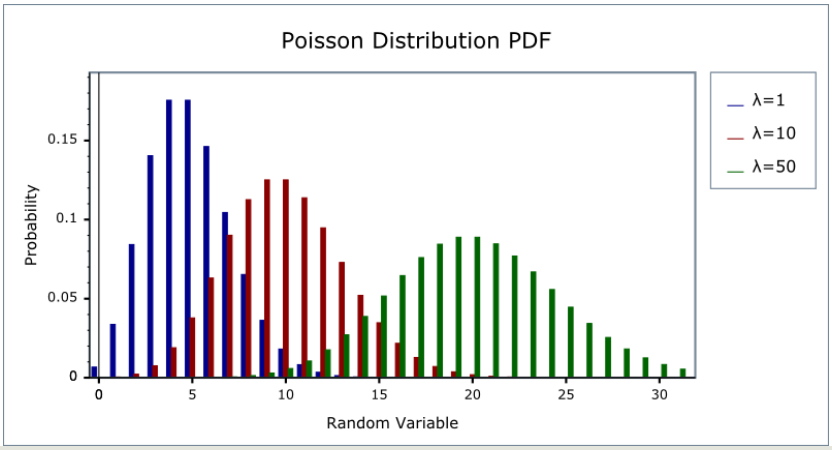
- Random variable Y: the number of observed events in a unit time interval.
- Parameter λ: the mean rate of events, or the mean number of events in a unit time interval
- Assumptions: the time between two events is independent of that of another two events

**Probability (mass) function:**

$$f(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

**Properties of Poisson distribution:**

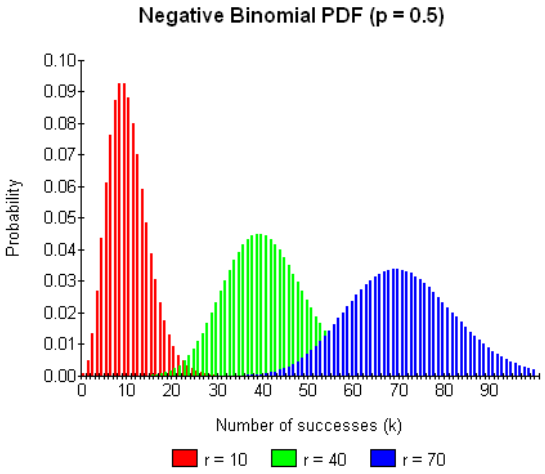
- mean: E(y) = λ
- variance: Var(y) = λ
- mean = variance



# Negative binomial (NB) distribution

- **Definition:** In a sequence of independent and identical Bernoulli trials with success probability *p*, we observe *Y=y* success trials before the *r*-th failure
- **Probability (mass) function NB(r, p):**

$$f(y;r,p) = \binom{y+r-1}{y} p^y (1-p)^r = (-1)^y \binom{-r}{y} p^y (1-p)^r$$




- **Properties:**

mean: $E(Y) = \frac{pr}{1-p}$

variance: $Var(Y) = \frac{pr}{(1-p)^2}$

mean ≤ variance



# Reparameterize negative binomial

Original NB(r,p) using r and p:

$$f(y;r,p) = \binom{y+r-1}{y} p^y (1-p)^r$$

By plugging into the definition with

$$r = a \qquad p = \frac{\mu}{a+\mu}$$

New form of NB(μ,a) using mean μ and dispersion $\alpha$:

$$f(y;\mu,\alpha) = \binom{y+\alpha-1}{y} \left(\frac{\mu}{\alpha+\mu}\right)^y \left(\frac{\alpha}{\alpha+\mu}\right)^\alpha$$

# Generalized linear model (GLM)

**Mathematical linear model**: Y = X β

- X: an *n* x *p* design matrix of *n* observed values from *p* predictors, independent variables
- β: a vector representing p unknown parameters
- Y: a vector of n observed values of a response, or dependent variable

**Statistical linear model (lm):**

- E(Y)=μ=Xβ
- Y is normally distributed
- Implicitly: g(z)=z is a linear link function

**Statistical Generalized linear model (glm):**

- E(Y)=μ=$\mathbf{g}^{-1}$(Xβ)
- Y has a distribution from the exponential family, including negative binomial
- **g**() is a link function, often nonlinear, such as **log**()



# GLM with a negative binomial distribution

**Differential expression: The gene read count is a generalized linear function of experimental conditions**

*i* – gene index  *j* – sample index  *r* – covariate (treatment) index

| Quantity | Definition |
|---|---|
| Observed read count $Y_{ij}$ of gene i sample j: | $Y_{ij} \sim \mathrm{NB}(\mathrm{mean} = \mu_{ij}, \mathrm{dispersion} = \alpha_i)$ |
| Mean read count proportional to the true read count $q_{ij}$: | $E(Y_{ij}) = \mu_{ij} = s_{ij} q_{ij}$ |
| Nonlinear (log) link function: | $\log \frac{E(Y_{ij})}{s_{ij}} = \log q_{ij} = \sum_r x_{jr} \beta_{ir}$ |
| Gene and sample specific factor: | $s_{ij}$ |
| Elements of the design matrix X – treatment r of sample j: | $x_{jr}$ |
| Logarithmic fold change for gene *i* contributed by covariate *r*: | $\beta_{ir}$ |

# The Pasilla gene RNA-seq experiment

- Pasilla (PS), the *Drosophila melanogaster* ortholog of mammalian NOVA1 and NOVA2
- Pasilla gene regulates alternative splicing of pre-mRNA
- Experiment: Pasilla is depleted (treated) and RNA-seq is measured
- Control: Wild type (untreated) RNA-seq is measured
- What genes are differentially expressed in response to Pasilla depletion?

**References**: Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S, Brenner SE, Graveley BR, "Conservation of an RNA regulatory map between Drosophila and mammals." Genome Res. 2011 Feb;21(2):193-202. https://doi.org/10.1101/gr.108662.110



# Experimental design specification

- specify experimental design
  - condition: treated, untreated
  - type: single-read, paired-end
  - design:
    - ~ condition  # only treatment effect
    - ~ condition + type  # treatment and technical factor
    - ~ condition + type + condition:type  # treatment and interaction factor

| Sample Name | condition | type | SampleFiles |
|---|---|---|---|
| 1 | treated | single-read | treated1fb.txt |
| 2 | treated | paired-end | treated2fb.txt |
| 3 | treated | paired-end | treated3fb.txt |
| 4 | untreated | single-read | untreated1fb.txt |
| 5 | untreated | single-read | untreated2fb.txt |
| 6 | untreated | paired-end | untreated3fb.txt |
| 7 | untreated | paired-end | untreated4fb.txt |

# The design formula

- Include as many covariates as possibly influencing results
- Identify factors biologically important for decision making
- Extract p-values & fold-changes contributed by those biological factors
- Generate volcano plot (y=-log(p-value) versus x=log fold-change)
- Choose differentially expressed genes whose adjusted p-values are less than 0.05 and log fold change is ≤-1 or ≥1



# References


- Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, **15**, 550. doi: 10.1186/s13059-014-0550-8.
- Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S, Brenner SE, Graveley BR, "Conservation of an RNA regulatory map between Drosophila and mammals." Genome Res. 2011 Feb;21(2):193-202. https://doi.org/10.1101/gr.108662.110