Differential Expression

Bioinformatics Applications (PLPTH813)

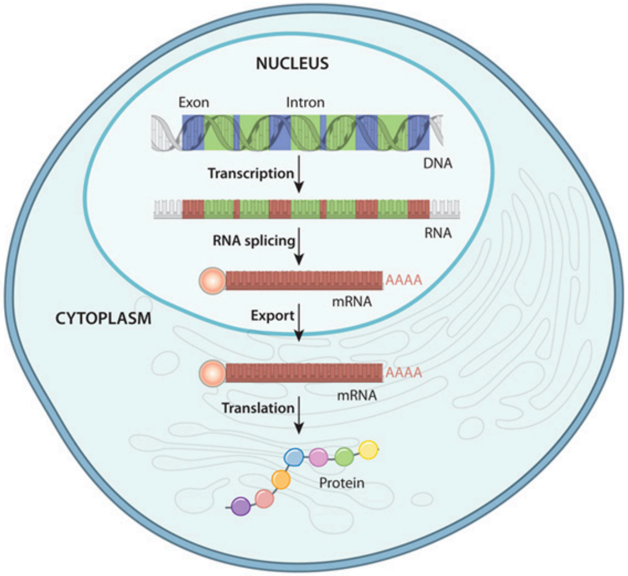
Sanzhen Liu

4/11/2017

Outline

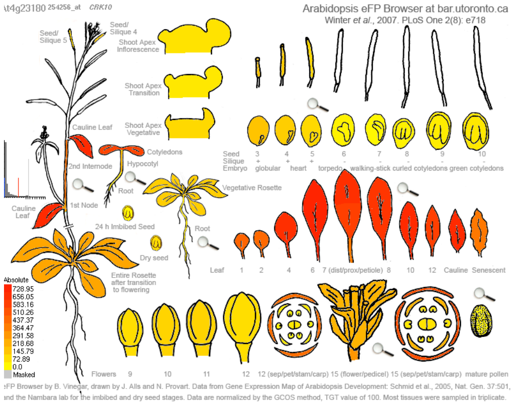
- Introduction of RNA-Seq
- RNA-Seq procedure
- Data normalization
- Statistical test of differential expression
- Multiple testing correction

Gene expression



DNA to protein in eukaryote

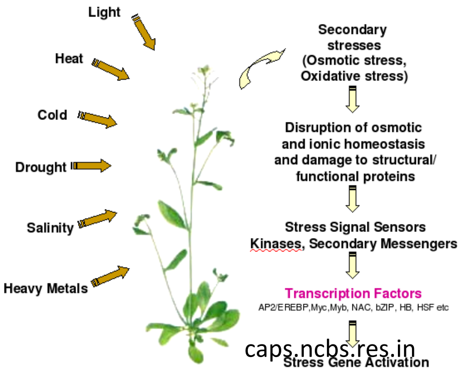
nature.com/scitable/topicpage/gene-expression-14121669



Arabidopsis eFP Browser at bar.utoronto.ca
Winter *et al.*, 2007. PLoS One 2(8): e718

eFP Browser by B. Vinegar, drawn by J. Alls and N. Provart. Data from Gene Expression Map of Arabidopsis Development: Schmid et al., 2005, Nat. Gen. 37:501, and the Nambara lab for the imbibed and dry seed stages. Data are normalized by the GCOS method, TGT value of 100. Most tissues were sampled in triplicate.

Expression profiles in different tissues



Light

Heat

Cold

Drought

Salinity

Heavy Metals

Secondary stresses (Osmotic stress, Oxidative stress)

Disruption of osmotic and ionic homeostasis and damage to structural/ functional proteins

Stress Signal Sensors Kinases, Secondary Messengers

Transcription Factors
AP2/EREBP,Myc,Myb, NAC, bZIP, HB, HSF etc

Stress Gene Activation

caps.ncbs.res.in

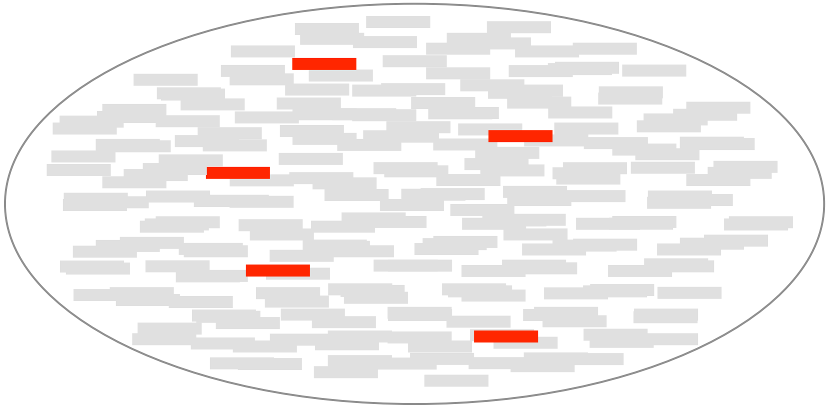
Adaptation to environmental change

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

cragenomica.es

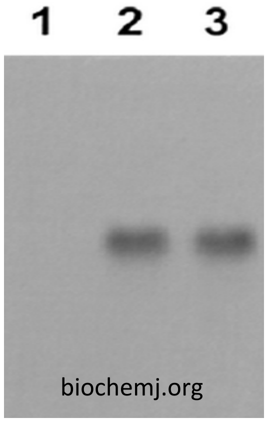
Response to biotic stress

Approaches for quantification of gene expression



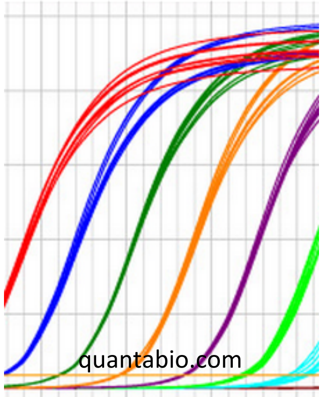
How can we measure the accumulative level of transcripts of **a given gene** in millions/ billions of transcripts?

Northern blot



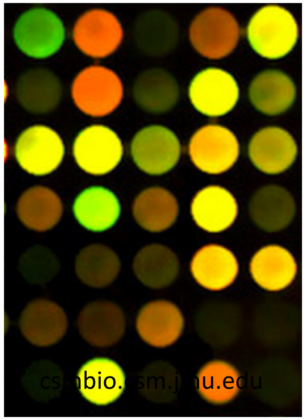
biochemj.org

qRT-PCR



quantabio.com

microarray

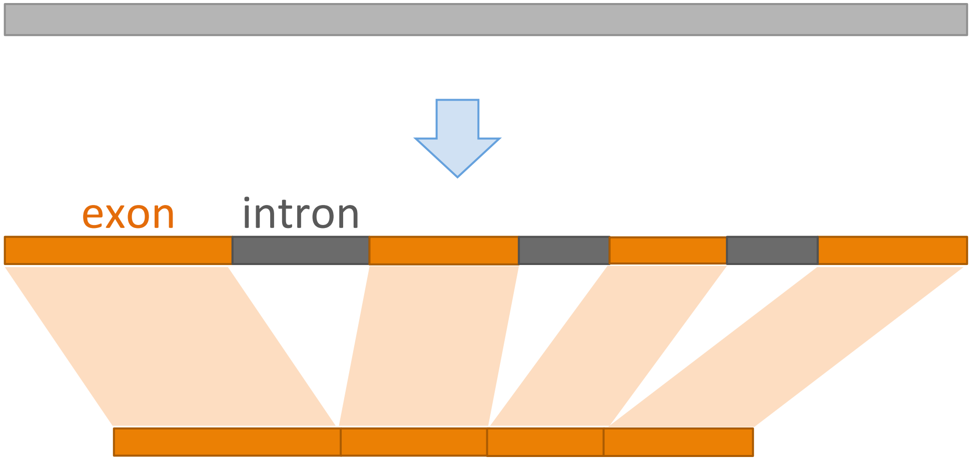


csmbio.csm.jmu.edu

RNA-Seq

Rationale of RNA-Seq (mRNA sequencing)

genomic DNA



exon intron

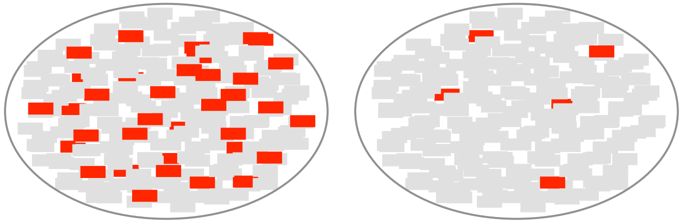
mRNA/transcript

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

10 millions of transcripts in each



100 gene of interest 5

sequence 1,000 transcripts

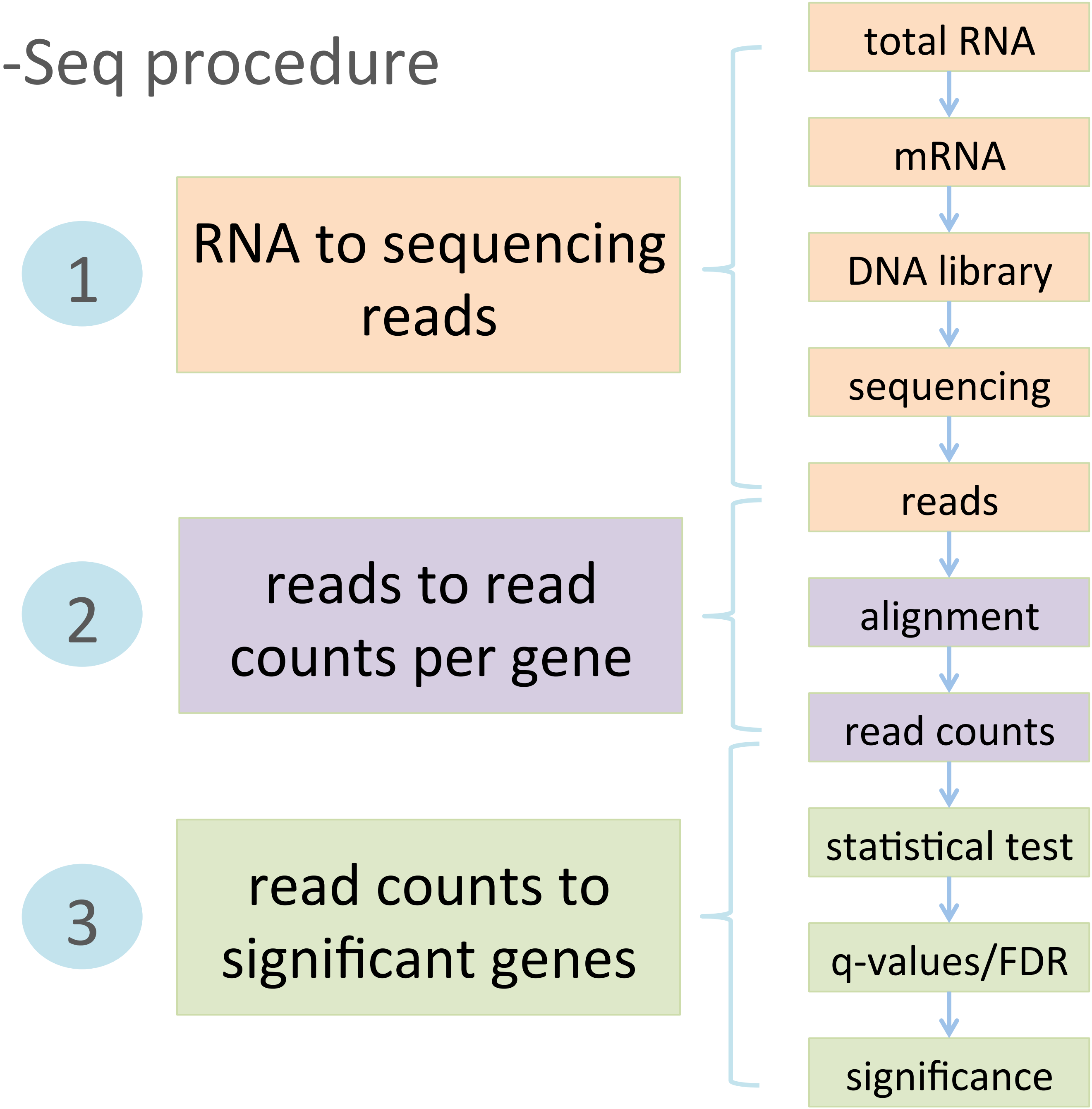
0 0

sequence **1 million transcripts**

10 1

Differential expression?

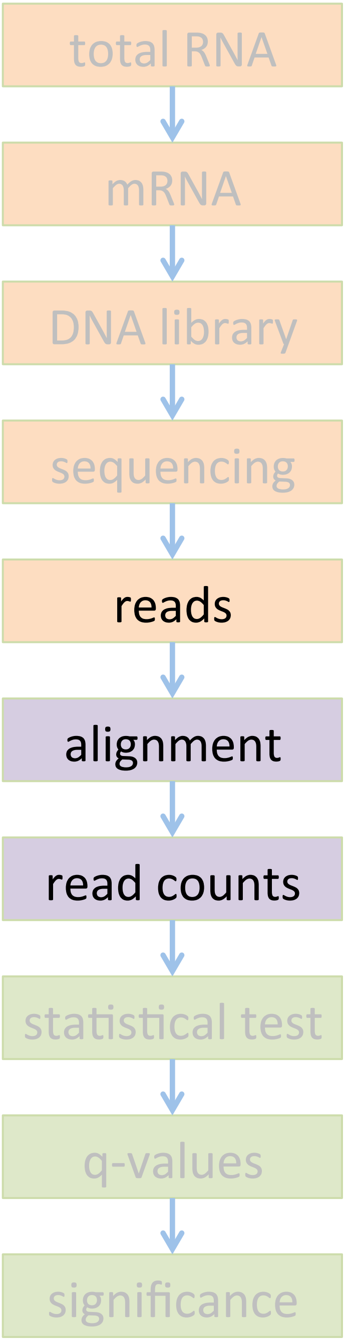
RNA-Seq procedure

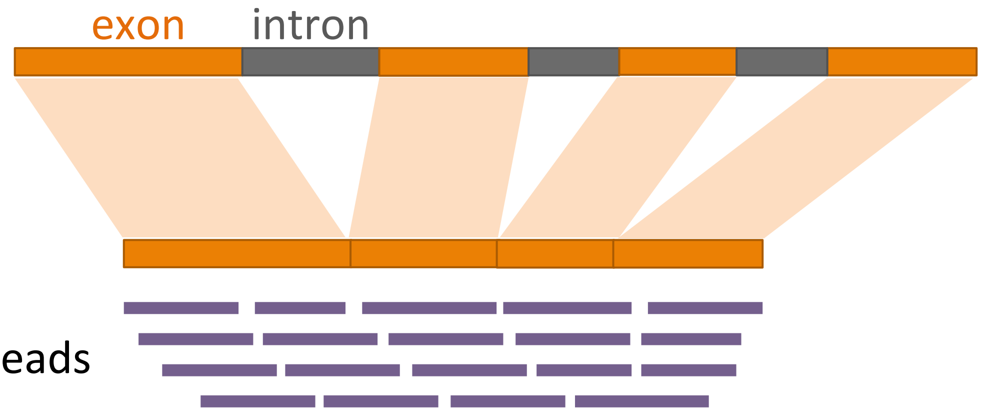


1. RNA to sequencing reads
2. reads to read counts per gene
3. read counts to significant genes

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

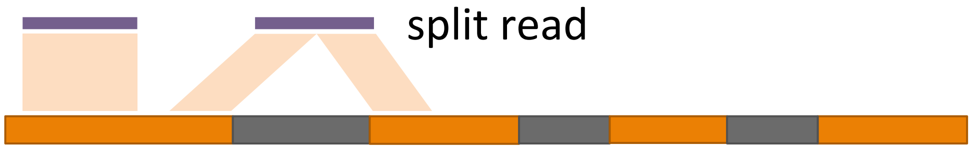
Reads to read counts per gene





1. reads

2. alignment to the reference genome (DNA sequence)

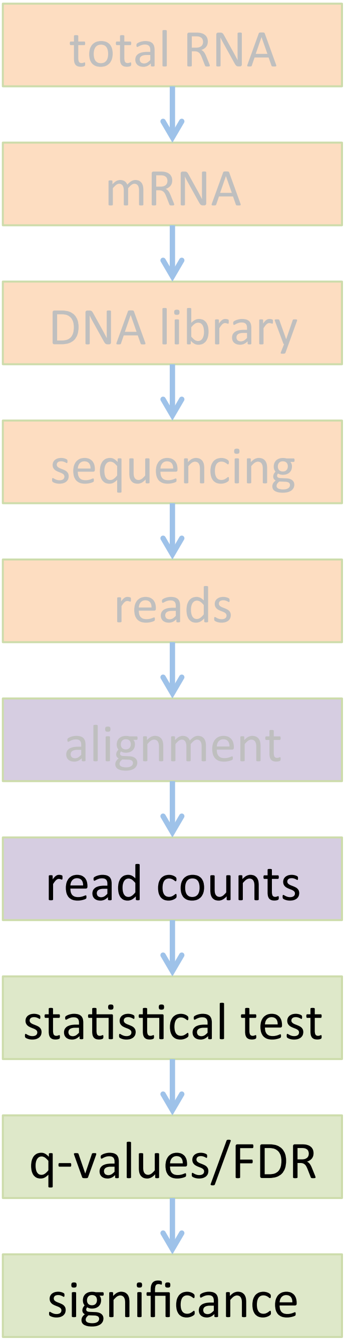


An **intron-aware** aligner is important for RNA-Seq reads alignment e.g., Tophat, GSNAP, star

3. read counts

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

Read counts to significant genes



total RNA

mRNA

DNA library

sequencing

reads

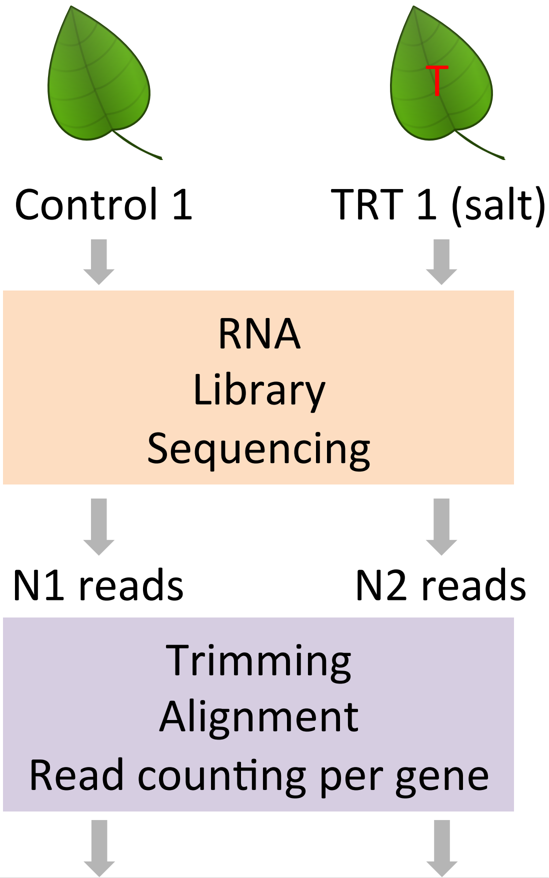
alignment

read counts

statistical test

q-values/FDR

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 χ^2 test on Gene 1

A p-value for Gene 1

- Repeat on all the genes

p-values

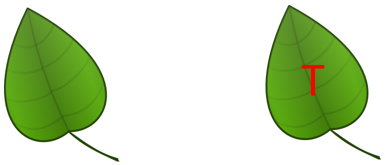
- Multiple testing correction

q-values

- Declaration of significance

a significant gene set

An RNA-Seq experiment – source of variance



Control 1 TRT 1 (salt)

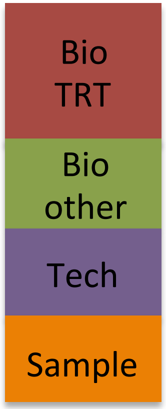
Gene 1	**C1**	**T1**
…		
Sum	N1	N2

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

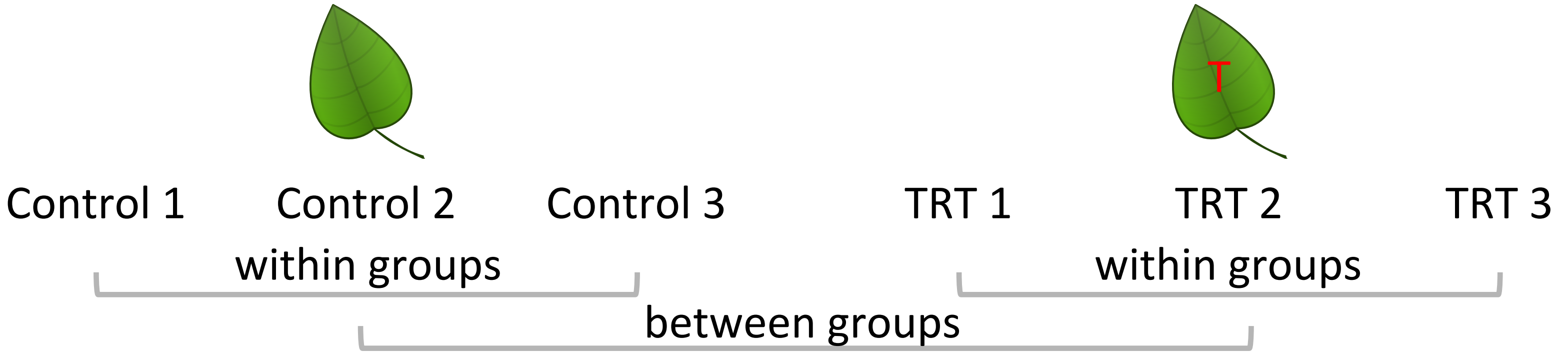

Bio TRT

Bio other

Tech

Sample

Technical replication



Control 1 Control 2 Control 3

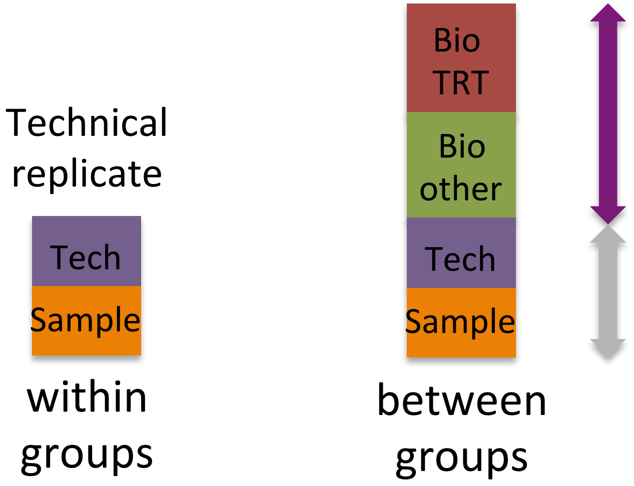
within groups

TRT 1 TRT 2 TRT 3

within groups

between groups

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



Technical replicate

Tech

Sample

within groups

Bio TRT

Bio other

Tech

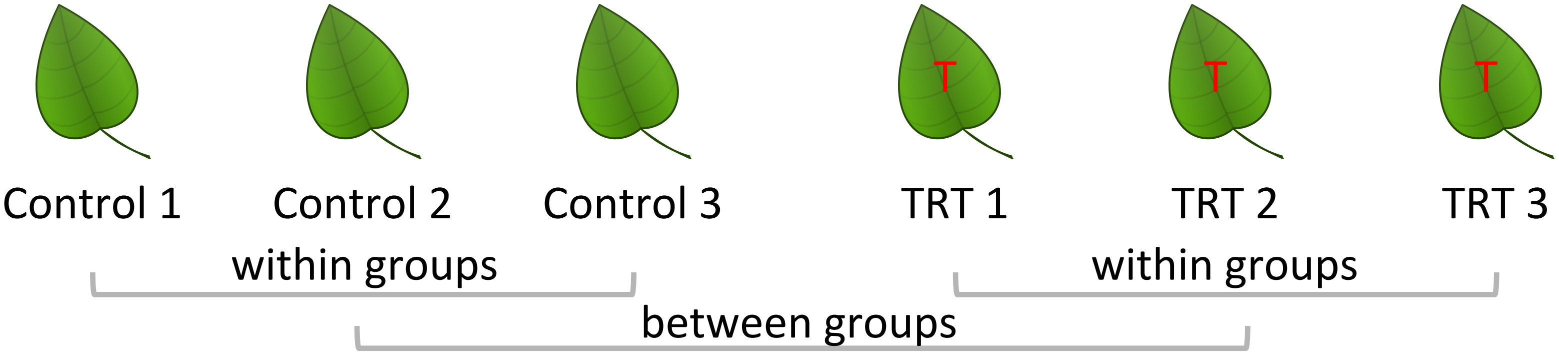
Sample

between groups

Compare to declare the significance

False power

Biological replication



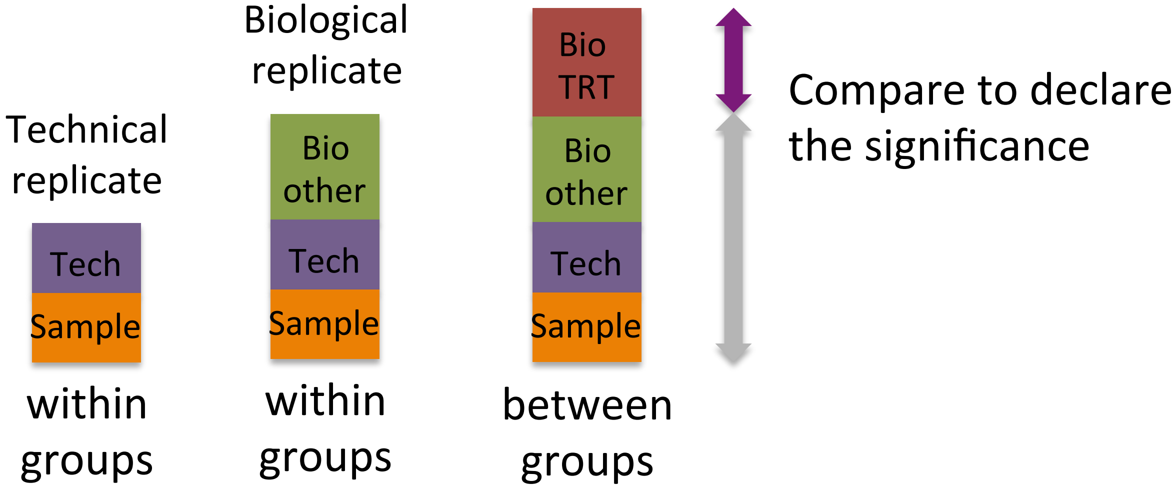
Control 1 Control 2 Control 3 TRT 1 TRT 2 TRT 3

within groups

within groups

between groups

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



Technical replicate

Tech

Sample

within groups

Biological replicate

Bio other

Tech

Sample

within groups

Bio TRT

Bio other

Tech

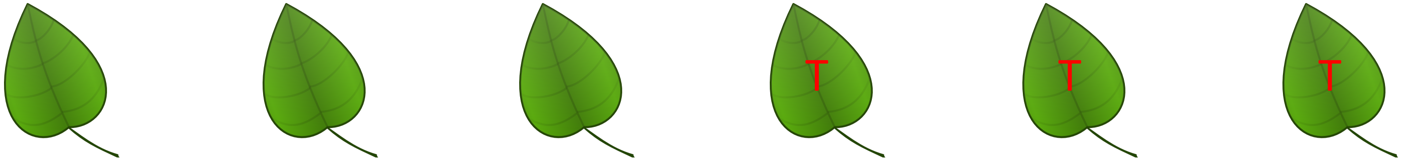
Sample

between groups

Compare to declare the significance

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

Comparison among read counts



	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

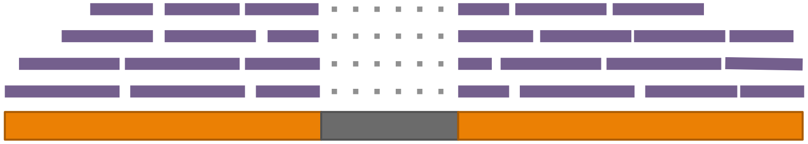
Sequence depth (total read number) influences read counts.

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

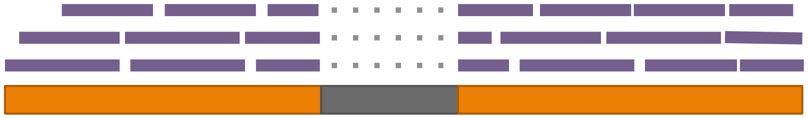
Control 1 read count = **23**



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

RPKM of X = ? = **3.1**

Treatment 1 read count = **18**



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

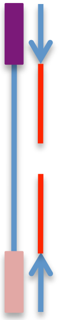
gene X

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

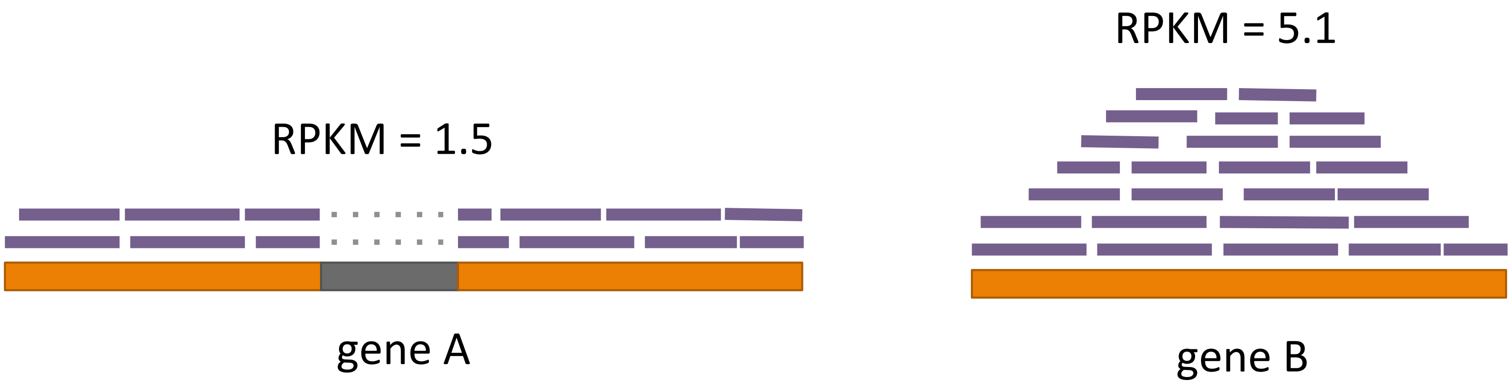
RPKM of X = ? = **3.6**

- **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

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

  - **Small n problem**: a few number of replication

  Borrowing information across all the genes to estimate gene-specific variation

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

single test vs. multiple tests

- **Single test:**

p = 0.03

At the 5% significant level (P-value threshold = 0.05), we can reject the null hypothesis.

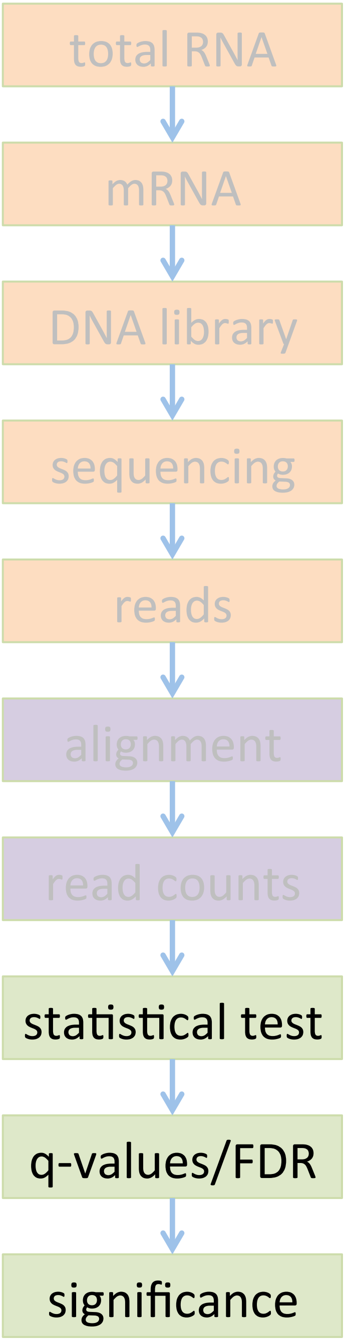
- **Multiple tests:**

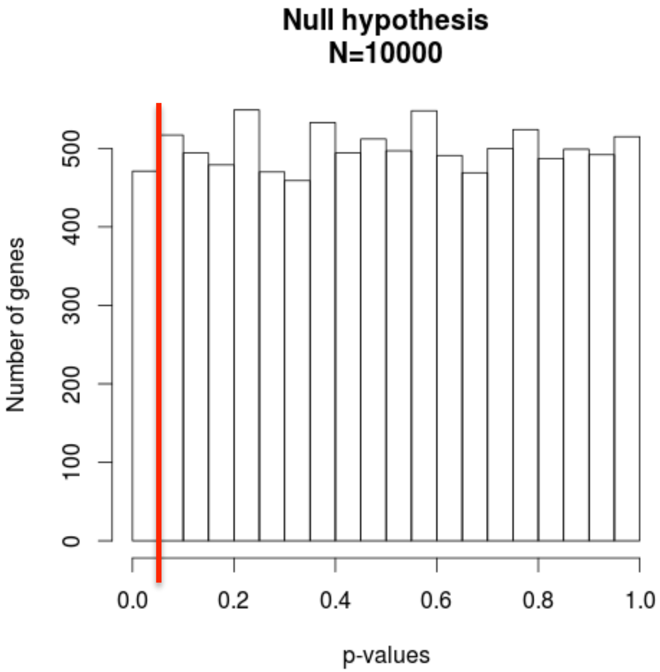
p1 = 0.8; p2 = 0.1; p3 = 0.3; p4 = 0.5; ...; p20 = 0.03

At the 5% significant level (P-value threshold = 0.05), we will reject the null hypothesis for p20.

Anything wrong here?

Multiple testing problem


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



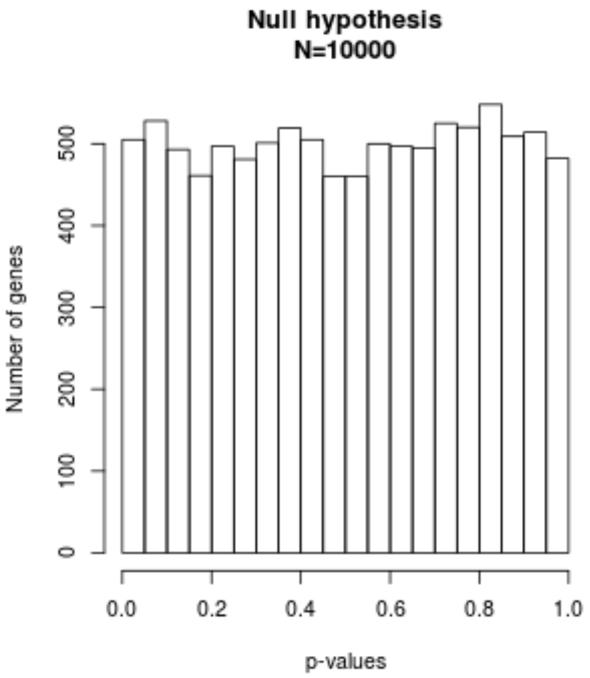
10,000 tests in total

200 (5% * 10,000) tests are expected to show p-values smaller than 0.05.

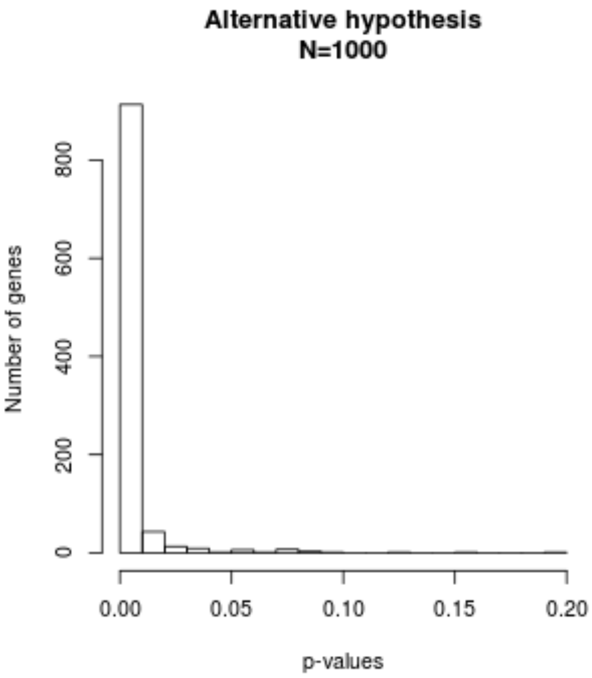
When the null hypothesis is true for every tests and these tests are independent, P-values are distributed uniformly from 0 to 1.

p1 = 0.8; p2 = 0.1; p3 = 0.3; p4 = 0.5; ...; p20 = 0.03

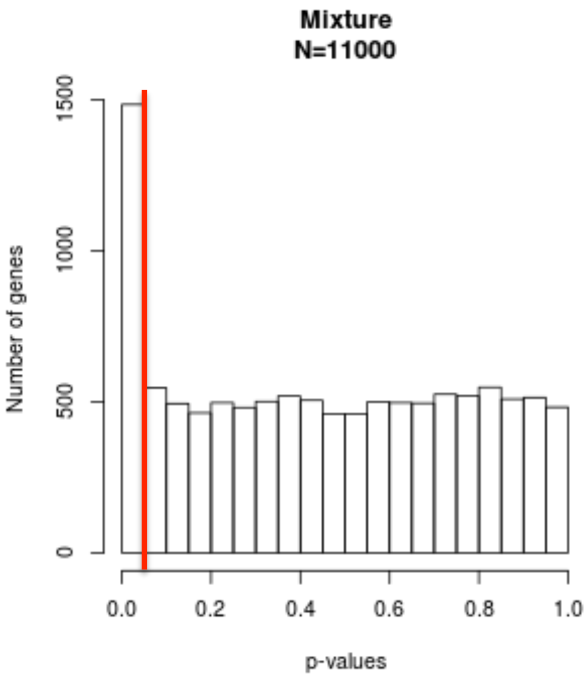
P-value distribution under both the null and alternative hypotheses



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

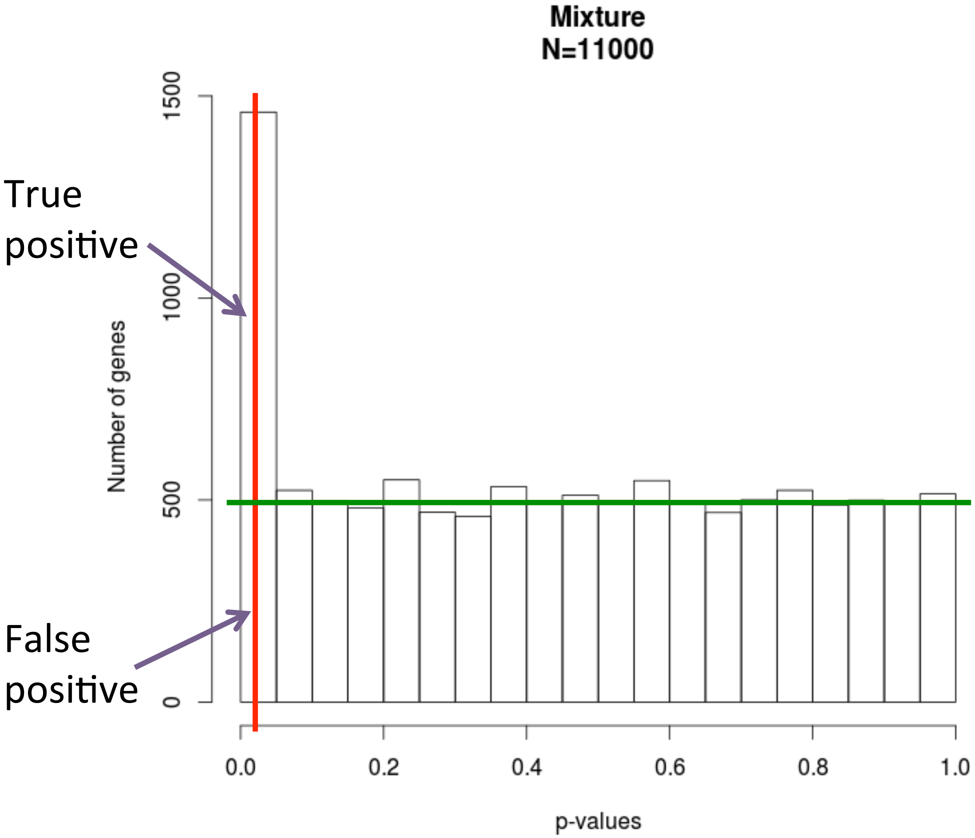


When the null hypothesis is false, the P-value distribution is skewed toward 0.



P-value cutoff: p=0.05? p=0.01? or others?

False discovery rate (procedure)



The FDR "procedure"

1. set up a FDR level
2. determine a P-value cutoff*
3. Any P-values smaller than the cutoff will be rejected.

FDR 10%
P-values < 0.00009
DE=992
False DE=99

* If a P-value cutoff fails to be determined, no tests should be rejected.

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.

k	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

$$q(\mathrm{i}) = \min \{\, p(k)\, m \,/\, k : k = i,\ldots,m \,\}$$

5% FDR, q-values < 0.05

10% FDR, q-values < 0.1

20% FDR, q-values < 0.2

False discovery rate (concept)

FDR: the expected error rate of a set of genes declared to be DE.

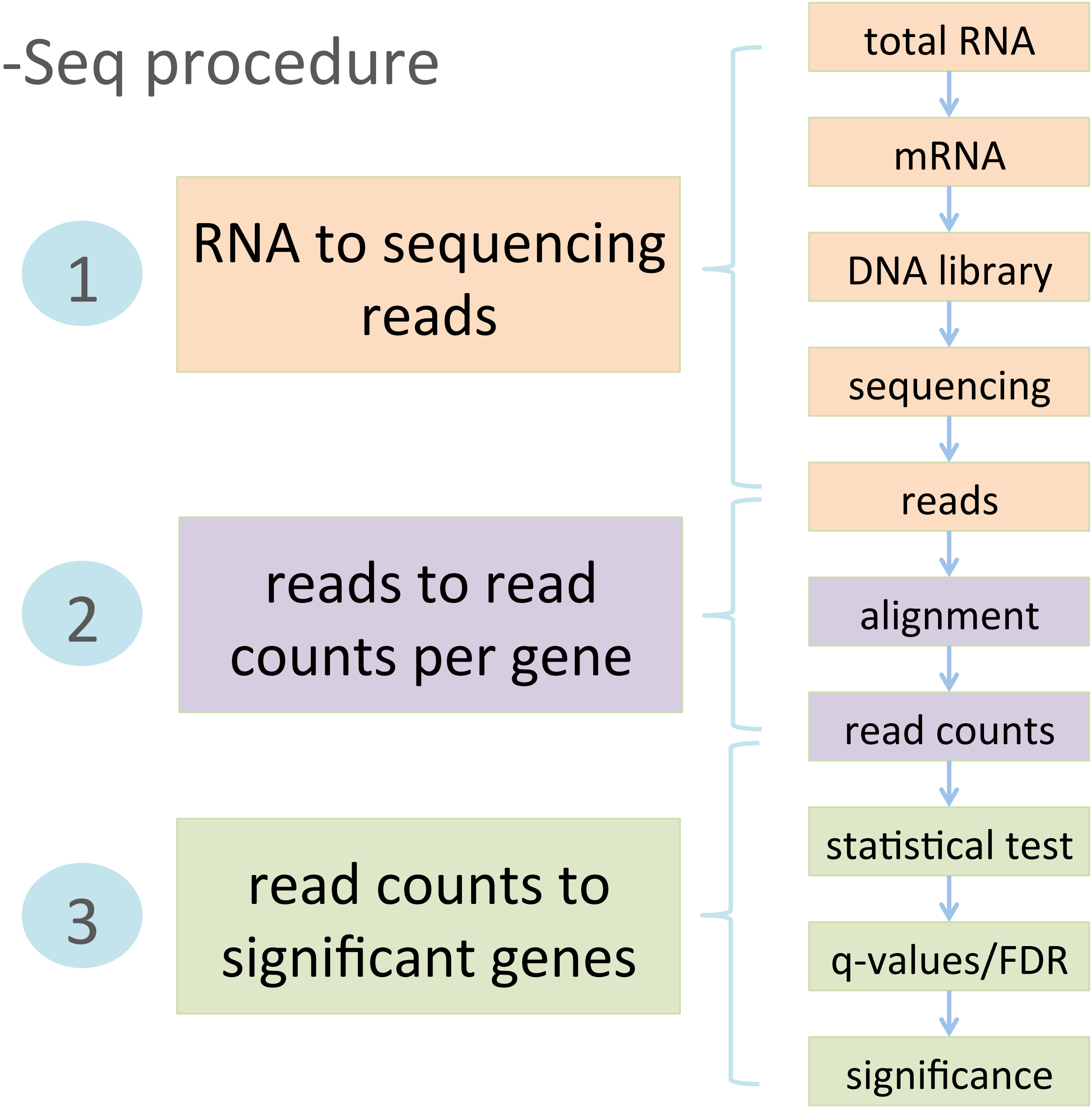
	True null hypothesis (H_0)	False null hypothesis (H_1)	Total
Rejected (Declared significant)	**V**	S	R

FDR: the expected value of V/R

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 significant)	**10**	90	100

RNA-Seq procedure



1 RNA to sequencing reads

2 reads to read counts per gene

3 read counts to significant genes

total RNA

mRNA

DNA library

sequencing

reads

alignment

read counts

statistical test

q-values/FDR

significance

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.