SevenBridges

RNA-seq analysis

Apr 2021

Dajana Panovic

dajana.panovic@sbgenomics.com

Transcriptomics

Recap

Topics that will be covered today

- Central dogma of molecular biology
- RNA-seq analysis:
  - RNA quantification
  - Differential expression

Processes

Central dogma of molecular biology:

- Transcription (DNA to RNA)
- mRNA maturation (splicing and polyadenylation)
- Translation (RNA to amino acids)

Rate(s) and genes involved are different for different cells

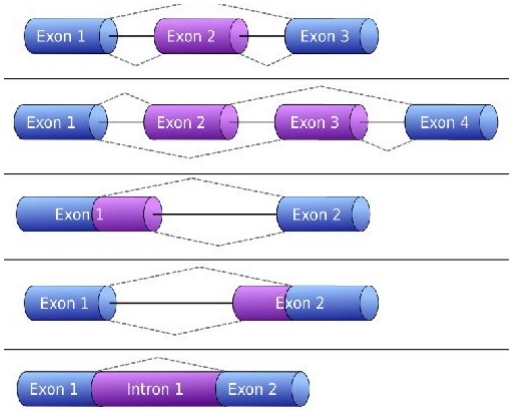
Central dogma of molecular biology

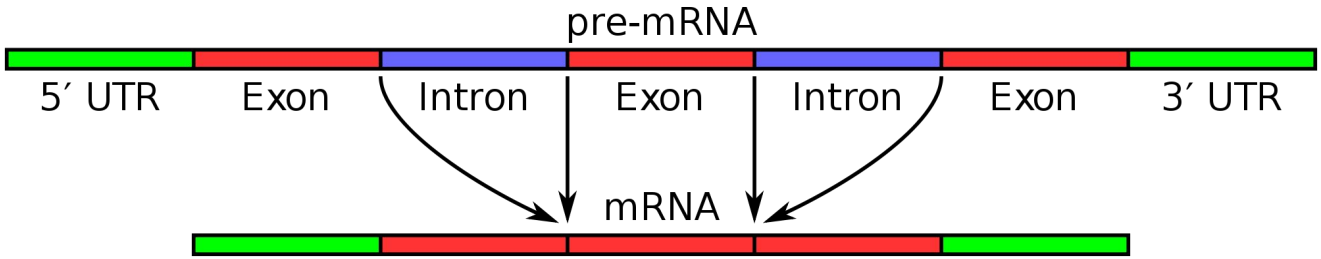
Following topics:

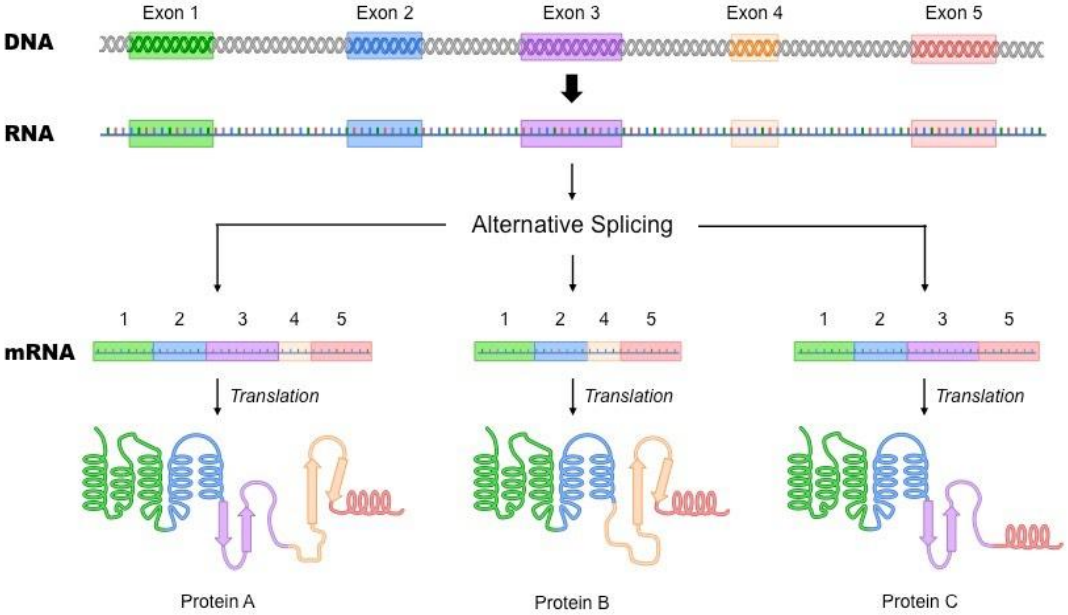
- **RNAs: transcription and translation, types of RNA**
- **mRNA: splicing, transcripts/isoforms**
- **RNA alignment**
- **RNA quantification**
- **Differential expression**

Terms

- **Transcripts**: All RNAs that are transcribed from DNA
- **mRNA**: Protein-coding transcripts
- **Isoforms**: Different transcripts from same gene







RNA-seq analysis

- RARELY: (splice-aware) alignment -> variant calling
- EVEN MORE RARELY: transcriptome assembly

RNA-seq analysis

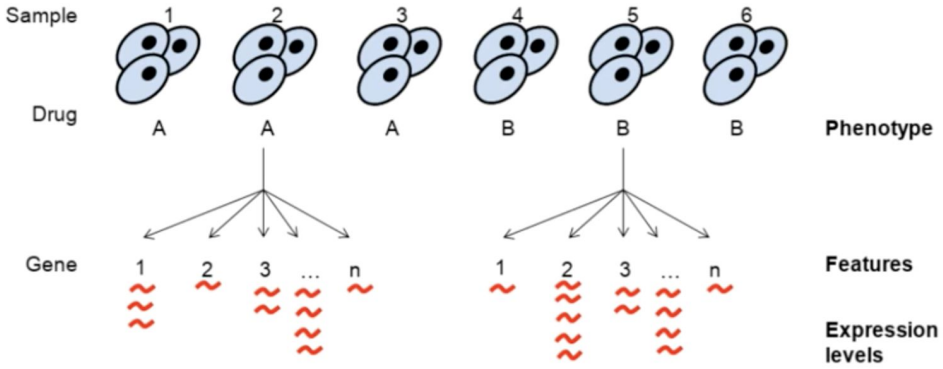
- OFTEN: **relative abundance (quantification)** of RNAs and testing for **differential expression**

New term:

- When genes give final products (proteins through transcription and translation) we say that gene is **expressed**

Why we analyze RNA

- All cells in the body have the same DNA
- However, set of RNA molecules between different cell types significantly differ



Motivation for RNA quantification

- We (usually) want to check if there is **change in transcription (expression)** between conditions (healthy/sick, treated/untreated, different tissues, etc..)

Transcriptomics

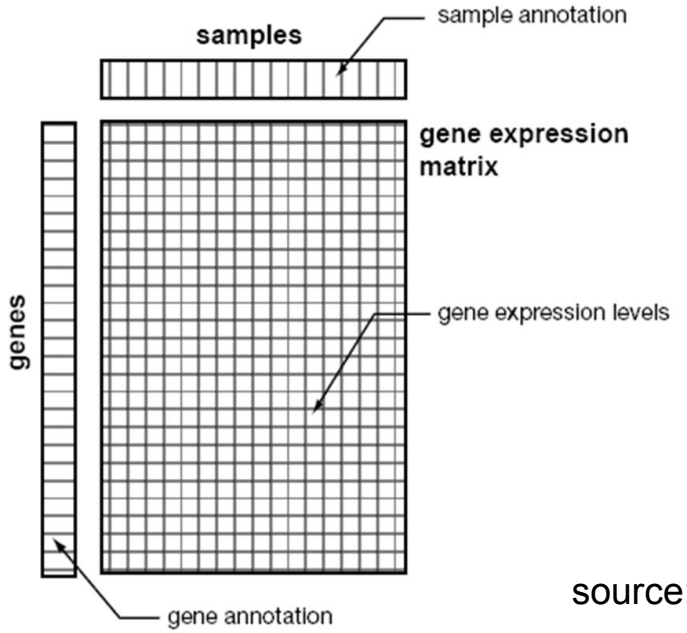
Quantification

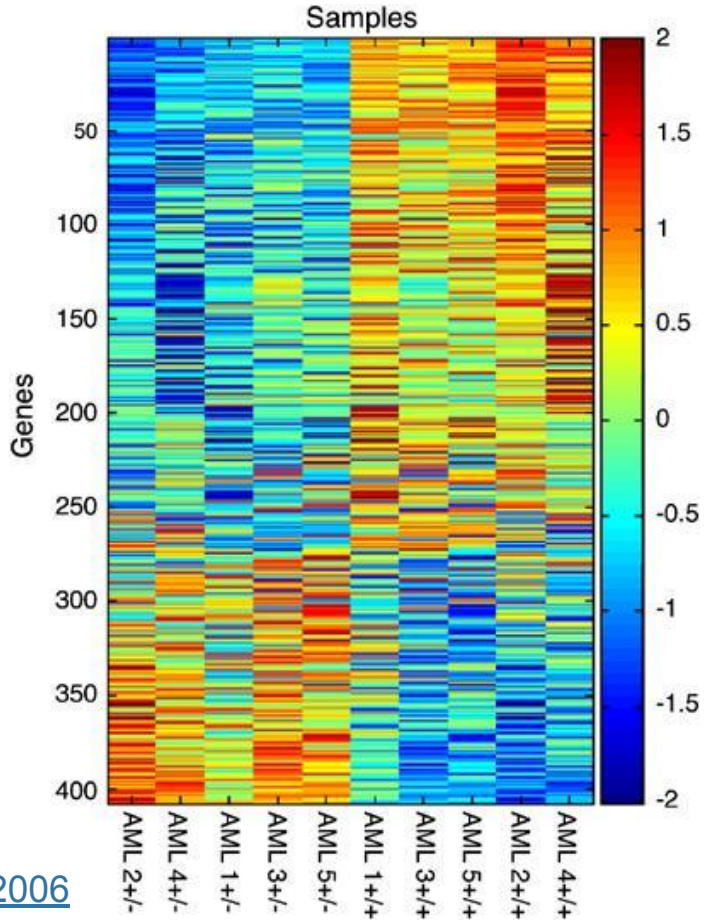
We will talk about:

- RNA quantification
- Differential expression

RNA quantification result

- Expression profiles





source: [Nature Leukemia 2006]

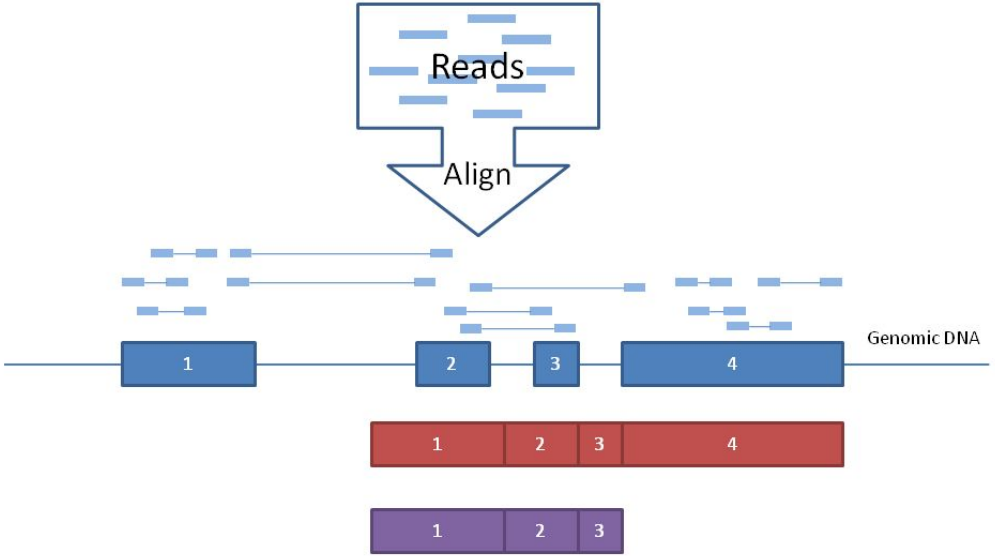
Quantification - problems

- Quantification = Counting reads?
- We can be interested in gene expression quantification, but also in transcript quantification

(1) RNA-seq: abundance estimation

Problem statement:

How to resolve alignment ambiguity?



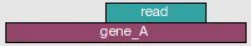
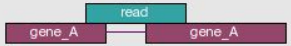
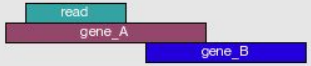
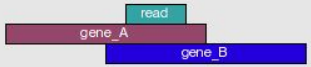
Source: http://dx.doi.org/10.13070/mm.en.3.203

(1) RNA-seq: abundance estimation

Raw counting

vs.

probabilistic estimation

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

HTSeq counting model

(2) RNA-seq: abundance estimation

- For transcript quantification we usually use different probabilistic methods

- E.g. Expectation Maximization algorithm (EML or EM), Maximum Likelihood estimation

(2) RNA-seq: abundance estimation

Maximum likelihood example

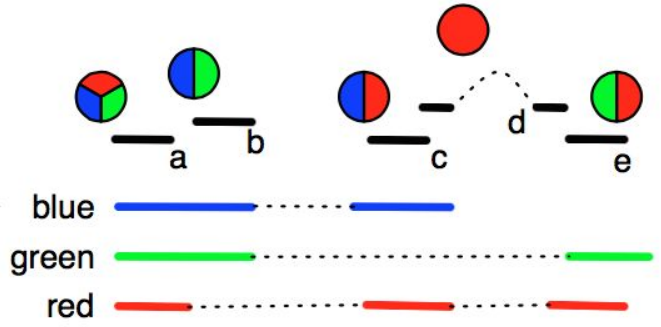
$i = 5$ single-end, equal-length reads (a,b,c,d,e)
$k = 3$ transcripts (blue, green, red)
$\rho = (\rho_{blue}, \rho_{green}, \rho_{red})$ relative abundances of transcripts
$\sum_k \rho_k = 1$, multinomial distribution

$P_i = \sum_k y_{i,k} \cdot \rho_k$, probability of detecting i-th read

where $y_{i,k} = 1$ if i-th read aligns to k-th transcript, otherwise 0

$$L(\rho) = \prod_i \sum_k y_{i,k} \cdot \rho_k$$

Analytical solution $\rho = (0.18, 0.18, 0.64)$



Adapted from: Lior Pachter 2011, arxiv: 1104.3889v2

(2) RNA-seq: abundance estimation

EM example

$(\rho_{blue}, \rho_{green}, \rho_{red}) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, uniform prior

E1 step: Proportional assignment

$p_a = (1/3, 1/3, 1/3),\ p_b = (1/2, 1/2, 0),$

$p_c = (1/2, 0, 1/2),\ p_d = (0, 0, 1),\ p_e = (0, 1/2, 1/2)$

M1 step: recalculate abundances

$\rho_{blue} = (1/3 + 1/2 + 1/2 + 0 + 0)/5 = 0.27$

E2 step: prior $= (0.27, 0.27, 0.46)$

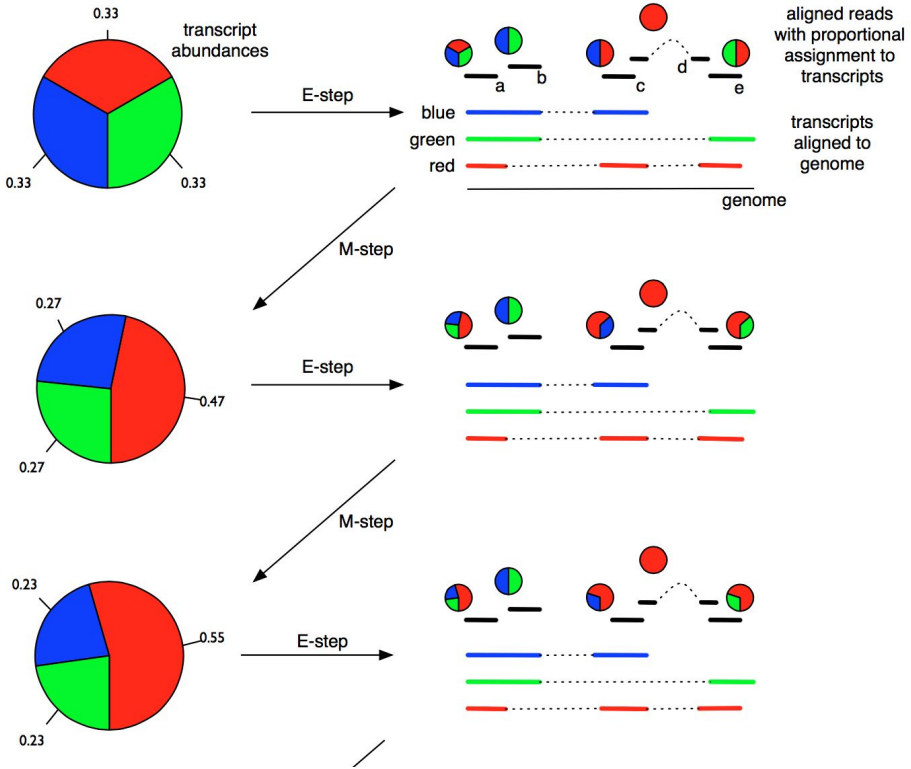
$p_a = (0.27, 0.27, 0.46),\ p_b = (1/2, 1/2, 0),$

$p_c = (\frac{0.27}{0.46 + 0.27}, 0, \frac{0.46}{0.46 + 0.27}),\ p_d = (0, 0, 1), ...$

M2 step:

$\rho_{blue} = (0.27 + 1/2 + 0.37 + 0 + 0)/5 = 0.23$

Iterative convergance $\rho_{blue} = 0.33, 0.27, 0.23, ..., 0.18$



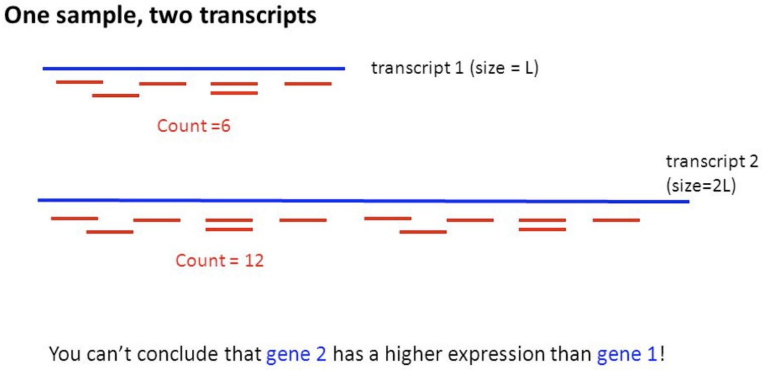
RNA-seq: data normalization

Problem statement:

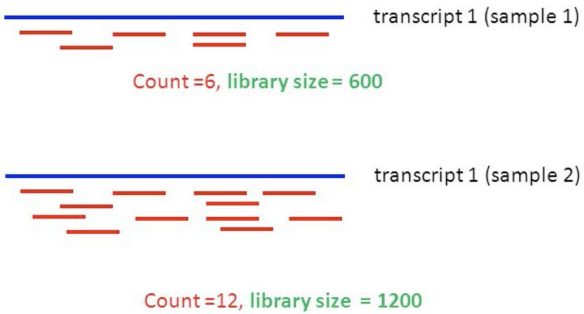
Can we compare expression of genes (within and between samples)

if we observe reads from sampled transcripts?

RNA-seq: data normalization



You can't conclude that gene 2 has a higher expression than gene 1!



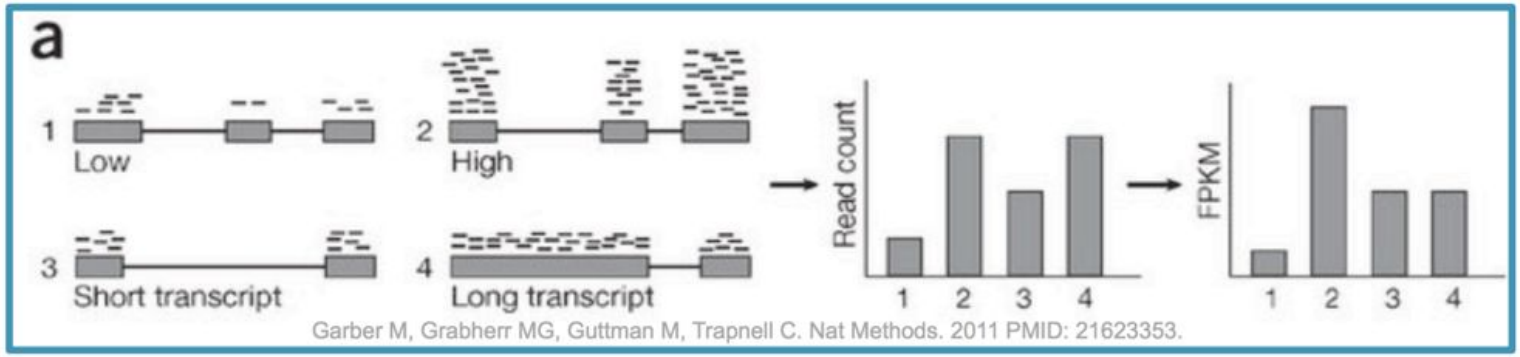
You can't conclude that gene 1 has a higher expression in sample 2 compared to sample 1!

- We need to account for gene length and library size

RNA-seq: data normalization

Let X_i be number of reads aligned to ith transcript

$$\sum_i X_i \neq \text{expression of a gene}$$



Garber M, Grabherr MG, Guttman M, Trapnell C. Nat Methods. 2011 PMID: 21623353.

(2) RNA-seq: data normalization

Relative units (adjust for transcript length and sequencing depth):

- Transcripts per million (TPM)
- Fragments per kilobase of exon per million reads (FPKM)

$$FPKM_i = \frac{X_i}{\frac{N}{10^6} \cdot \widetilde{l_i}}$$

$$TPM_i = \frac{\frac{X_i}{\widetilde{l_i}} \cdot 10^6}{\sum_i \frac{X_i}{\widetilde{l_i}}}$$

X_i - number of reads aligned to transcript ‘i’

N - total number of reads

l_i - read length

$\widetilde{l_i} = l_i/10^3$ - read length in kilobases

Transcriptomics

Differential expression

Differential expression:

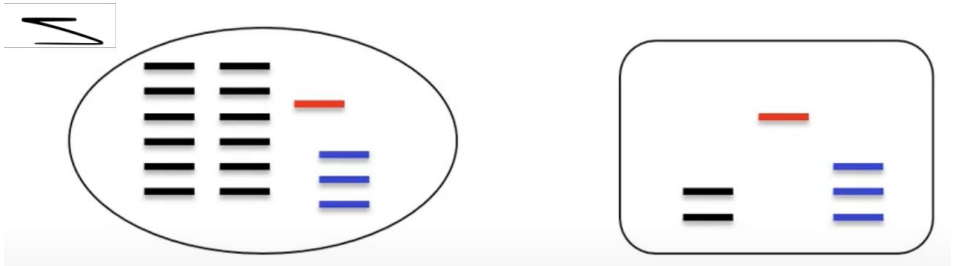
Problem statement:

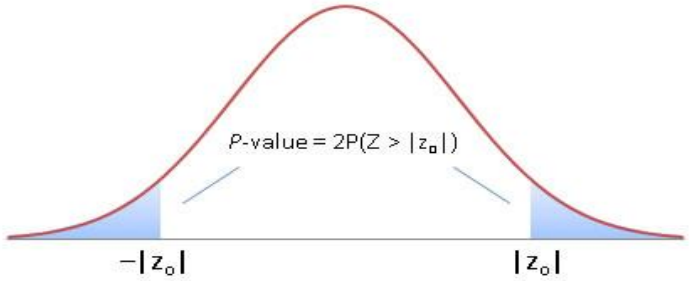
From thousands of genes, how do we know which ones are really differentially expressed and not observed changed by coincidence?

(3) RNA-seq: multiple testing

Measure of statistical significance

- **Null hypothesis**: there is no significant difference between specified populations, any observed difference being due to sampling or experimental error.
- The **p-value** is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true.
- The **alternative hypothesis** is considered true if the statistic observed would be an unlikely realization of the null hypothesis according to the p-value.





(3) RNA-seq: multiple testing

- In genomic studies you don't usually fit just one regression model or calculate just one p-value. You calculate many p-values.
- *human_hg19_genes_2015.gtf* has about 26,000 genes and 54,000 transcripts.
- Suppose 1200 out of 20,000 genes are found significant at 0.05 level.
    - No correction: you should expect 0.05 * 20,000 = 1000 false positives
    - Solution: Multiple testing correction

(3) RNA-seq: multiple testing

Multiple testing correction procedures:

- Bonferroni correction
  - p_value * total_number_of_tests_performed

For more info see also:

- BH (Benjamini-Hochberg) procedure
- BY (Benjamini–Yekutieli) procedure