# Differential Expression

Bioinformatics Applications (PLPTH813)

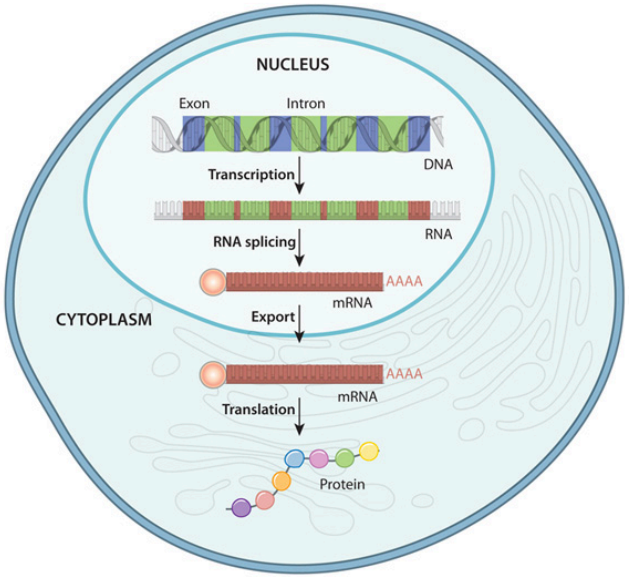
Sanzhen Liu

4/30/2019

# Outline

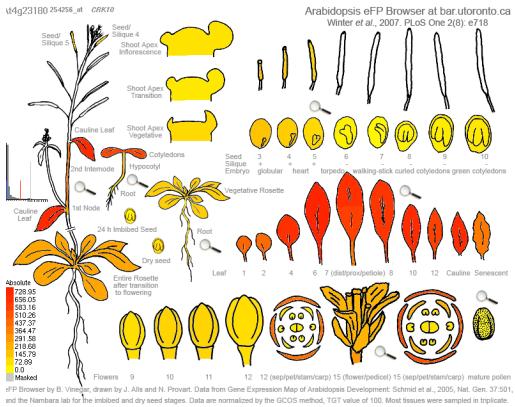
- Introduction of RNA-Seq
- RNA-Seq procedure
- Statistical test of differential expression
- Multiple testing correction

# Gene expression

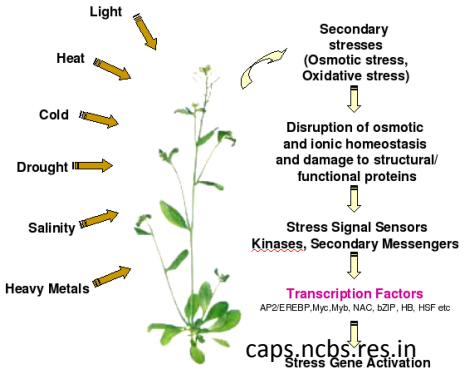



DNA to protein in eukaryote

nature.com/scitable/topicpage/gene-expression-14121669




Expression profiles in different tissues

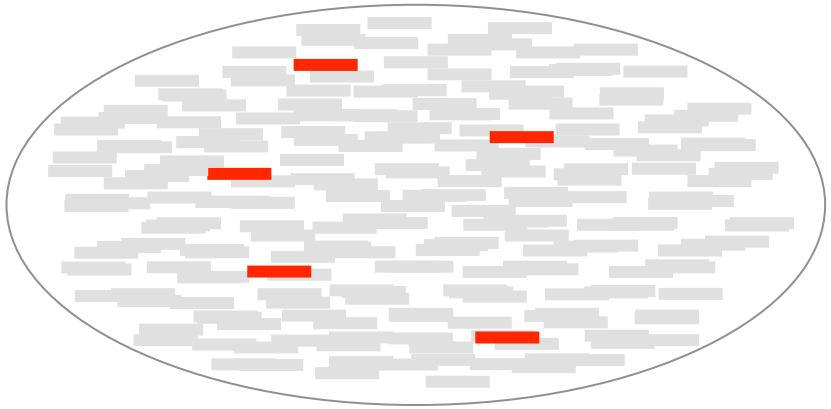



Adaptation to environmental change

1. What are sequences of transcripts?
2. What is the expression level of each transcript?

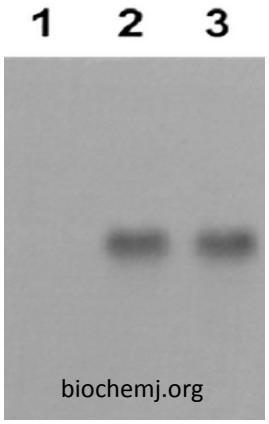

Response to biotic stress

# Approaches for quantification of gene expression

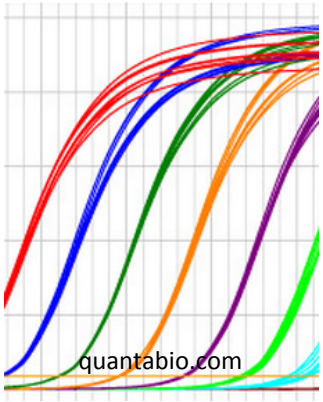


How can we measure the accumulative level of transcripts of **a given gene** in millions/ billions of transcripts?

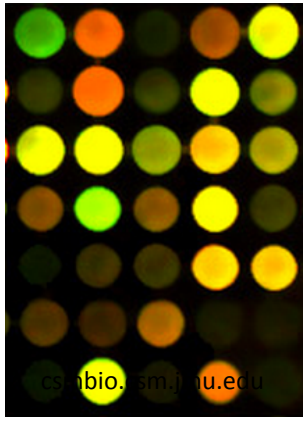
Northern blot




qRT-PCR

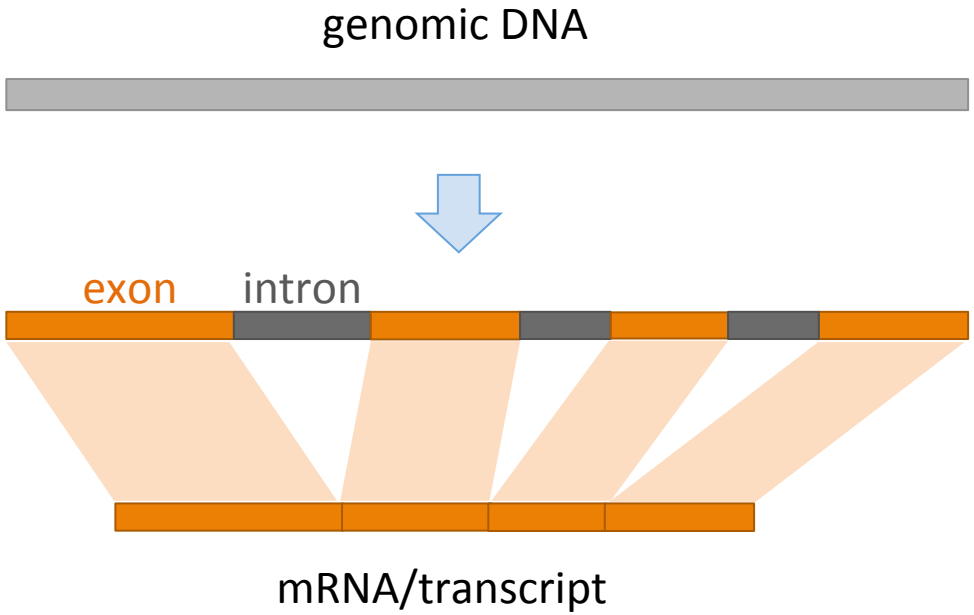



microarray



RNA-Seq

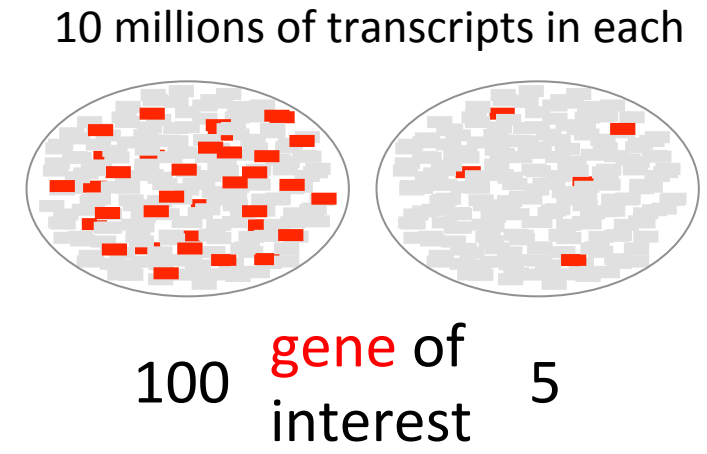
# Rationale of RNA-Seq (mRNA sequencing)



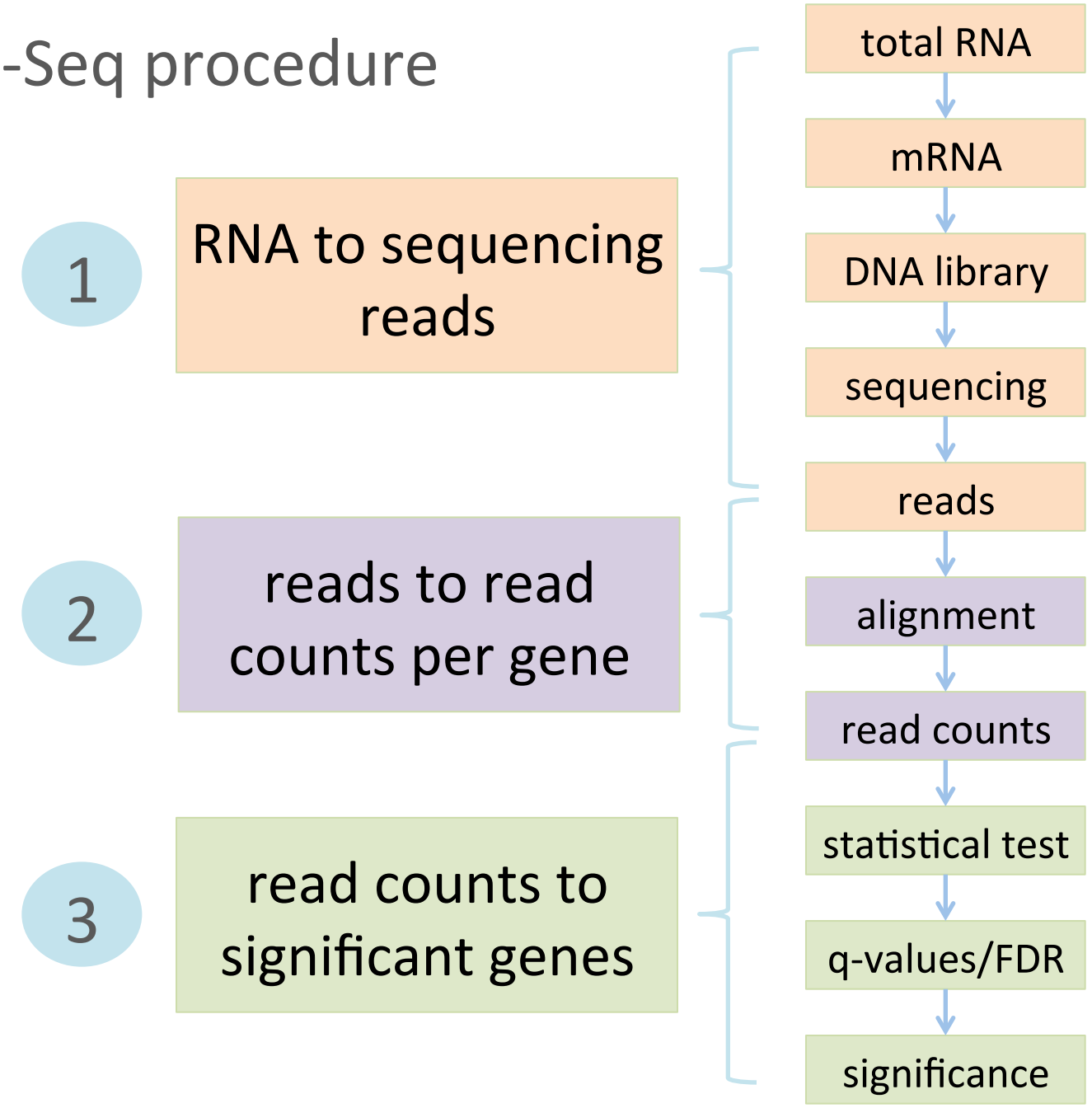

Essentially, RNA-Seq is designed to measure mRNA accumulation levels of genes by

1) recognizing transcripts based on sequences

2) and quantifying transcripts of each gene



# RNA-Seq procedure



total RNA

mRNA

DNA library

sequencing

reads

alignment

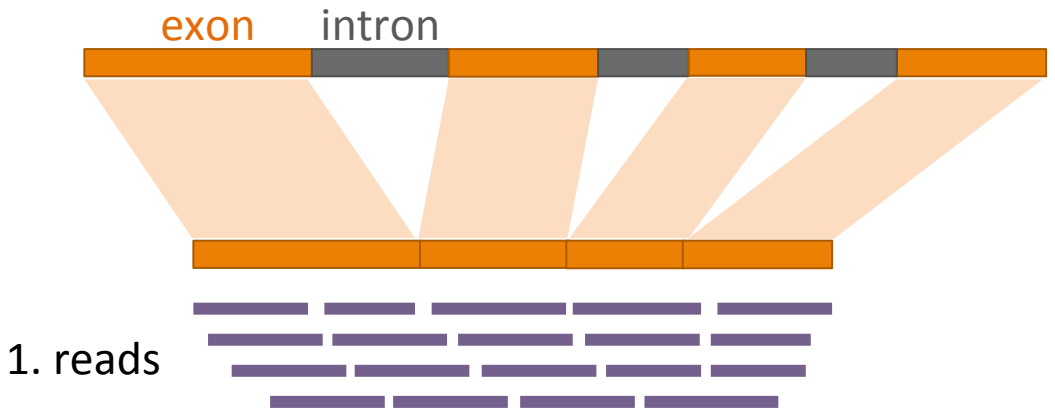
read counts

statistical test

q-values

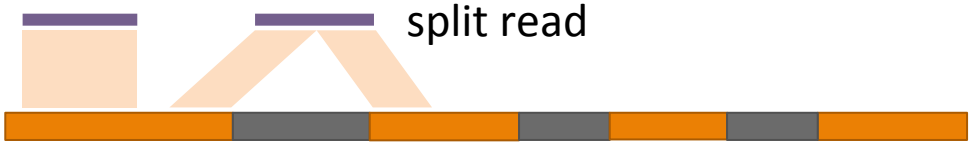
significance

# Reads to read counts per gene




1. reads

2. alignment to the reference genome (DNA sequence)

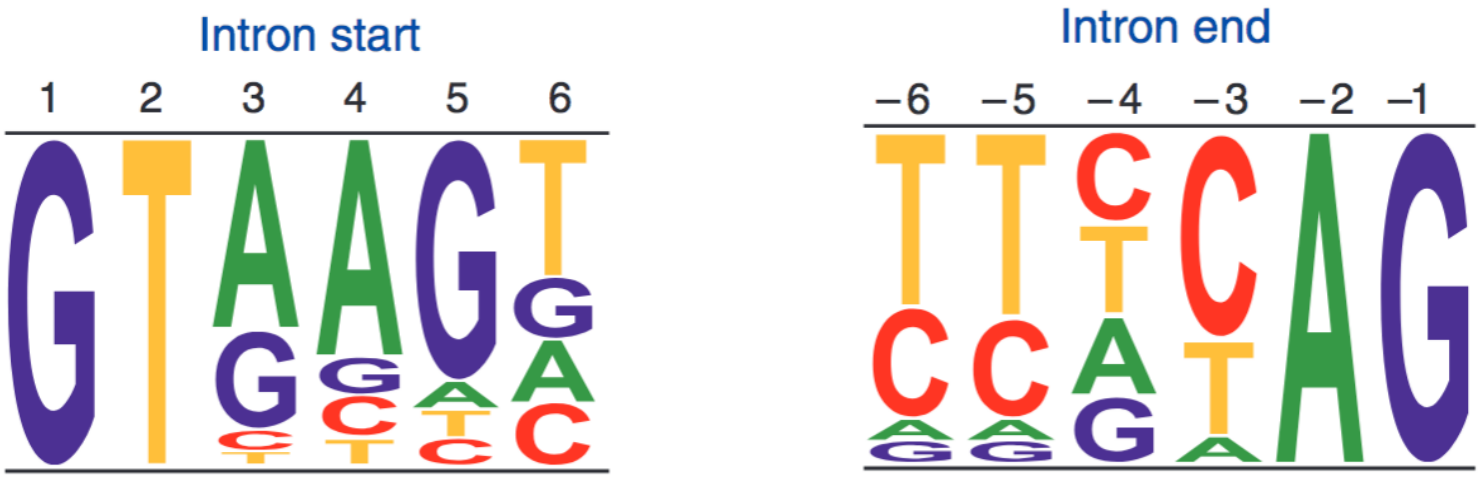



An **intron-aware** aligner is important for RNA-Seq reads alignment
e.g., Tophat, GSNAP, star

3. read counts

N = 19 if all reads can be confidently mapped to the reference genome

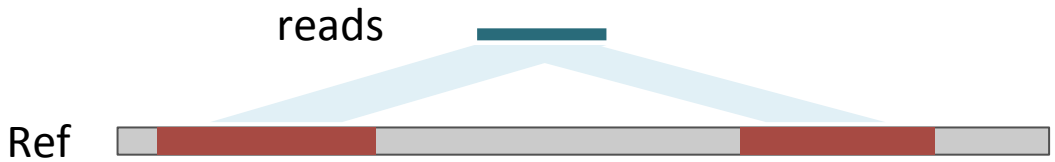
# Exon-intron structure




Sequence logos representing weight matrices for the first six bases of an intron (left) and the last six bases of an intron (right). In plants and animals, ~99% of introns begin with GT.

# Alignment issues

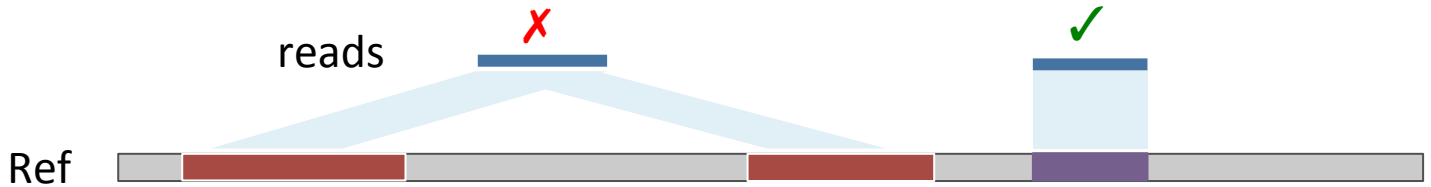
- Repeats



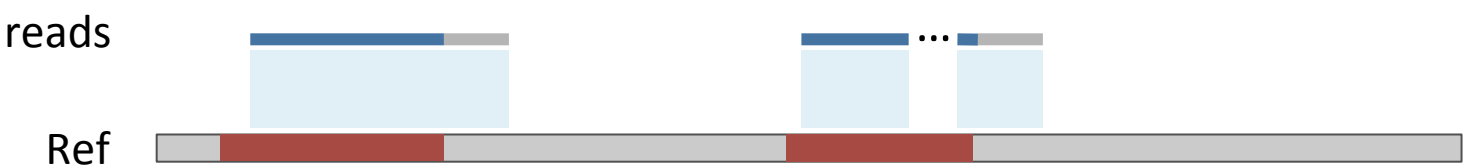

- Sequencing errors
- Polymorphisms (reference and sequenced sample)
- Quality of reference genomes (mis-assembly and incomplete genome)

# Solutions to mitigate problems

- Unique mapped reads

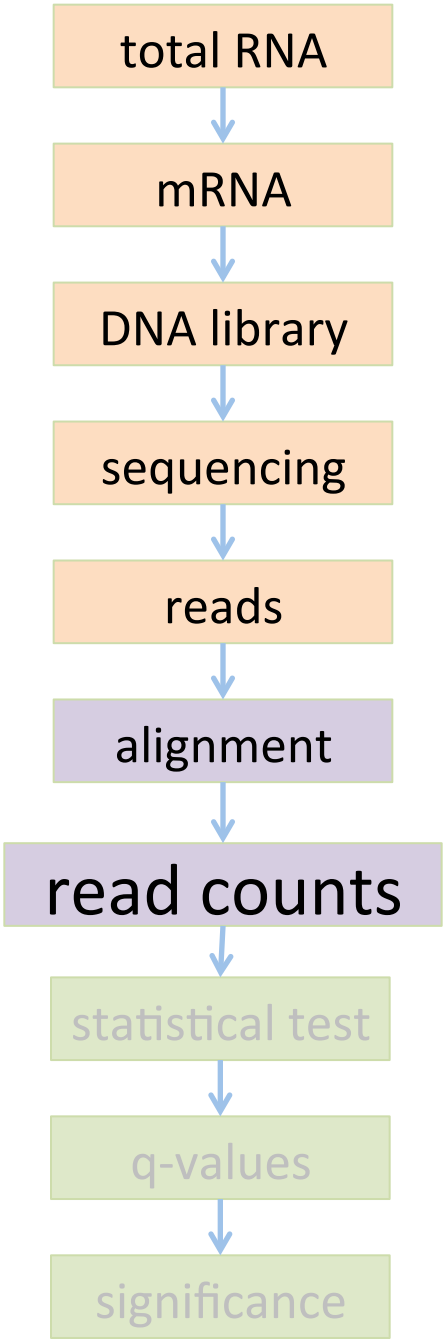



- Longer reads or Paired-end reads




- Tolerance of mismatches or gaps for each alignment

- Better reference genome

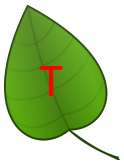



# Count matrix
# Read counts (Raw) per gene

| Gene | sample 1 | sample 2 | sample 3 |
|---|---|---|---|
| gene 1 | 6,075 | 5,934 | 3,370 |
| gene 2 | 295 | 377 | 169 |
| ... | ... | ... | ... |

# Read counts to significant genes

total RNA
↓
mRNA
↓
DNA library
↓
sequencing
↓
reads
↓
alignment
↓
read counts
↓
statistical test
↓
q-values
↓
significance




Control 1 TRT 1 (salt)

RNA
Library
Sequencing

N1 reads N2 reads

Trimming
Alignment
Read counting per gene

| Gene | Control 1 | TRT 1 |
|---|---|---|
| 1 | C1 | T1 |
| 2 | C2 | T2 |
| … | | |
| Sum | N1 | N2 |

2x2 Table for Gene 1

| | Gene 1 | Others |
|---|---|---|
| Control 1 | C1 | N1 – C1 |
| TRT 1 | T1 | N2 – T1 |

- Fisher's Exact Test or $\chi^2$ test on Gene 1

A p-value for Gene 1

- Repeat on all the genes

p-values

- Multiple testing correction

q-values

- Declaration of significance

a significant gene set

# Statistical test for differential expression

- Statistical test to discover differential expression (DE)
  - **Count data**: Generalized Linear Model (GLM) to deal with count data

  e.g., Poisson GLM could handle count data but overdispersion exits
  - **Dispersion issue**: Using **negative binomial GLM** to incorporate dispersion into the model

edgeR (Robinson and Smyth, 2007), **DESeq** (Anders and Huber, 2010), NBPSeq (Di et al., 2011), and QuasiSeq (Lund 2012)

Conesa *et al. Genome Biology* (2016) 17:13
DOI 10.1186/s13059-016-0881-8

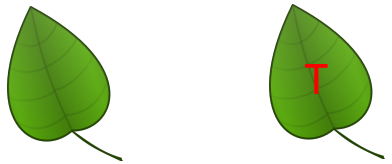
Genome Biology

REVIEW Open Access

A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15] and Ali Mortazavi[16,17*]

# An RNA-Seq experiment – source of variance




Control 1 TRT 1 (salt)

| | | |
|---|---|---|
| Gene 1 | **C1** | **T1** |
| … | | |
| Sum | N1 | N2 |

Our interest:
the effect of the salt treatment on gene expression

**Question**: what would cause the difference between two values, C1 and T1?

- **Treatment effect**
- Plant difference
- RNA quality
- Library preparation
- Sequencing
- Sampling

- Sequencing depth

# Source of variance in RNA-Seq - sampling

- **Sampling variance** derived from the inherent nature of counting experiments

total molecules: $10^9$
gene X: 1000 molecules

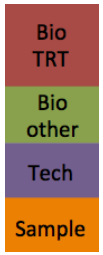
Randomly sample $10^7$

| | |
|---|---|
| First sampling | 6 |
| Second sampling | 13 |
| Third sampling | 8 |

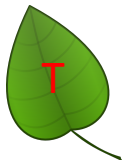
Randomly sample $10^8$

| | |
|---|---|
| First sampling | 102 |
| Second sampling | 93 |
| Third sampling | 97 |

Sequence depth (sampling number) matters.




# Technical replication




Control 1 Control 2 Control 3 TRT 1 TRT 2 TRT 3

within groups within groups

between groups

***Technical replication*** refers to the sequencing of multiple libraries derived from **the same biological sample**.

Technical replicate

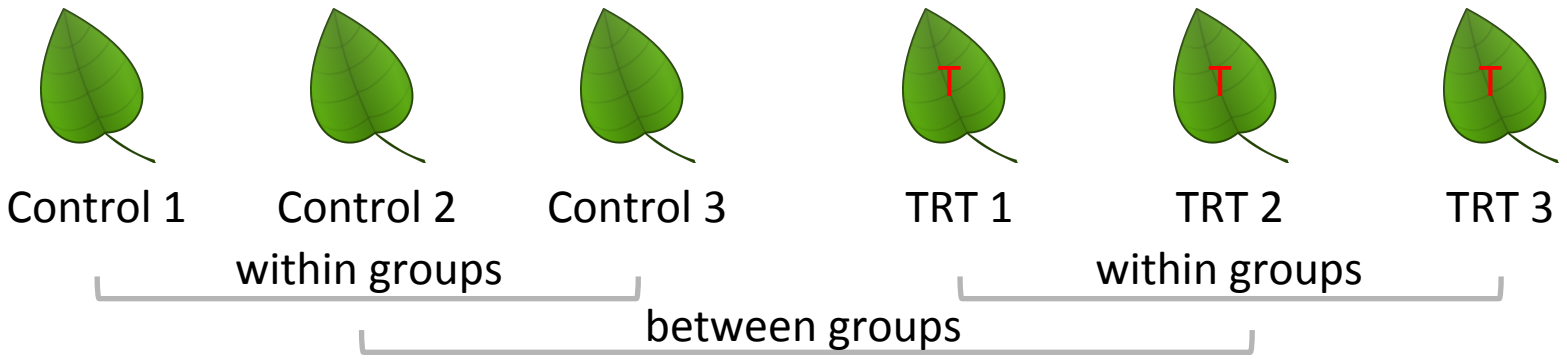

within groups


between groups

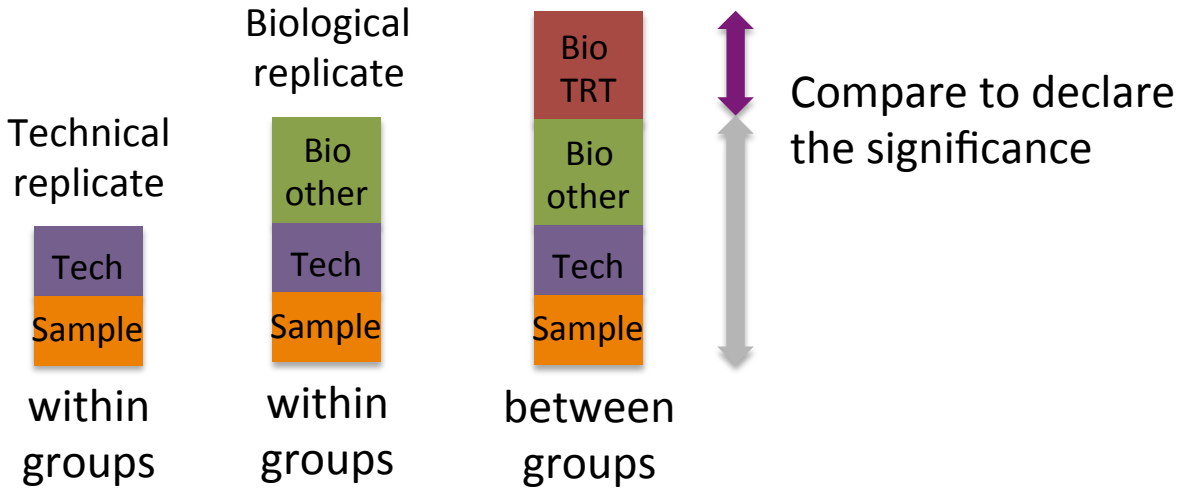
Compare to declare the significance

False power

# Biological replication




***Biological replication*** refers to the sequencing of multiple libraries derived from **different biological samples**.




1. Use biological replication instead of technical replication unless you have your own interest.
2. More replicates increase the power to detect small treatment effect

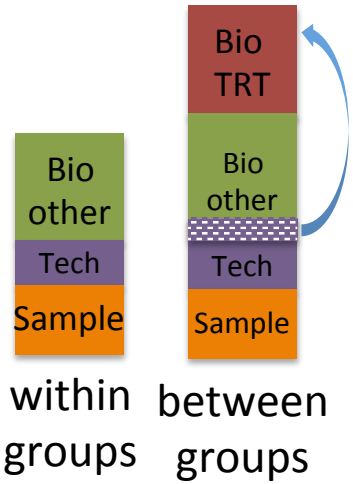
# Question I

Goal: to identify the DEs between two biological groups

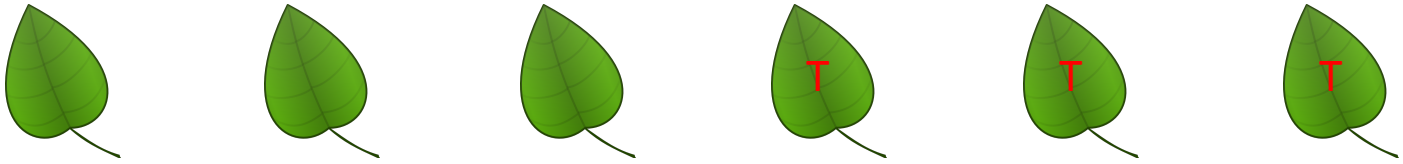
Design: Each group has five biological replicates

To avoid messing up samples across groups. The experiment of each group was conducted separately.

Is this a sound experimental design? Why?



# Comparison of read counts among different samples




| | Control 1 | Control 2 | Control 3 | TRT 1 | TRT 2 | TRT 3 |
|---|---|---|---|---|---|---|
| Gene 1 | C1 | C2 | C3 | T1 | T2 | T3 |
| … | | | | | | |
| Sum | N1 | N2 | N3 | N4 | N5 | N6 |

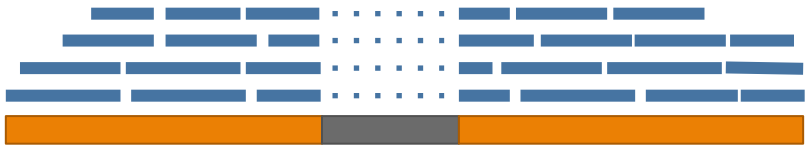
Sequence depth (total read number) influences read counts.
Therefore, raw read counts can not be compared directly.

Can we generate some comparable numbers among samples?

# A **normalization** method: RPKM and FPKM

- **RPKM**: **Read** number per kilobase of exons per million of total reads

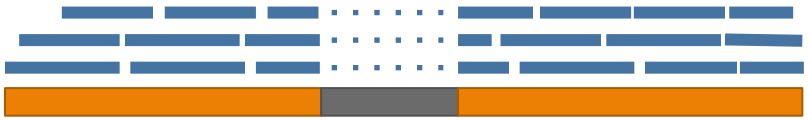
Control 1 read count = **23**



total reads: **15 millions** of total reads

RPKM of X = ? = **3.1**

Treatment 1 read count = **18**



total reads: **10 millions** of total reads

RPKM of X = ? = **3.6**

exon 1 (**220 bp**) exon 2 (**280 bp**)

gene X

- **FPKM**: **Fragment** number per kilobase per million of total reads.

Fragment = one pair of paired-end reads or one single-end read

# More about RPKM


Can we say that the gene B has higher expression than the gene A?

- RPKM is not an ideal indicator to compare the expression/ accumulation levels between two genes

1. amplification bias
2. alignment efficiency

# Experimental Design

- **Sequencing depth**

Increasing sequencing depth decreases sampling variance

- **Biological replication**

Reasonable number of biological replication helps accurately estimate variances to achieve reliable statistical inference.

- **Randomization and unbiasedness**

To avoid confounding effect

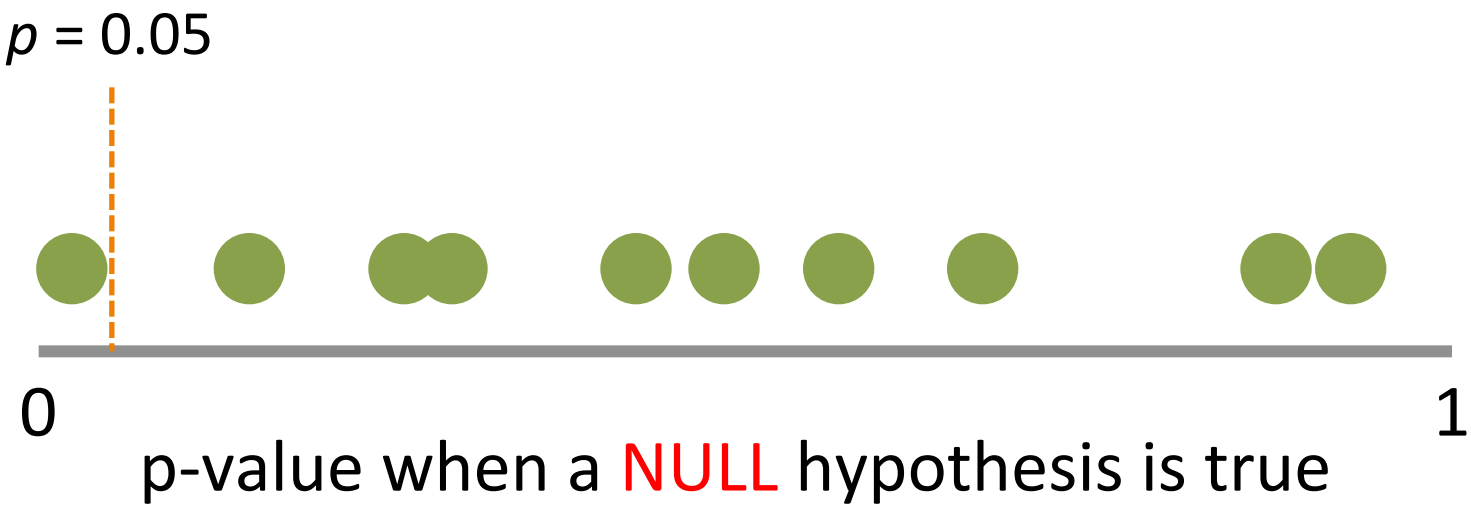
# DE result

| | DE Result | |
|---|---|---|
| GeneID | Log2FC* | p-value |
| 1 | -0.40 | 0.037 |
| 2 | 0.03 | 0.916 |
| 3 | -0.89 | 2.42E-05 |
| 4 | 0.30 | 0.130 |
| 5 | -0.36 | 0.140 |
| 6 | -0.07 | 0.811 |
| … | | |

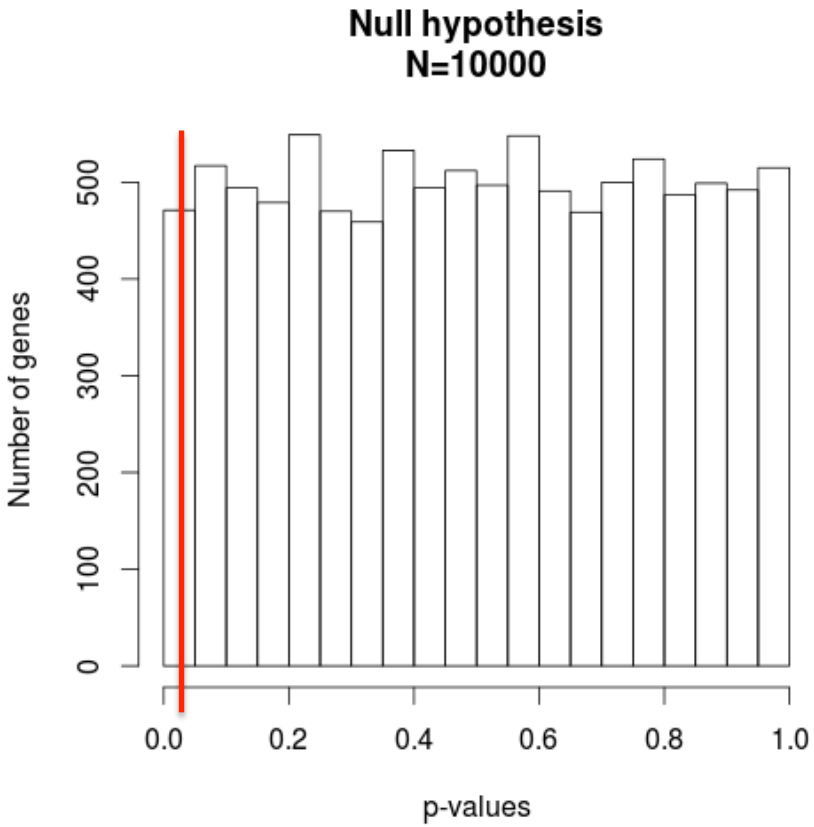
* Log2FC: log2 of fold change (trt / control)

# Multiple testing correction

"A p-value is only statistically valid when a single score is computed."



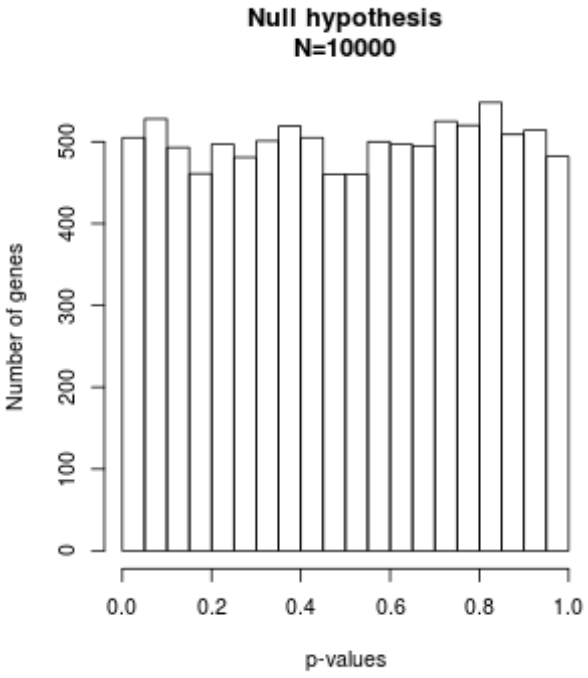
# P-value distribution under the null hypothesis (e.g., no treatment effect)

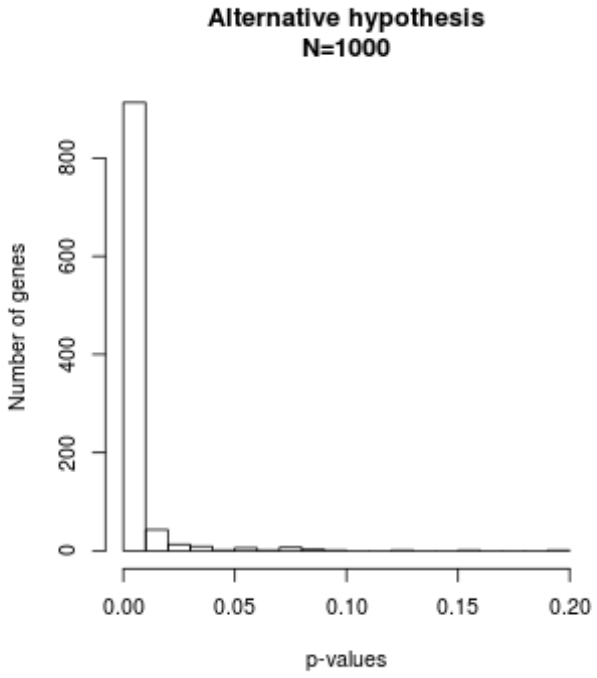


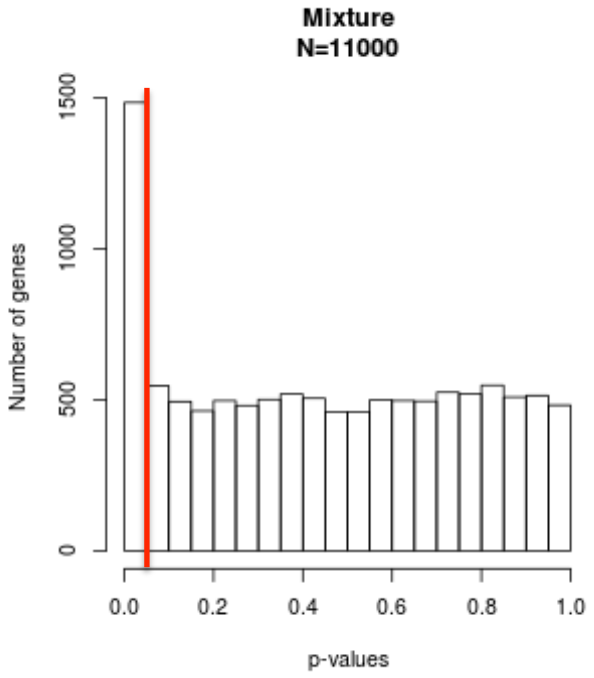

When the null hypothesis is true, a P-value is distributed uniformly from 0 to 1.

No matter how stringent the criteria are, you'll identify genes with very small p-values and the false discovery rate (FDR) is 100%.

# P-value distribution under both the null and non-null hypotheses



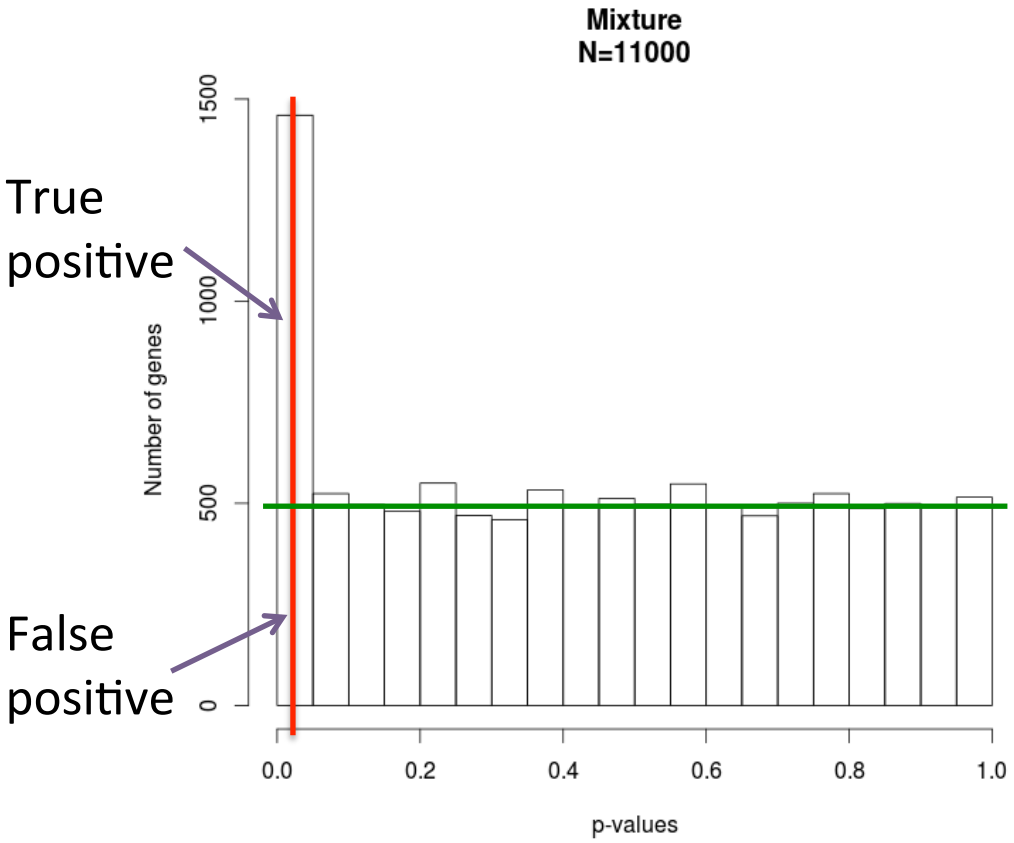

When the null hypothesis is true, a P-value is distributed uniformly.

When the null hypothesis is false, the P-value distribution is skewed toward 0.

Cutoff: p=0.05
FDR=471/(471+989)=32%

Cutoff: p=0.01
FDR=102/(102+912)=10%

# Multiple test correction – FDR method




P-values < 0.00009
DE=992
False DE=99

FDR 10%

# q-values

The **q-value** of a test in a set of tests is **the smallest FDR** for which we can reject the null hypothesis for that one test and all others with smaller p-values.

| Gene | p-values | q-values |
|---|---|---|
| 1 | 0.000 | 0.006 |
| 2 | 0.002 | 0.015 |
| 3 | 0.009 | 0.059 |
| 4 | 0.013 | 0.063 |
| 5 | 0.035 | 0.139 |
| 6 | 0.051 | 0.171 |
| 7 | 0.155 | 0.442 |
| 8 | 0.197 | 0.492 |
| 9 | 0.247 | 0.539 |
| 10 | 0.269 | 0.539 |
| 11 | 0.358 | 0.651 |
| 12 | 0.396 | 0.656 |
| 13 | 0.426 | 0.656 |
| 14 | 0.493 | 0.702 |
| 15 | 0.526 | 0.702 |
| 16 | 0.622 | 0.777 |
| 17 | 0.782 | 0.920 |
| 18 | 0.862 | 0.958 |
| 19 | 0.925 | 0.974 |
| 20 | 0.992 | 0.992 |

Total number of tests: m = 20

FDR method (BH) is a method to calculate q-values/adjusted p-values/ corrected p-values based on p-values

5% FDR, q-values < 0.05

10% FDR, q-values < 0.1

20% FDR, q-values < 0.2

# Question

If we identify 500 differential expression (DE) genes using the 5% FDR to account for multiple tests. Which one below is a better description?

1. I am 95% confident that 500 genes are DE.

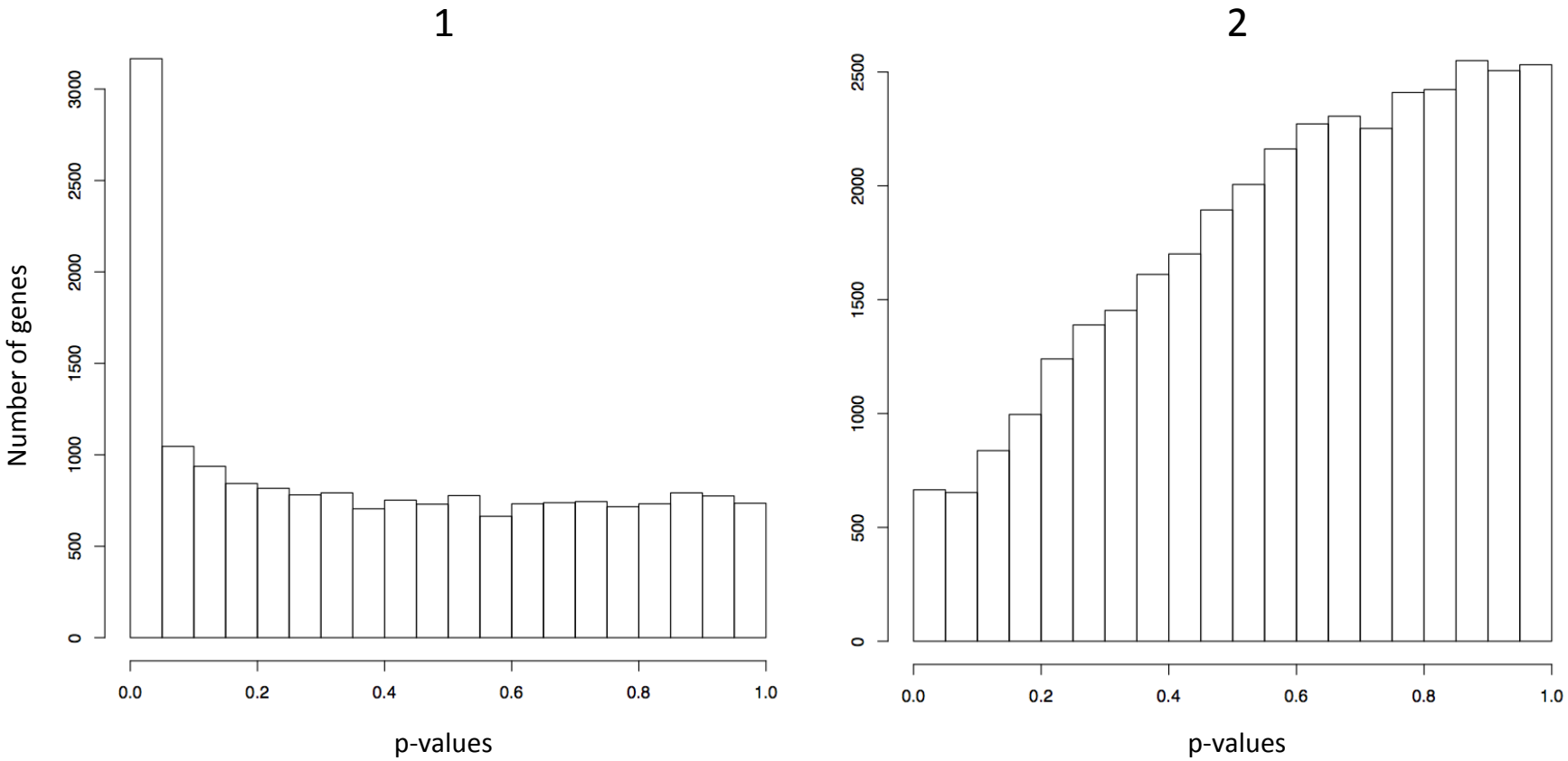
2. The 5% genes (25 genes) in the set are expected to be false DE genes.

# False discovery rate (concept)

For example, among 10,000 tests (10,000 genes), 100 significant genes are declared, in which 10 gene is falsely rejected. In this case, the false discovery rate is 10%.

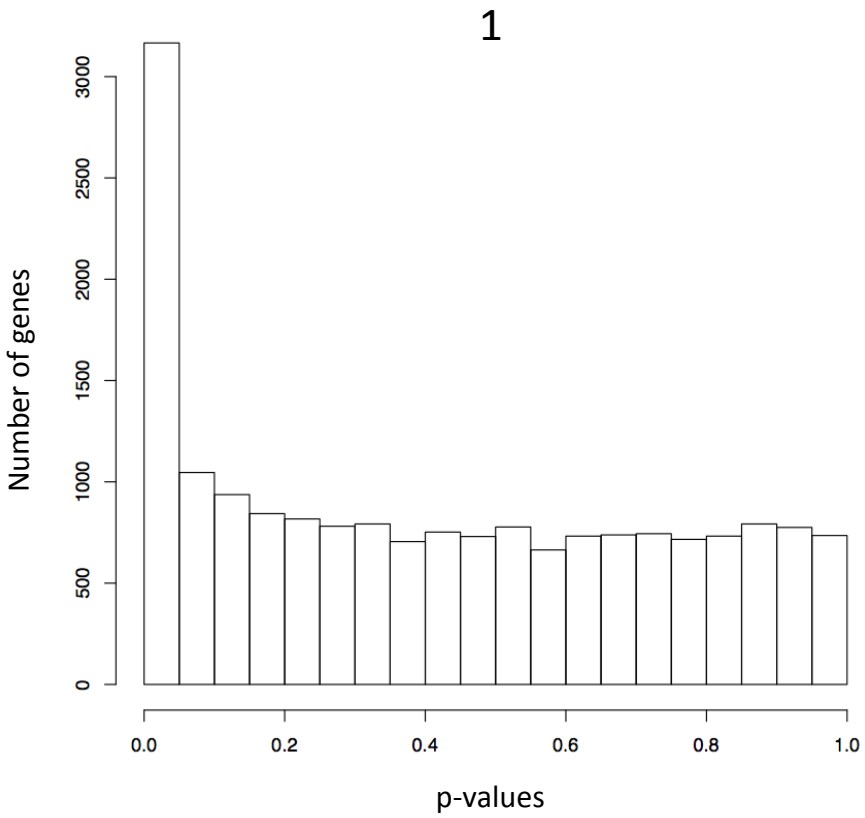
| | True null hypothesis ($H_0$) | False null hypothesis ($H_1$) | Total |
|---|---|---|---|
| Rejected (Declared significance) | **10** | 90 | 100 |

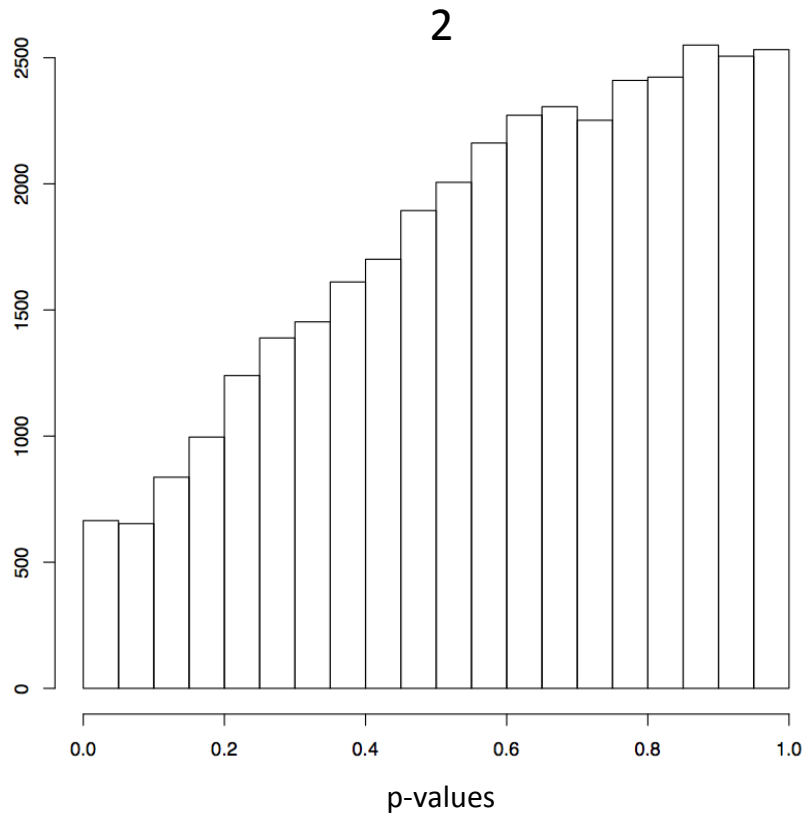
# P-value histograms from real studies




If you perform an RNA-Seq experiment, which one would you hope to obtain? Why?

# P-value histograms from real studies

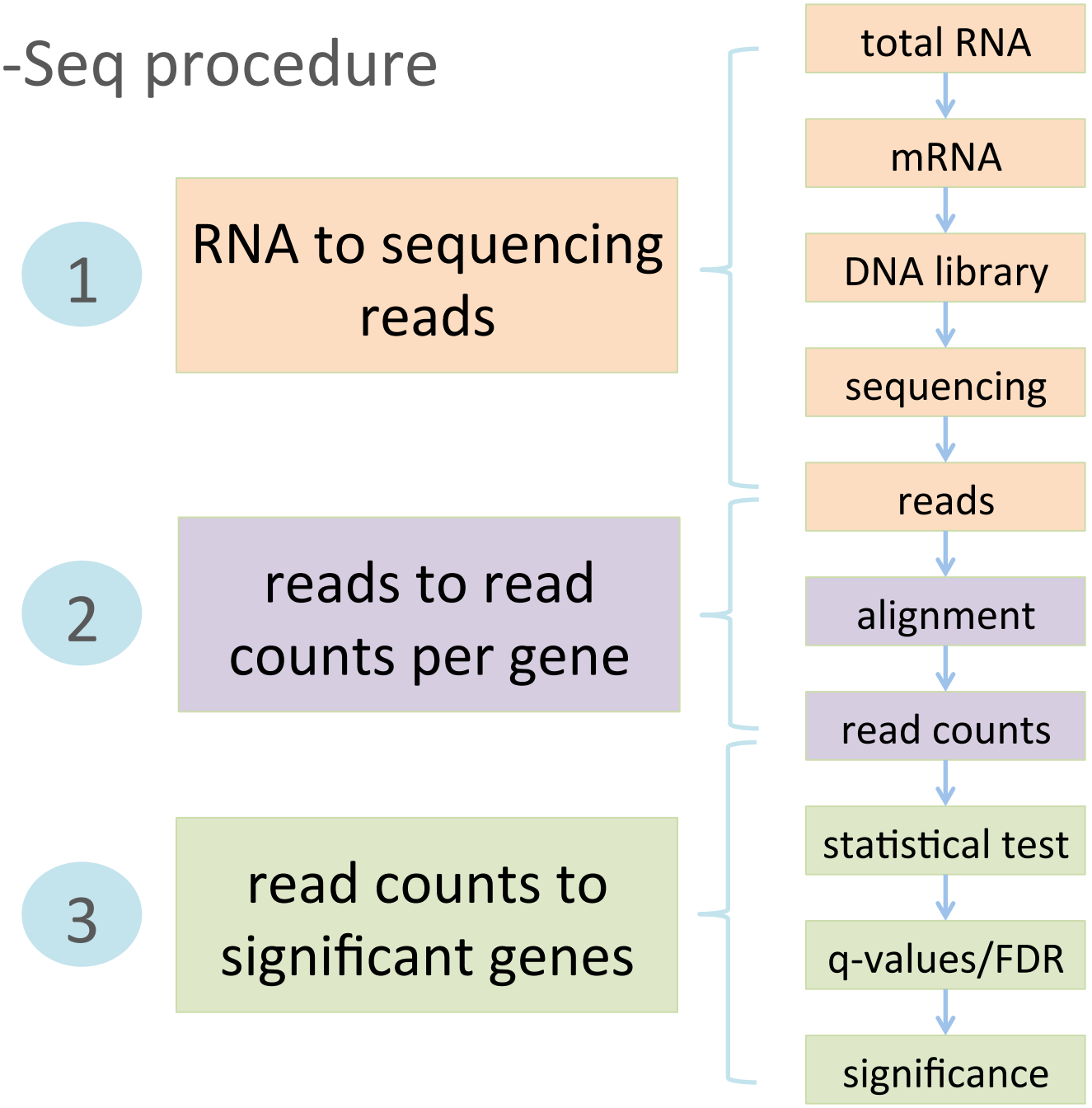


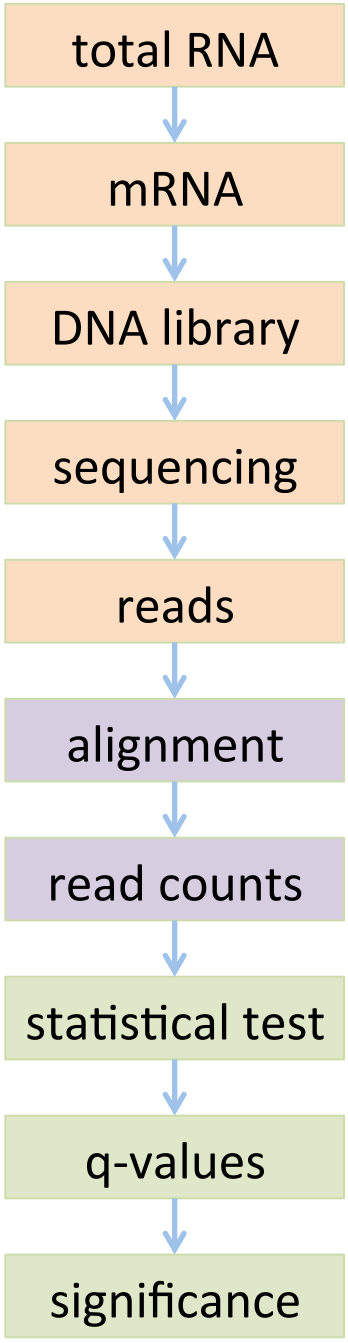

DE = 1,370, FDR=5%

DE = 0, FDR=20%

# RNA-Seq procedure



# Keywords




randomization, replication, RNA quality

short or long reads
single- or paired-end reads, read length
sequencing depths
(e.g., >20 million short reads for most experiments)

intron-spanning Aligner (e.g., GSNAP, STAR)

count data statistical analysis (DESeq2 & edgeR)
multiple test p-value adjustment

# Summary

- Biological replication rather than technical replication are typically needed for an RNA-Seq experiment.
- P-values need to be corrected to account for multiple tests. The FDR method is a reliable approach for the correction.
- Many bioinformatics pipelines and statistical methods have been developed. Most methods work fine but the parameters in each method need to be carefully selected.