Design of RNA-seq Experiments and Differential Expression Analysis

Genomic Technologies Workshop
(PLPTH885)

Sanzhen Liu

6/8/2022

Schedule

- 9:30 am **Lecture 6** - Sanzhen Liu

Design of RNA-Seq Experiments and Differential Expression Analysis

- 10:50 am **Break**

- 11:00 am **Computer Lab 2** - Guifang Lin, Sanzhen Liu

Introduction to R programming

- 12:30 pm **Lunch on your own**

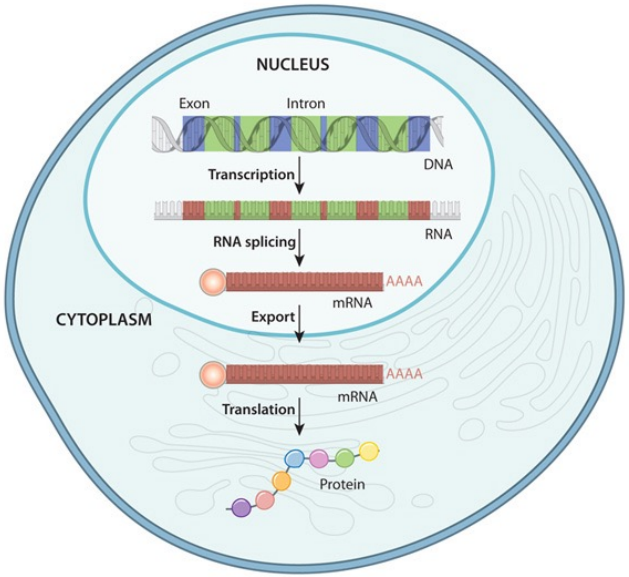
- 1:30 - 3pm **Computer Lab 3** - Sanzhen Liu, Guifang Lin

RNA-Seq data analysis using R

Outline

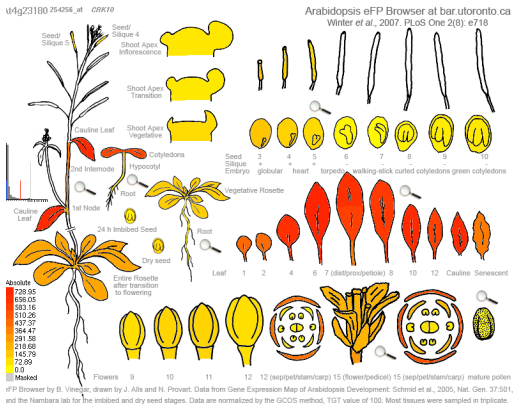
- RNA-seq procedure
- Experimental design
- Multiple testing correction
- Data visualization
- Gene ontology (GO) enrichment

Gene expression

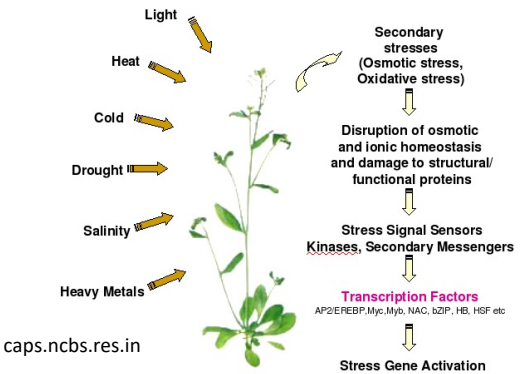


DNA to protein in eukaryote

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



Expression profiles in different tissues

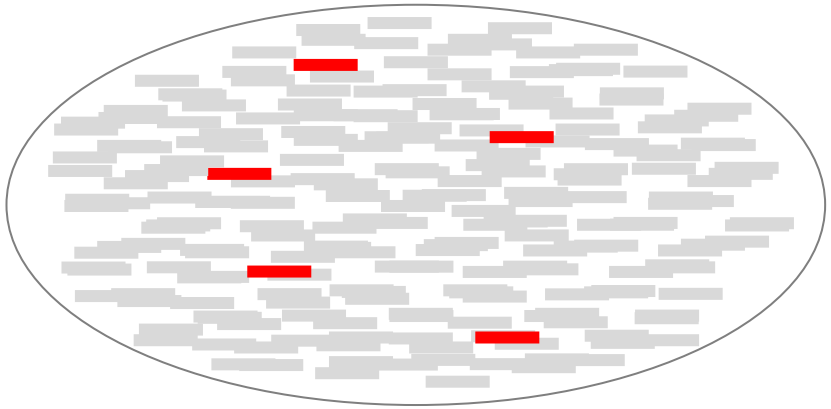


Adaptation to environmental change

What is the expression level of a transcript?

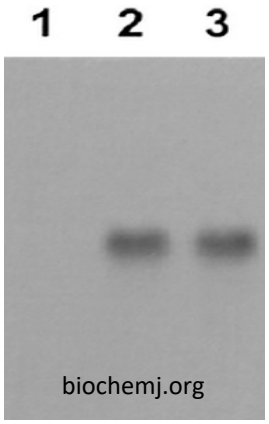
Response to biotic stress

Approaches for quantification of gene expression

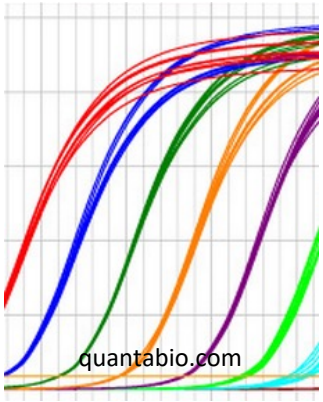


How can we measure the accumulative level of transcripts of **a given gene** in millions/billions of transcripts?

Northern blot



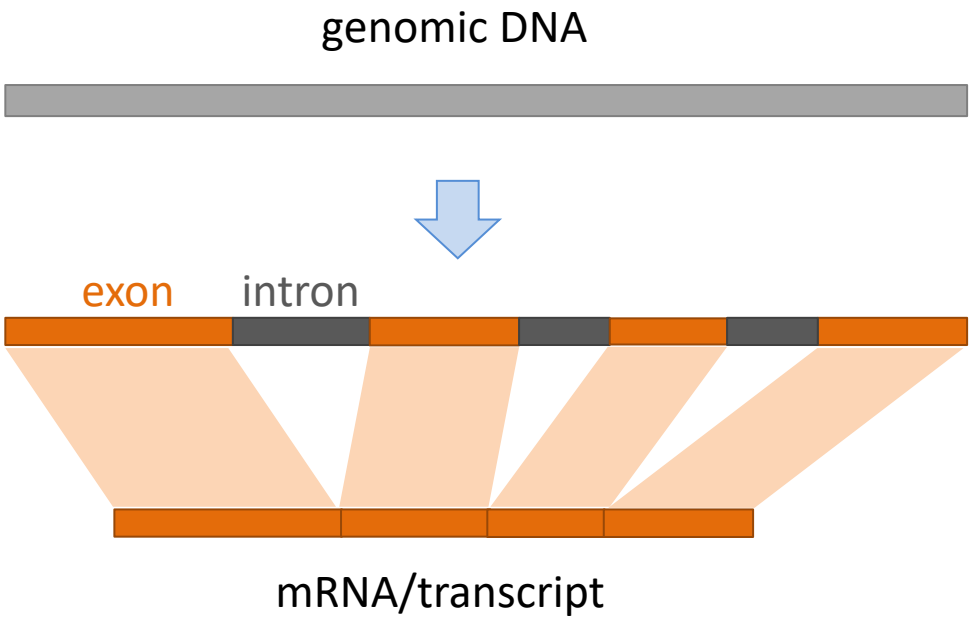
qRT-PCR



microarray

RNA-seq

Rationale of RNA-seq (mRNA sequencing)



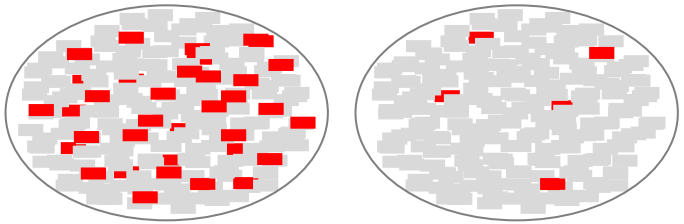
Essentially, RNA-seq is designed to measure mRNA accumulation levels of genes by

1) recognizing transcripts based on sequences

2) and quantifying transcripts of each gene

10 millions of transcripts in each sample

Including transcripts from a gene of interest



100 5

sequence 1,000 transcripts

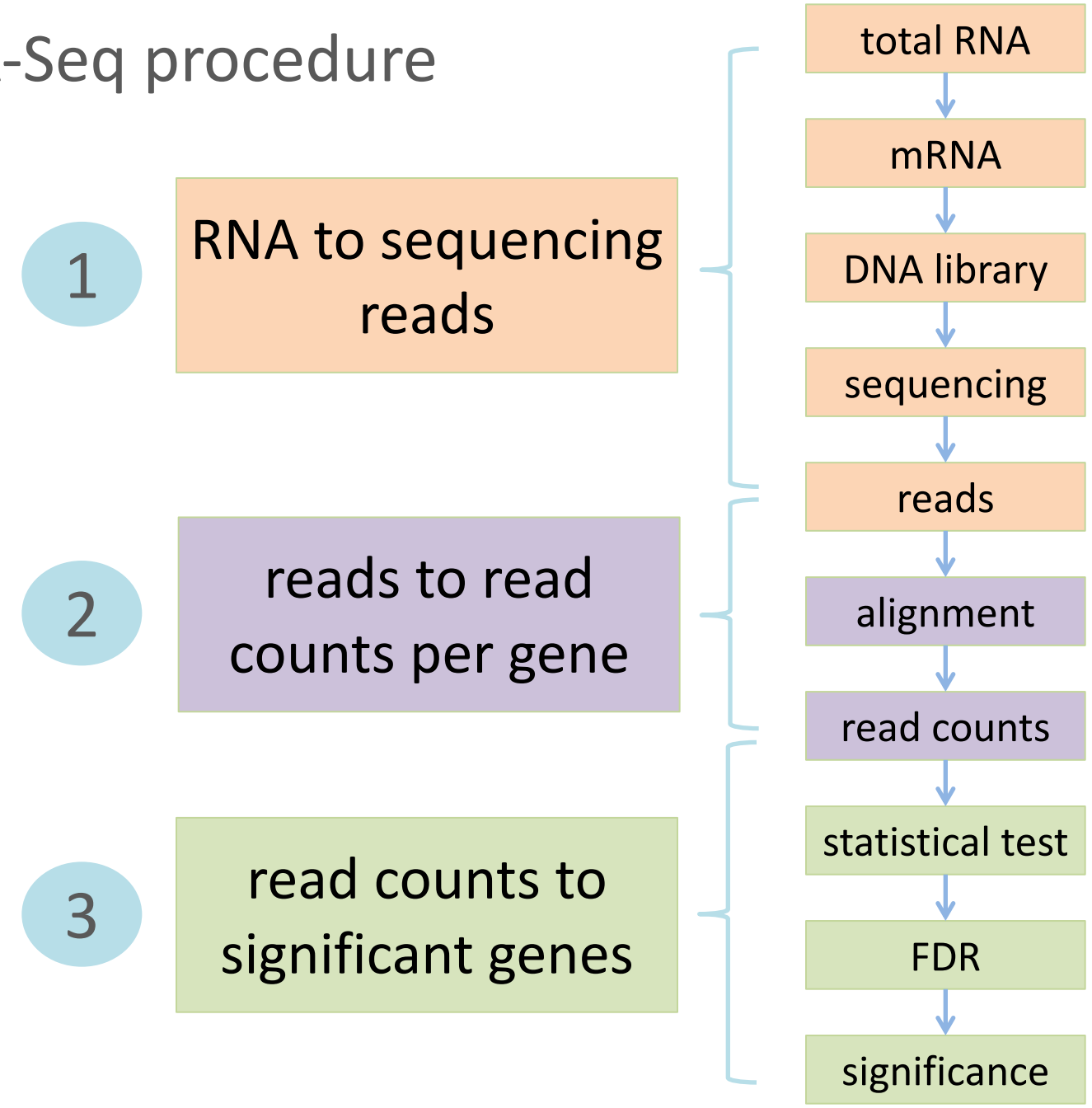
0 0

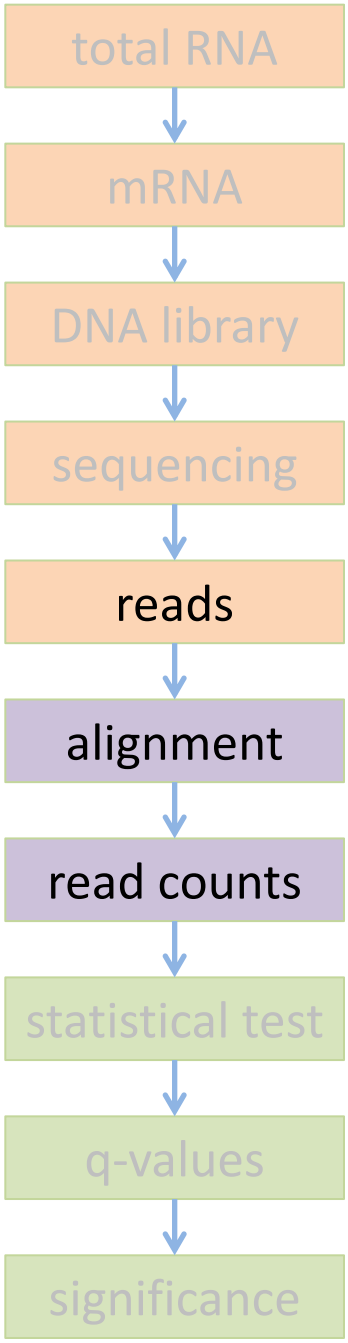
sequence **1 million transcripts**

10 1

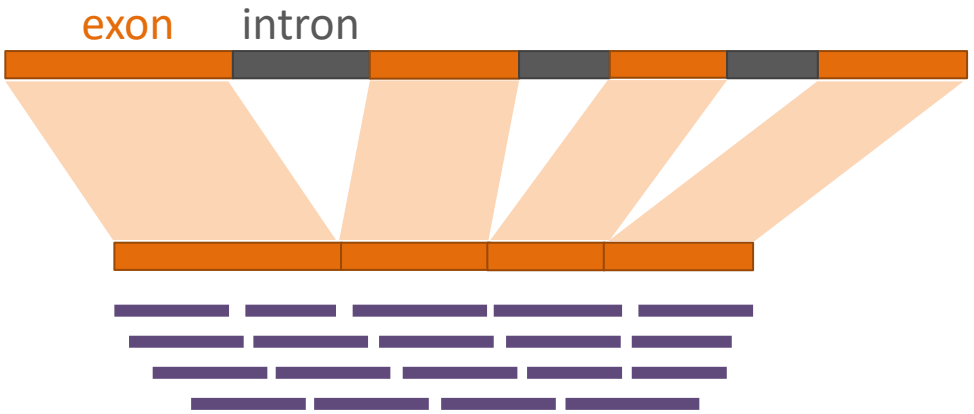
Differential expression (DE)?

RNA-Seq procedure



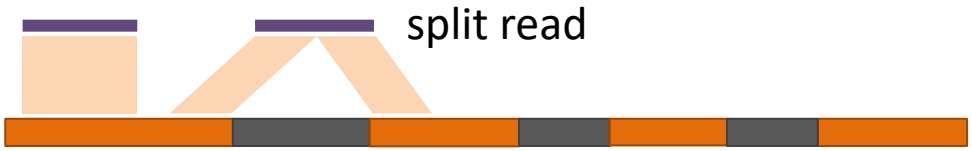


Reads to read counts per gene



1. reads

2. alignment to the reference genome (DNA sequence)



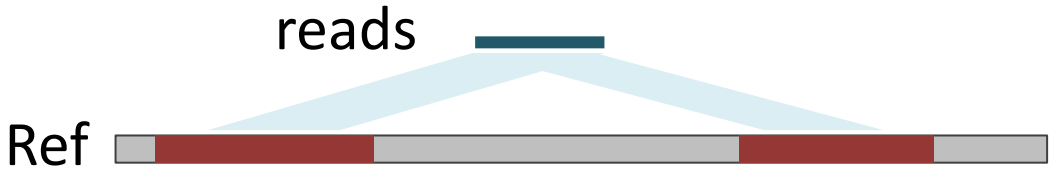
An **intron-aware** aligner is important for RNA-seq reads alignment e.g., STAR, HiSAT2

3. read counts

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

Alignment issues

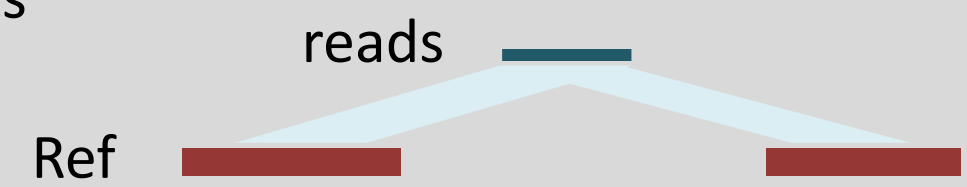
- Repeats



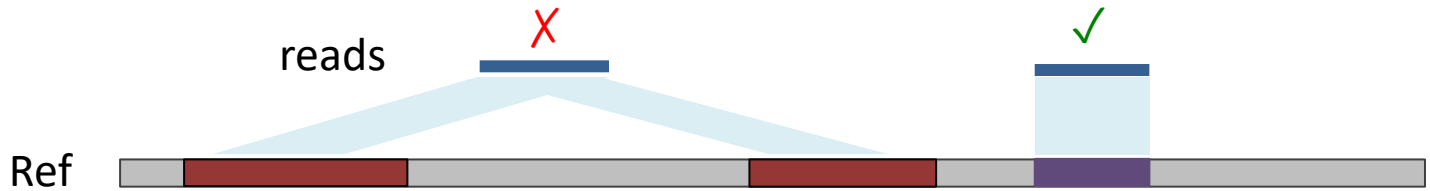
- Sequencing errors
- Polymorphisms (reference and sequenced individuals)
- Quality of reference genomes (mis-assembly and incomplete genome)

Solutions to mitigate problems - I

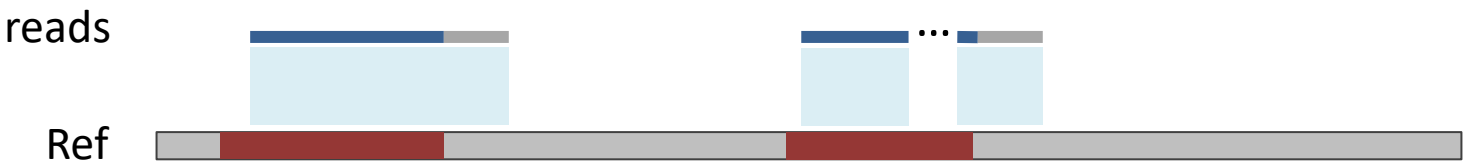
- Repeats



- Unique mapped reads



- Longer reads or Paired-end reads



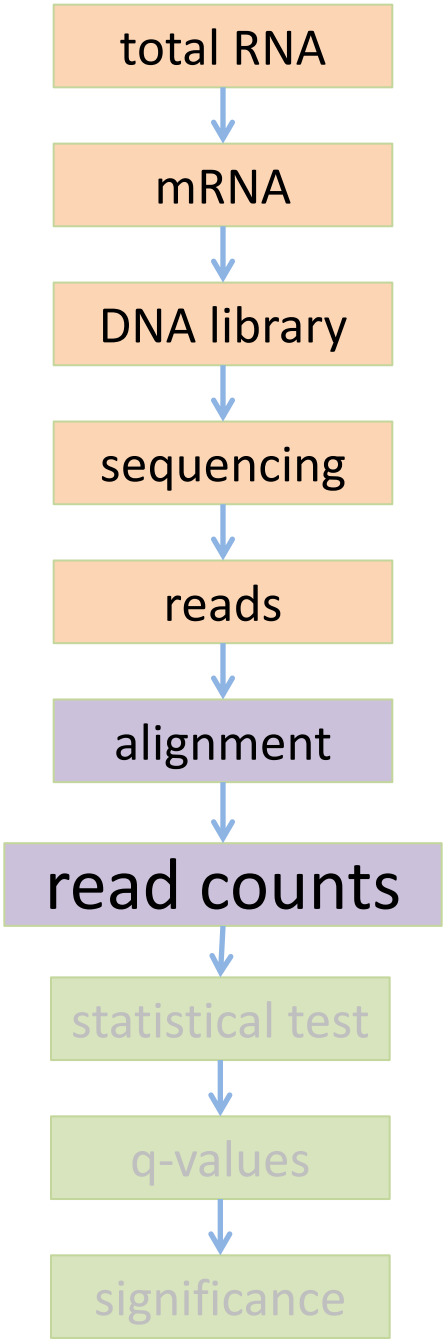
Solutions to mitigate problems

- Sequencing errors
- Polymorphisms (reference and sequenced individuals)

- Tolerance of mismatches or gaps for each alignment

- Quality of reference genomes (mis-assembly and incomplete genome)

- Better reference genome



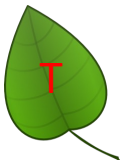
Count matrix
Read counts (Raw) per gene

Gene	sample 1	sample 2
gene 1	6,075	5,934
gene 2	295	377
...	...	...

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

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

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
  - **Overdispersion issue**: Using **negative binomial GLM** to incorporate a dispersion parameter 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

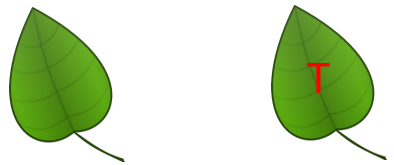
Genome Biology

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*]

Source of variance in counts



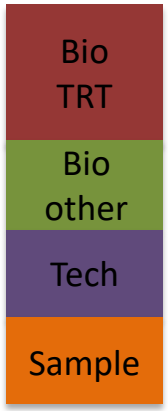
Control 1 TRT 1 (salt)

Gene 1	**C1**	**T1**
…		
Sum	N1	N2

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

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

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



- Sequencing depth

Sampling variance

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

total molecules: 10^9

gene X: 1000 molecules

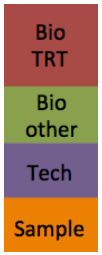
Randomly sample 10^7

First sampling	6
Second sampling	13
Third sampling	8

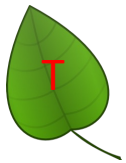
Randomly sample 10^8

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

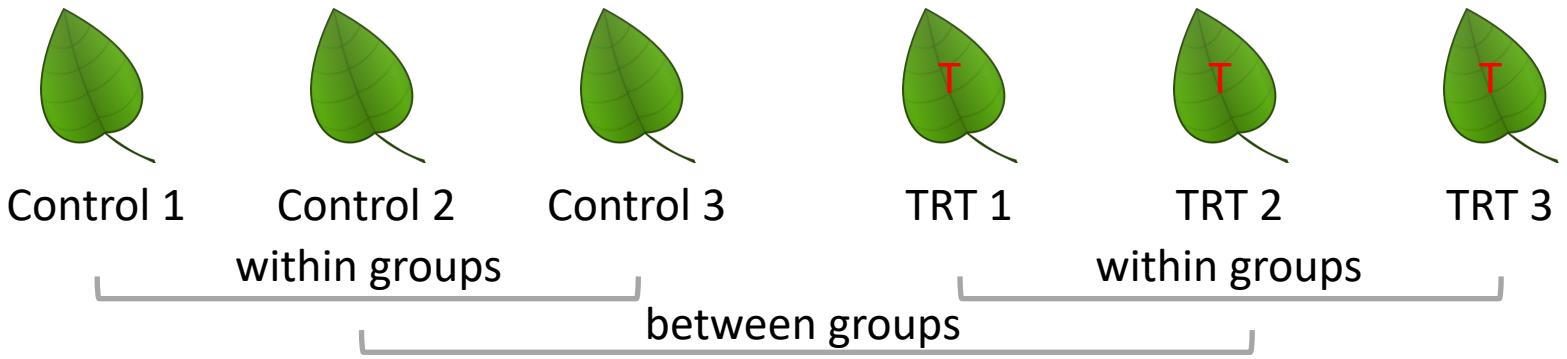
within groups

between groups

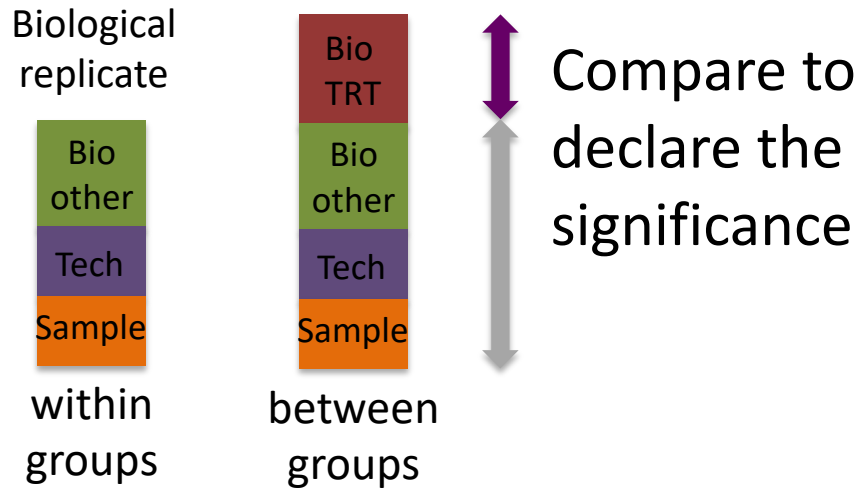
Compare to declare the significance

Spurious 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 effect

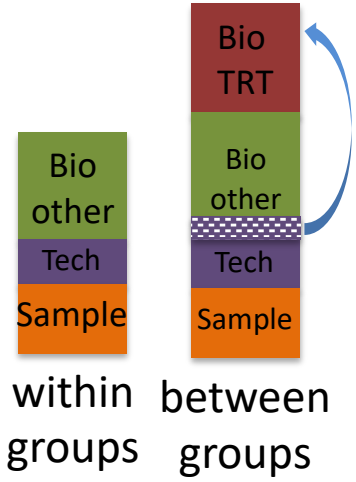
Question

Goal: to identify the DEs between two biological groups

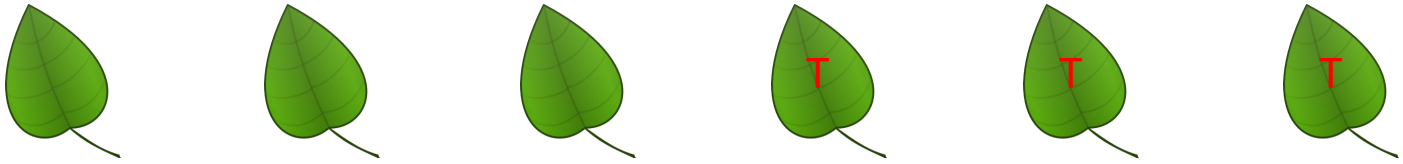
Design: Each group has five biological replicates

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

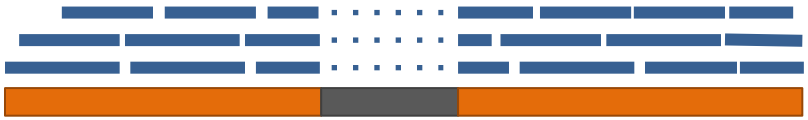
Control 1 read count = **28**

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

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

RPKM of X = ? = **3.7**

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

Experimental Design

- **Sequencing depth**

Increasing sequencing depth decreases sampling variance relative to the mean

- **Biological replication**

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

- **Randomization and unbiasedness**

Try to avoid confounding effect

Outline

- RNA-seq procedure
- Experimental design
- Multiple testing correction
- Data visualization
- Gene ontology (GO) enrichment

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)

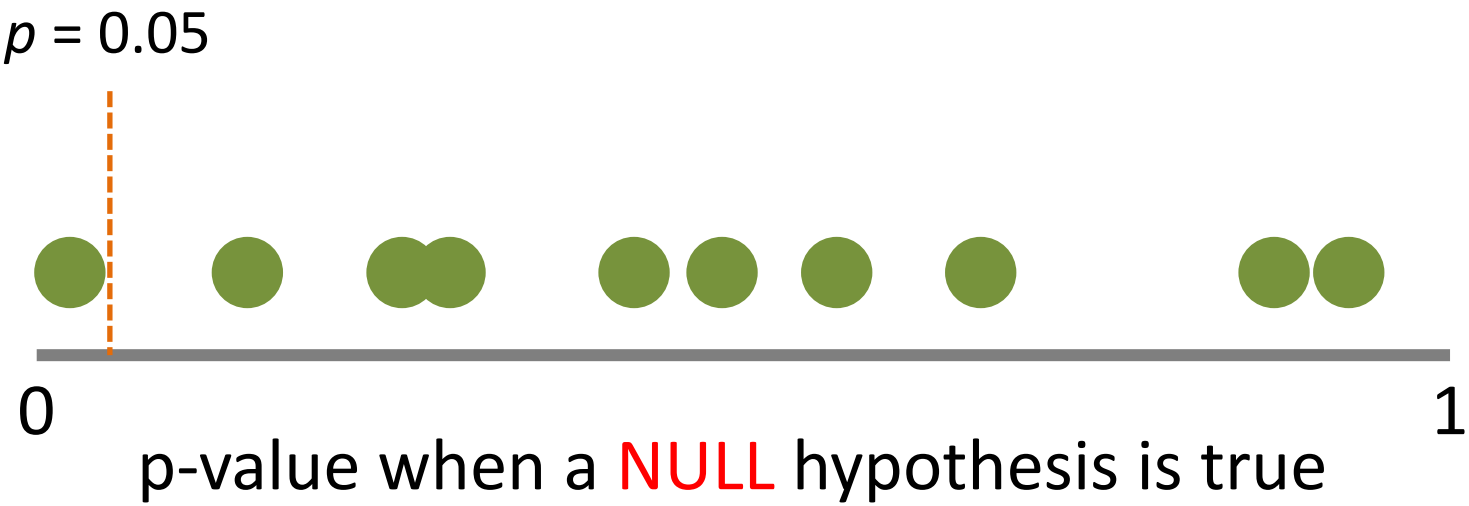
Single statistical test

H0: the null hypothesis

Single statistical test

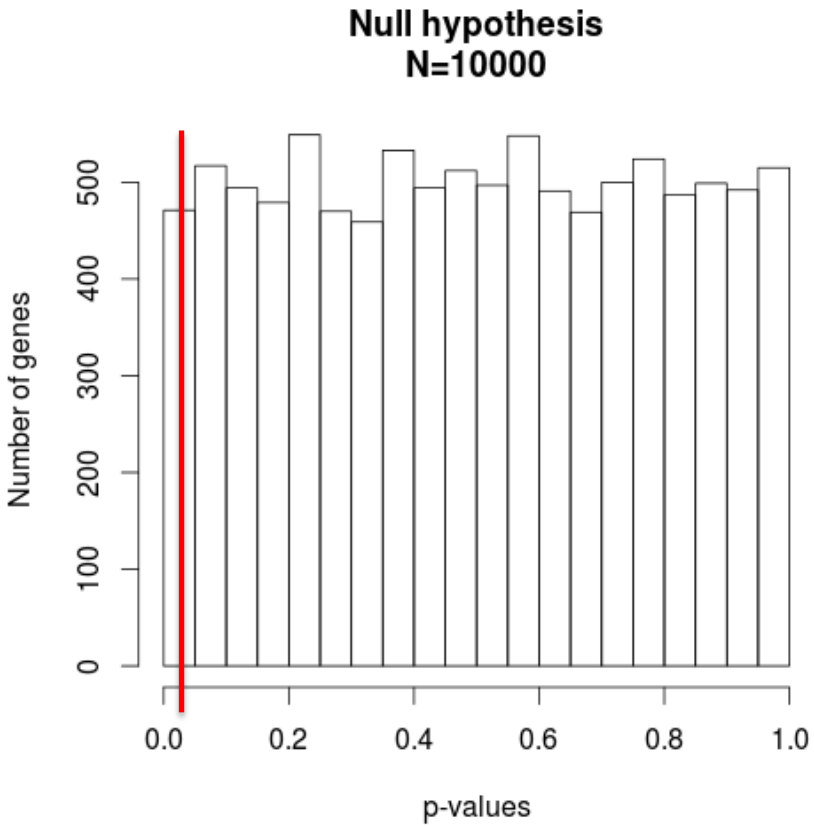
H0: the null hypothesis

Multiple testing correction



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

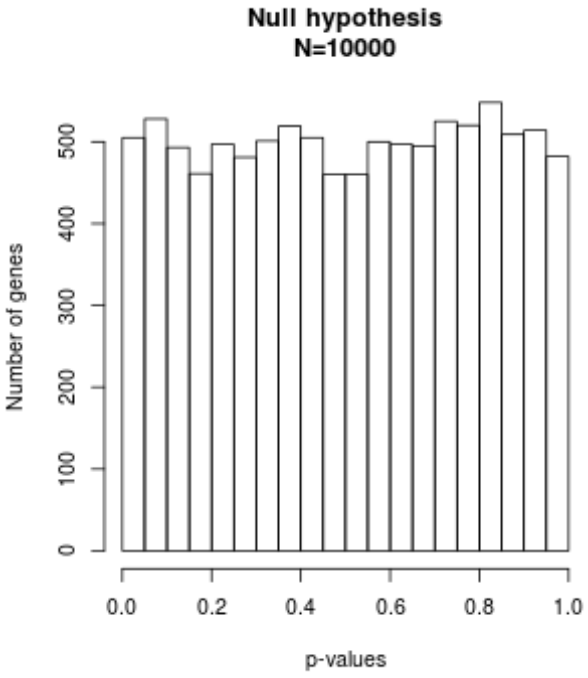
P-value distribution under the null hypothesis (e.g., no treatment effect)

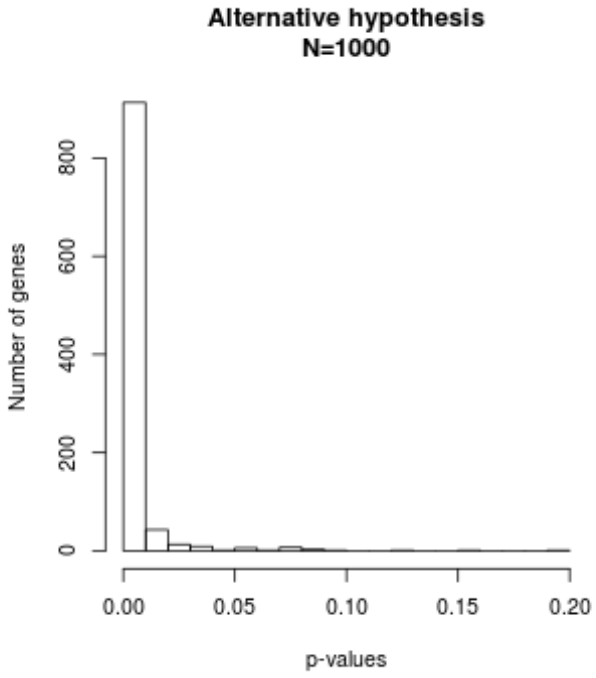


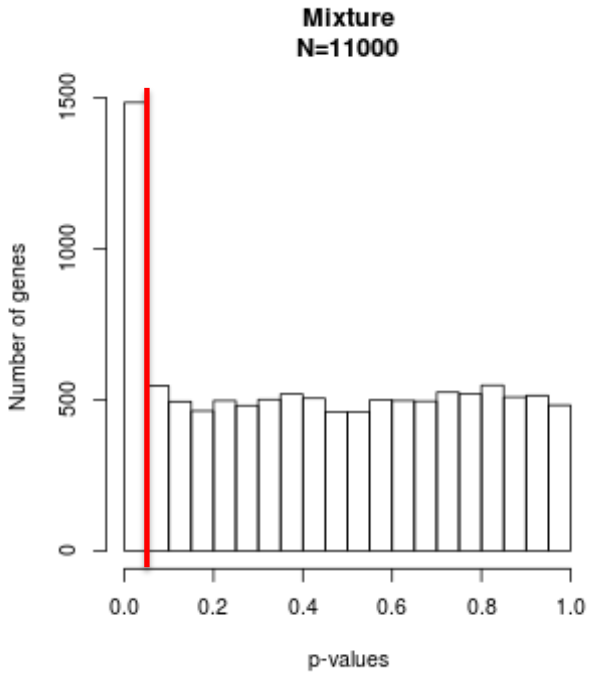
No matter how stringent the criteria are, you'll identify genes with very small p-values and the false discovery rate (FDR) is 100%.

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

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







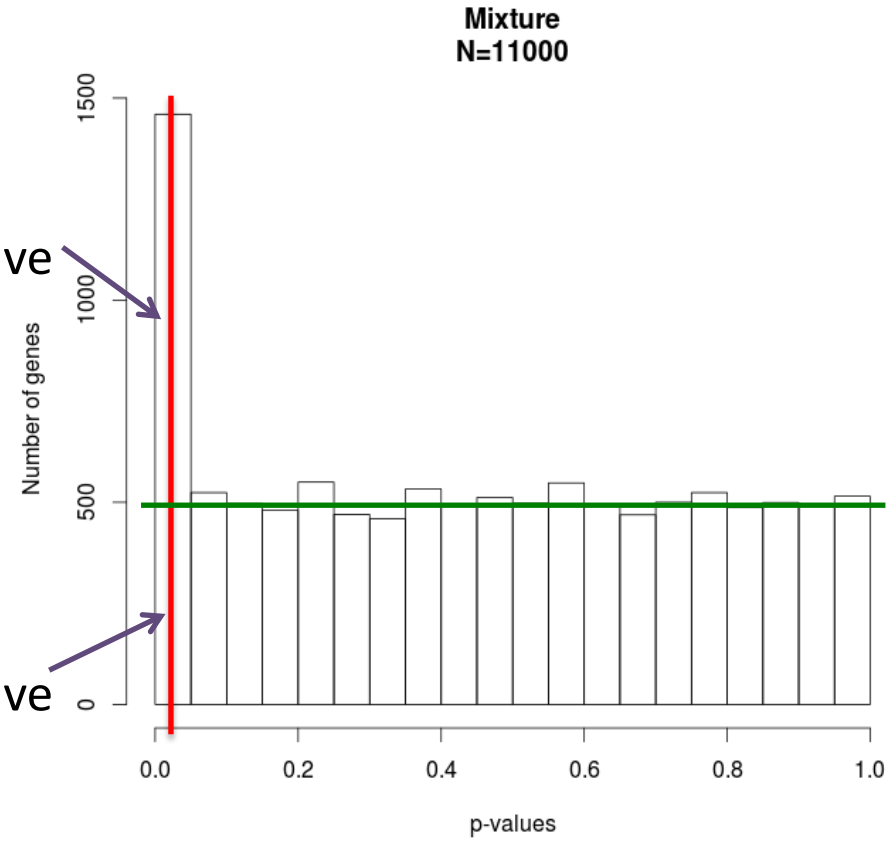
When the null hypothesis is true, the p-value is distributed uniformly.

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



True positive

False positive

P-values < 0.00009

DE=992

False DE=99

FDR 10%

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

q-values (adjusted p-values)

The **q-value** 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 (False Discovery Rate) 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

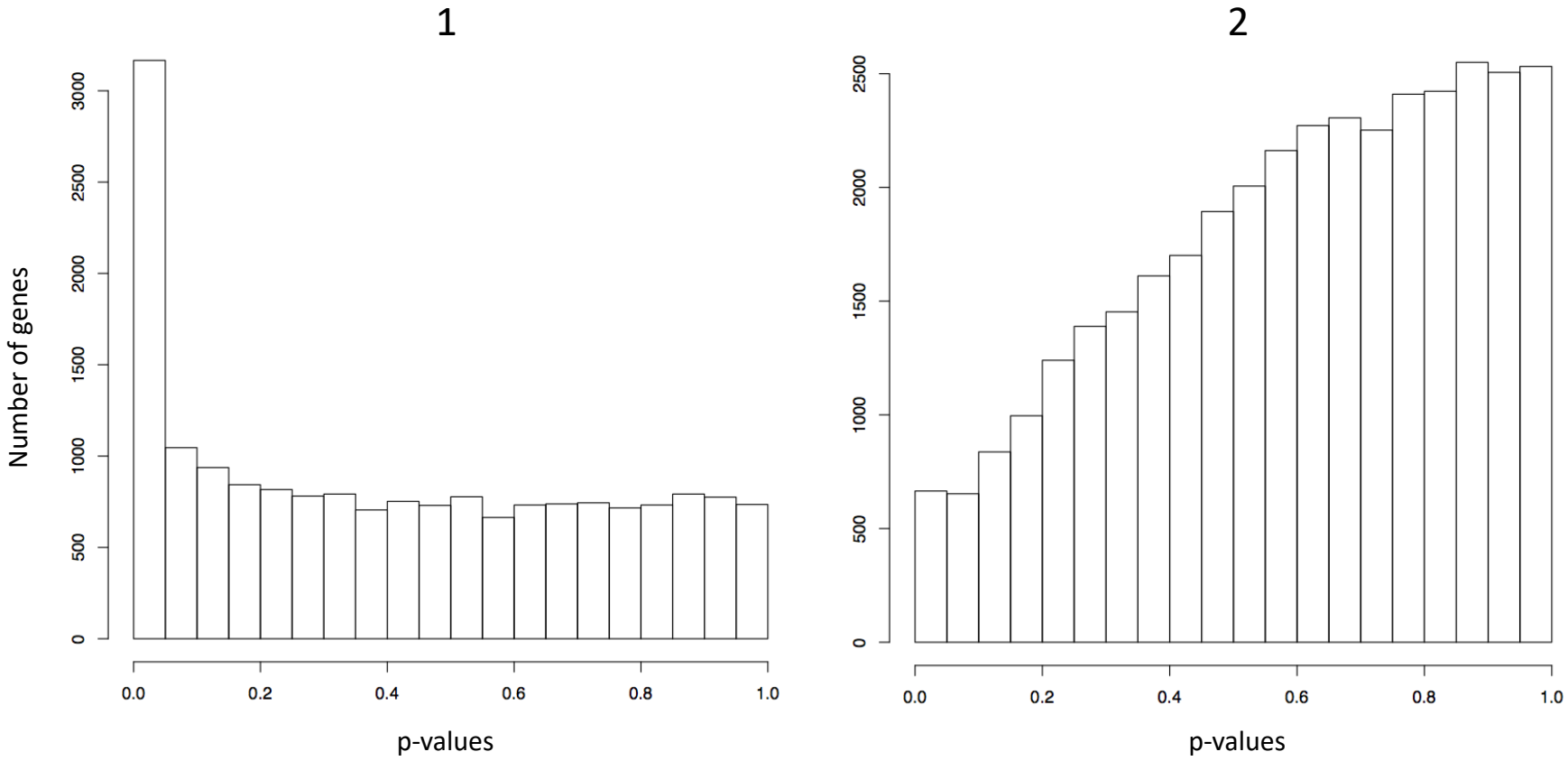
Question

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.

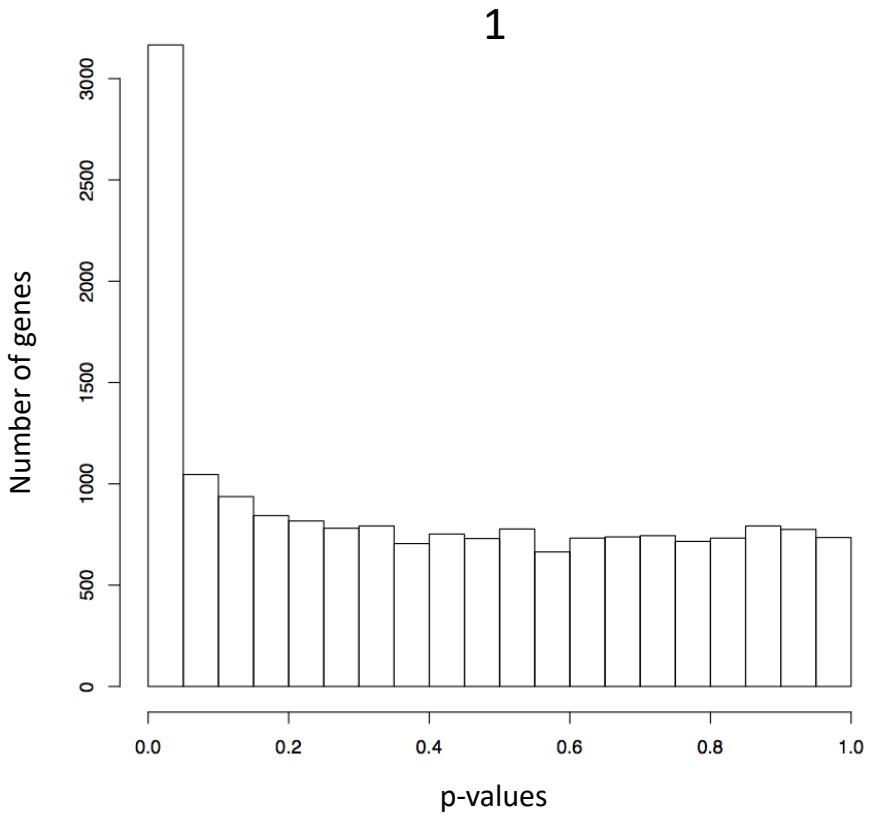
2. The 5% genes (25 genes) in the set are expected to be false DE genes.

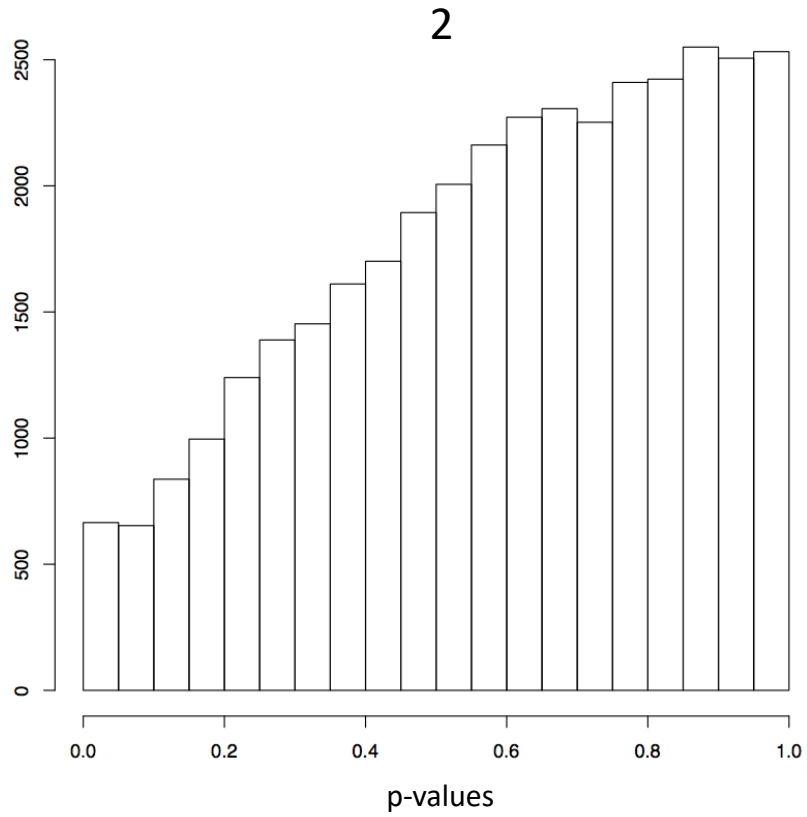
P-value histograms from real studies



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

P-value histograms from real studies

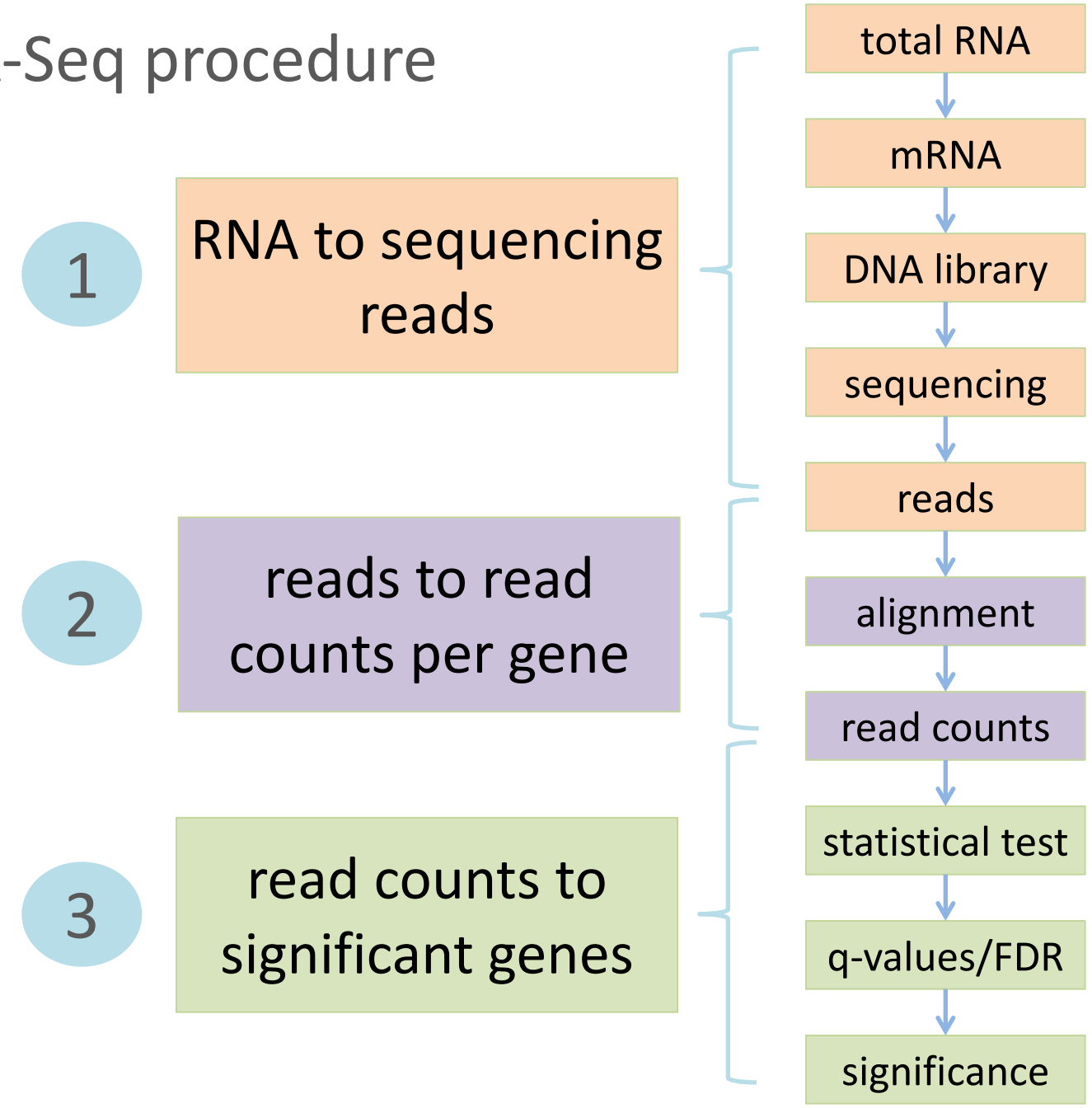




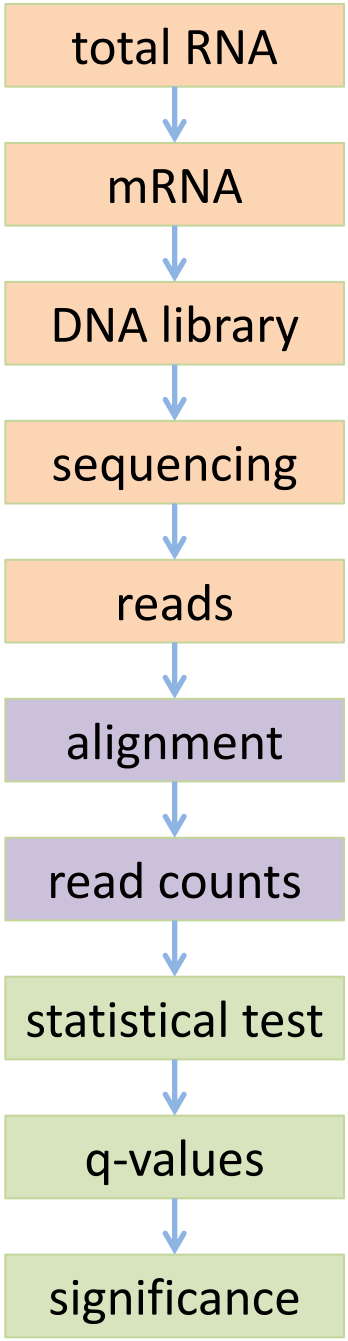
DE = 1,370, FDR=5%

DE = 0, FDR=20%

RNA-Seq procedure



Keywords



randomization, replication, RNA quality

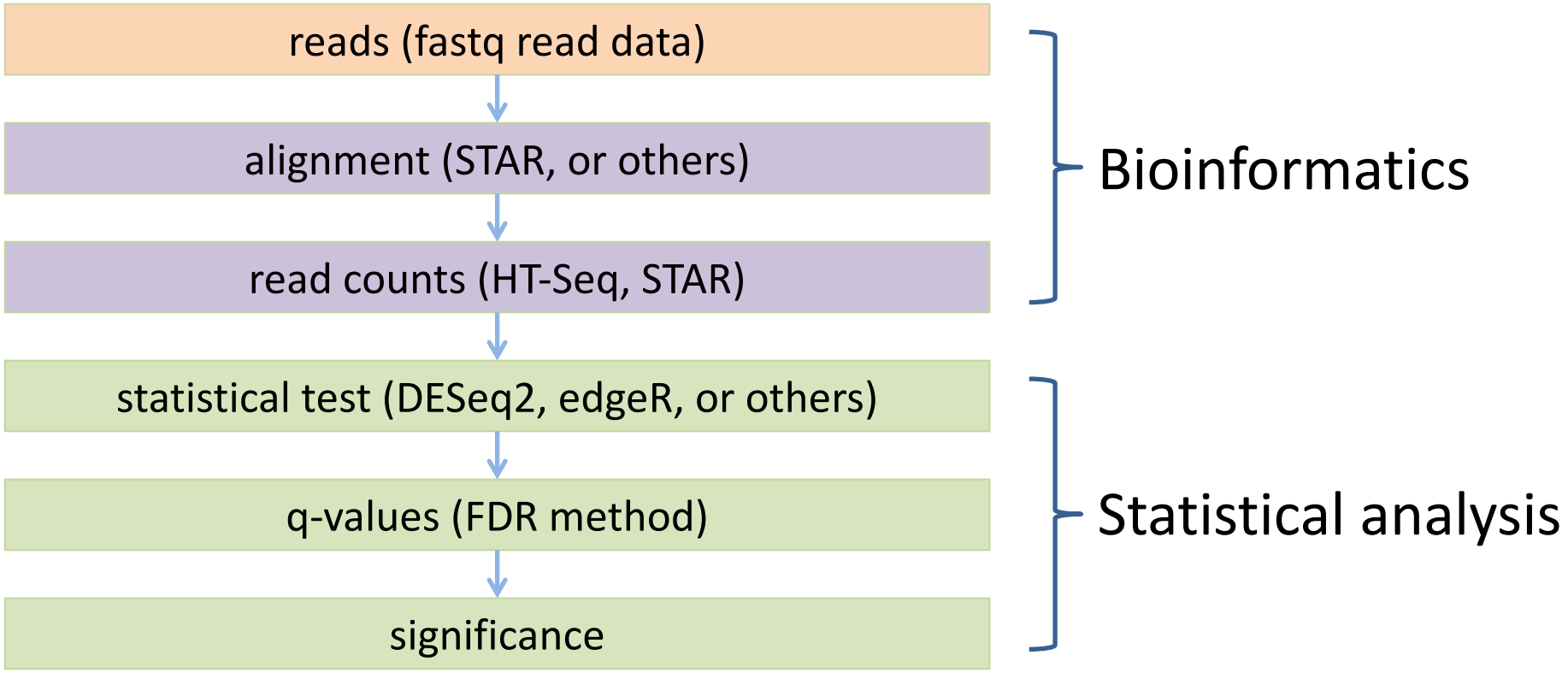
short or long reads
single- or paired-end reads, read length
sequencing depths
(e.g., >20 million short reads for most experiments)

intron-spanning aligners
(e.g., STAR, HiSAT2)

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

5-minute break

Bioinformatics and Statistics (Illumina data)



STAR pipeline – from reads to counts

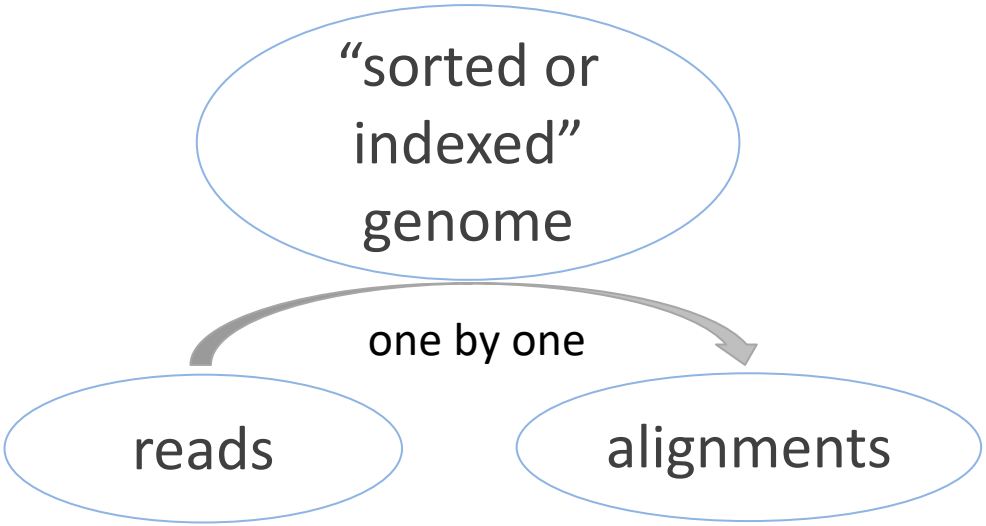
Required files:

1. Reference genome (fasta file)
2. Gene information (gff or gtf gene annotation)
3. Reads (fastq files) – your own data

Many reference genomes and gff/gtf files are available at:
http://ensembl.org/info/data/ftp

Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets
Human *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3
Mouse *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3
Zebrafish *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3

Reads to counts - **reference indexing**




```
STAR --runMode genomeGenerate \
     --genomeDir . \
     --genomeFastaFiles reference.fas \
     --sjdbGTFfile genes.gtf \
     --runThreadN 4
```

Reads to counts – **alignment and read counting**

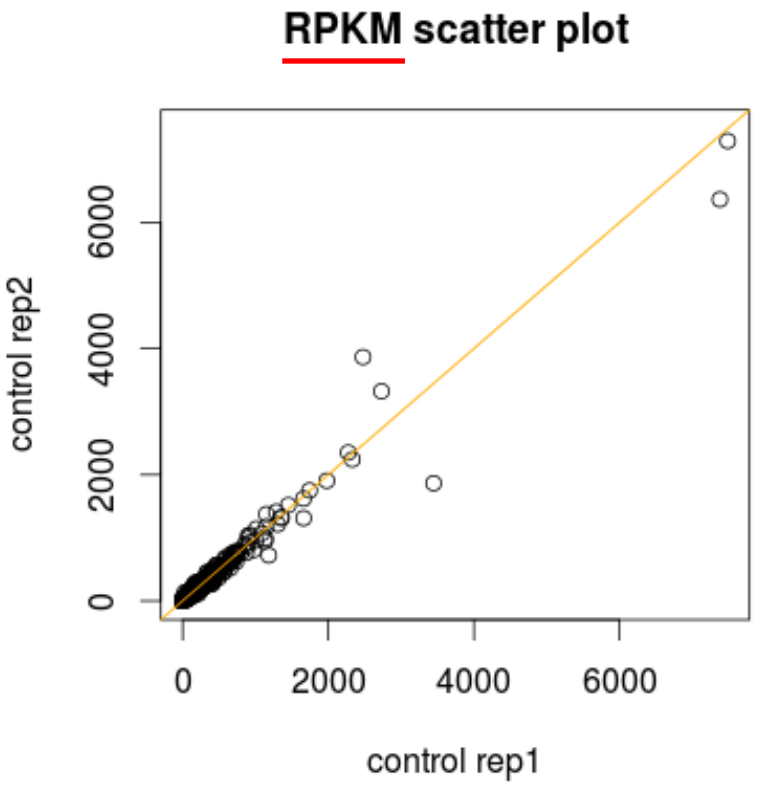
```
STAR --genomeDir reference.fas \
    --readFilesIn read1.fq read2.fq \
    --alignIntronMax 100000 \
    --alignMatesGapMax 100000 \
    --outFileNamePrefix output \
    --outSAMattrIHstart 0 \
    --outSAMmultNmax 1 \
    --outSAMstrandField intronMotif \
    --outFilterIntronMotifs RemoveNoncanonicalUnannotated \
    --outSAMtype BAM SortedByCoordinate \
    --quantMode GeneCounts \
    --outFilterMismatchNmax 5 \
    --outFilterMismatchNoverLmax 0.05 \
    --outFilterMatchNmin 50 \
    --outSJfilterReads Unique \
    --outFilterMultimapNmax 1 \
    --outSAMmapqUnique 60 \
    --outFilterMultimapScoreRange 2
```

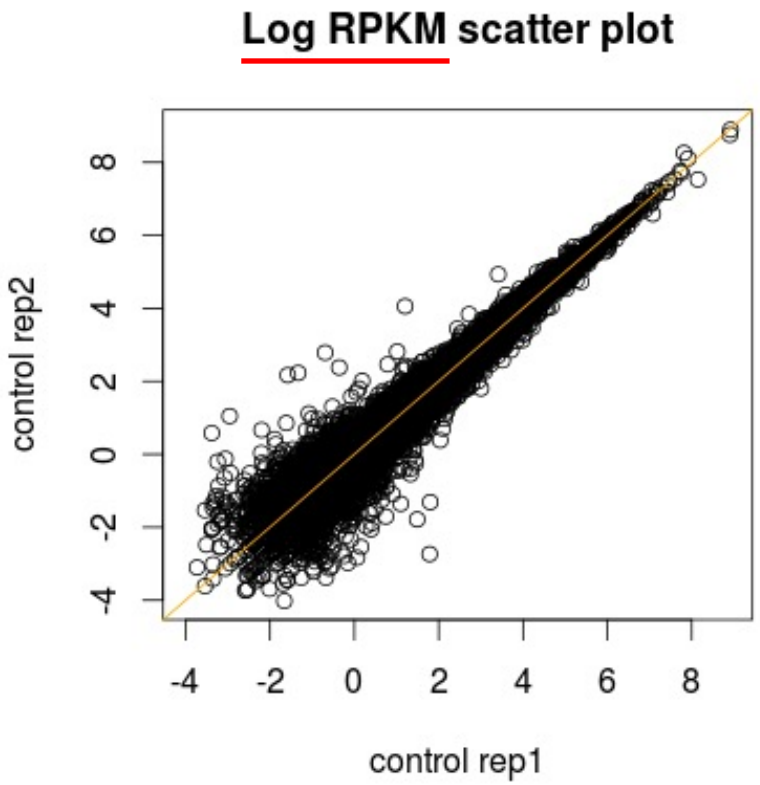
Count matrix: Read counts (Raw) per gene

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

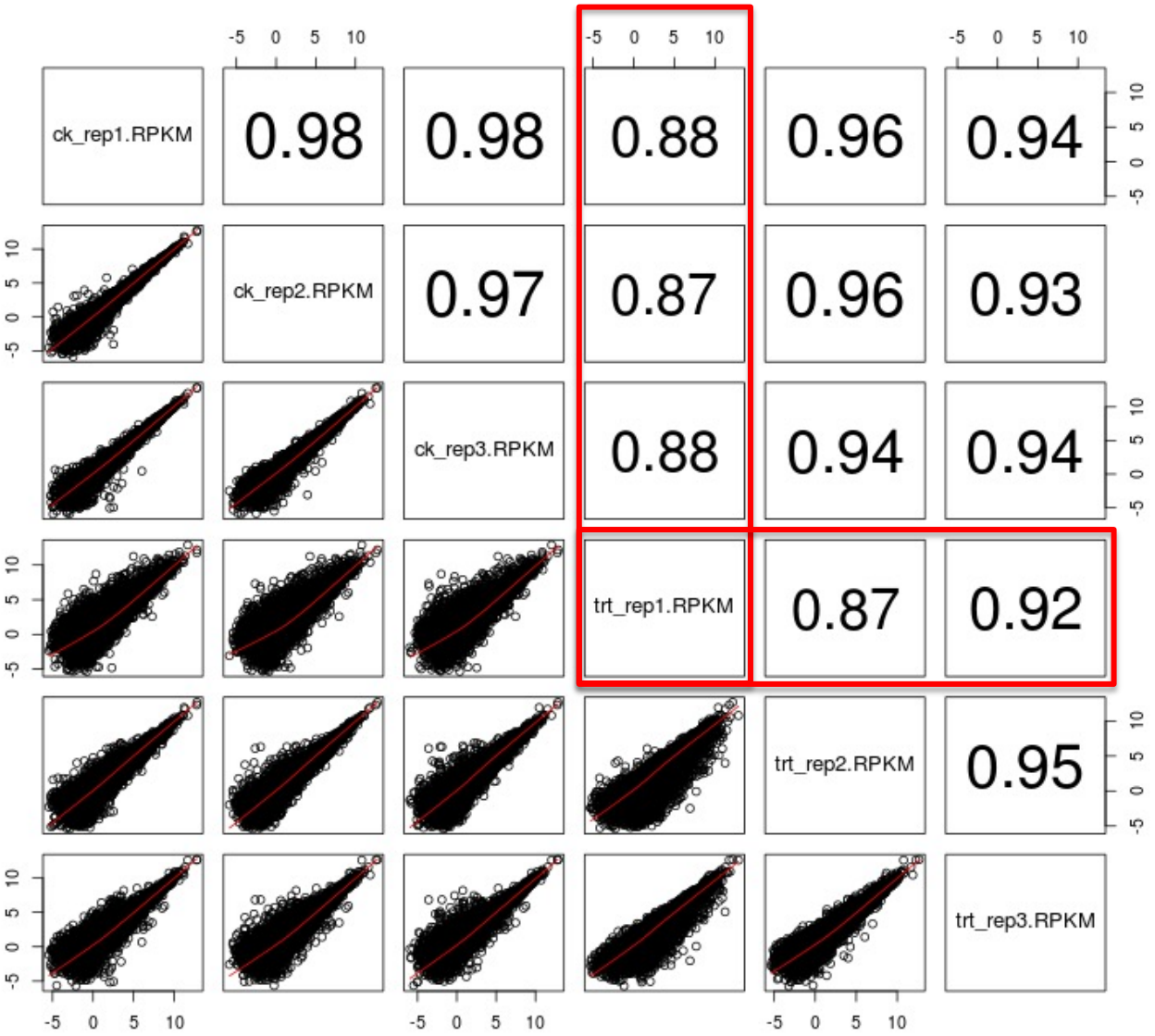
Overall comparisons of read counts among samples

Scatter plot





Pair-wise scatter plot



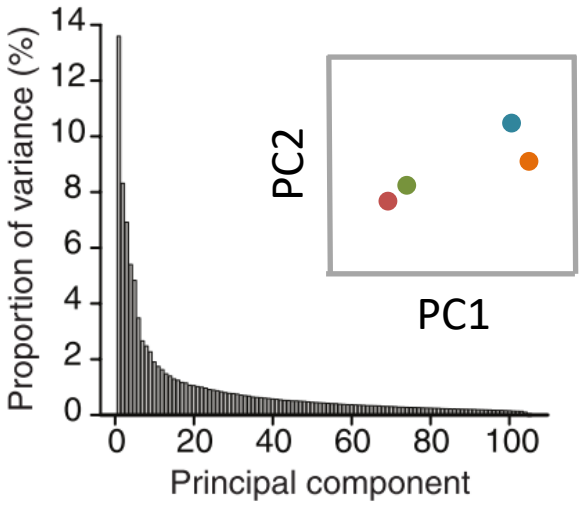
Principal Component Analysis (PCA)

PCA is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set.

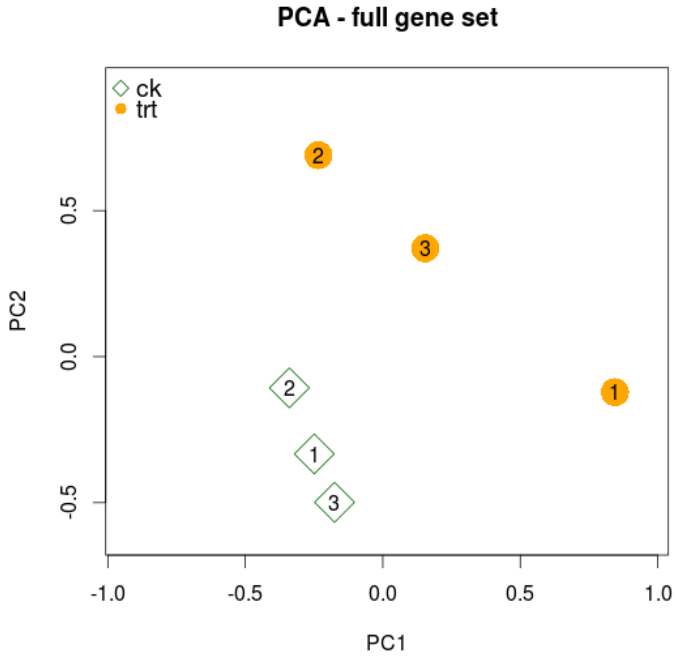
Feature/variable	John	Mike	Jack	Justin
Weight (lb)	150	243	186	128
Height (cm)	171	190	178	175
…				

	Control			Treatment		
GeneID	Rep1	Rep2	Rep3	Rep1	Rep2	Rep3
1	2679	2360	2573	2563	3398	3012
2	177	161	171	154	137	152
3	381	371	397	541	723	635
…						
30000	990	1073	1236	850	672	859

Normalized and standardized data



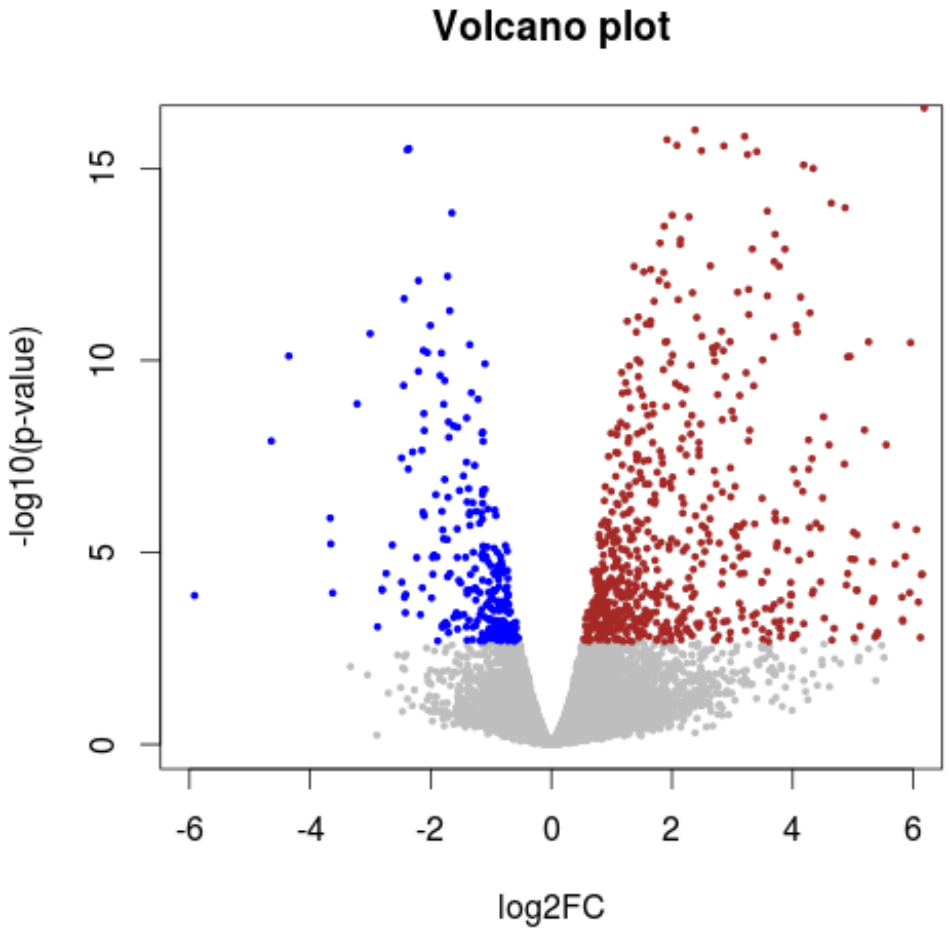
Nature Biotech, 2008, 26:303-4



Overview of differential expression

Volcano plot

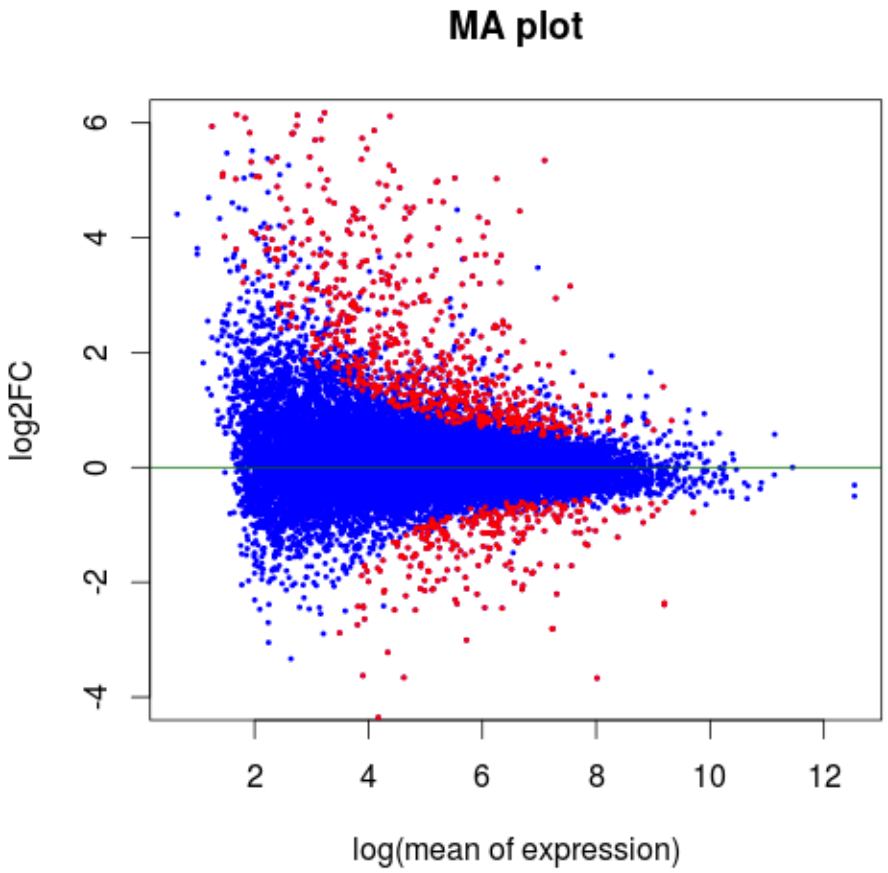
DE Result			
GeneID	Log2FC	p-value	-log10(pvalue)
1	-0.40	0.037	1.43
2	0.03	0.916	0.04
3	-0.89	2.42E-05	4.62
4	0.30	0.130	0.89
5	-0.36	0.140	0.85
6	-0.07	0.811	0.09
…			



MA plot

M (log ratios) and A (mean average)

GeneID	Mean RPKM	log mean	log2FC
1	0.51	-0.29	-0.40
2	1.25	0.10	0.03
3	3.52	0.55	-0.89
4	0.19	-0.72	0.30
5	2.34	0.37	-0.36
6	6.14	0.79	-0.07
…			

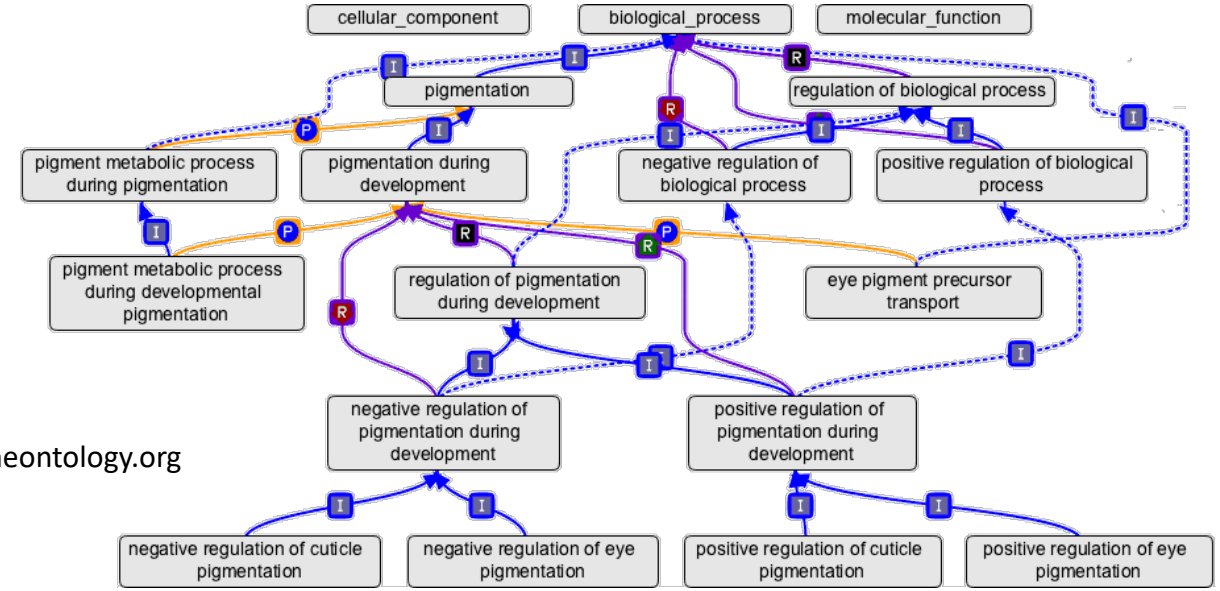


More at: en.wikipedia.org/wiki/MA_plot

Gene ontology enrichment analysis

Gene ontology (GO)

An ontology is a representation of a body of knowledge, within a given domain. Ontologies usually consist of a set of classes or terms with relations that operate between them.



Source: geneontology.org

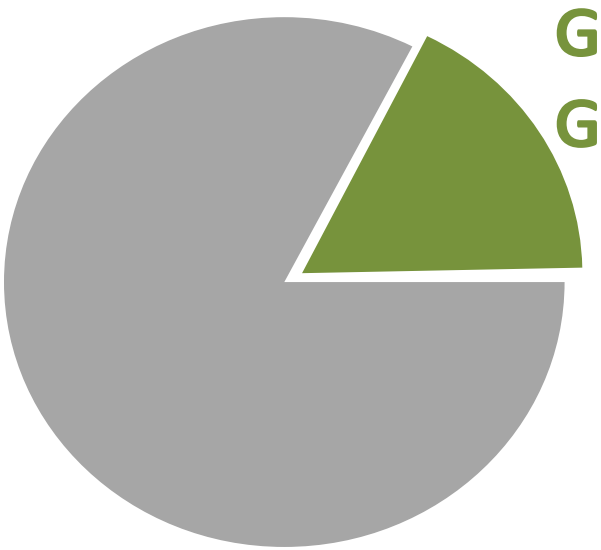
Three domains, three roots

Node: GO term (e.g., cell growth, GO:0016049, biological process)

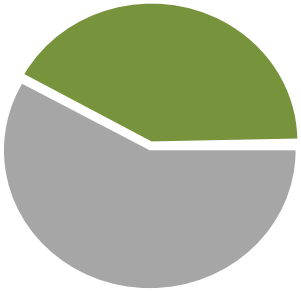
Edge: term-term connection

Each GO term can be traced back to a root

Category enrichment



All genes

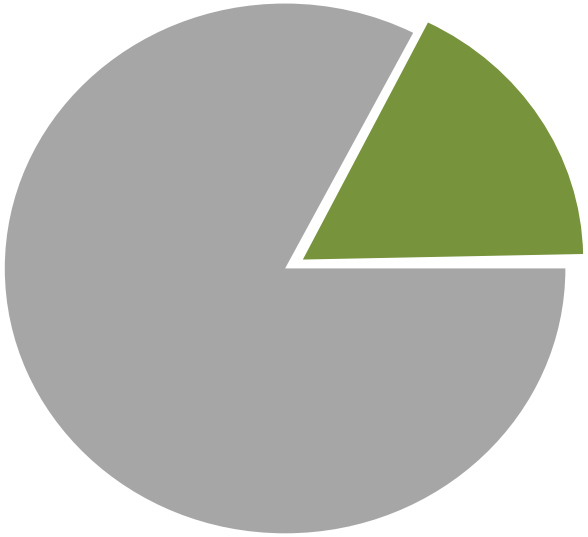


significant
genes

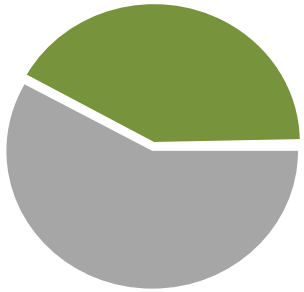
Is **Green GO** enriched in the significant gene set ?

Green GO is a GO ID (e.g., GO:0006519)

dance party



college students
N=10,000



students who are
in a dance party
N=200

Are **graduate students** over-represented (enriched) in the party?

GO enrichment test – Fisher's Exact test

40 significant genes

Gene	GO accession
GRMZM2G001475	GO:0006519
GRMZM2G001475	GO:0016831
GRMZM2G001500	GO:0005524
GRMZM2G001500	GO:0006457
GRMZM2G001500	GO:0051082
GRMZM2G001508	GO:0003993
GRMZM2G001514	GO:0003677
GRMZM2G001514	GO:0004879
GRMZM2G001514	GO:0005634
GRMZM2G001514	GO:0006355
…	…

GRMZM2G000012	1
GRMZM2G002342	2
GRMZM2G006480	3
…	…
GRMZM5G038156	40

GO:0006519

Gene	Significant?
GRMZM2G001475	no
GRMZM2G002652	no
GRMZM2G006480	**yes**
…	…
GRMZM5G868038	no

Question: Are the genes of this GO term enriched in the significant gene set?

Assumption: all genes are independent and equally likely to be selected as DEs.

2x2 Table for GO:0006519

	GO:0006519	Others
Significant	5	35
Not significant	210	39416

Fisher's Exact Test:
p-value = 2.5e-06

Name	cellular amino acid metabolic process
Ontology	Biological Process
Definition	The chemical reactions and pathways involving amino acids, carboxylic acids containing one or more amino groups, as carried out by individual cells.

GO enrichment test – Fisher's Exact test

Gene	GO accession
GRMZM2G001475	GO:0006519
GRMZM2G001475	GO:0016831
GRMZM2G001500	GO:0005524
GRMZM2G001500	GO:0006457
GRMZM2G001500	GO:0051082
GRMZM2G001508	GO:0003993
GRMZM2G001514	GO:0003677
GRMZM2G001514	GO:0004879
GRMZM2G001514	GO:0005634
GRMZM2G001514	GO:0006355
…	…

40 DE genes

GRMZM2G000012	1
GRMZM2G002342	2
GRMZM2G006480	3
…	…
GRMZM5G038156	40

GO:0006519

Gene	Significant?
GRMZM2G001475	no
GRMZM2G002652	no
GRMZM2G006480	**yes**
…	…
GRMZM5G868038	no

5 DE and 210 non-DE

Question: Are the genes of this GO term enriched in the significant gene set?

Randomly sample **40 DE**

GO:0006519

Gene	R
GRMZM2G001475	no
GRMZM2G002652	**yes**
GRMZM2G006480	no
…	…
GRMZM5G868038	no

1 DE and 214 non-DE

Randomly sample **40 DE**

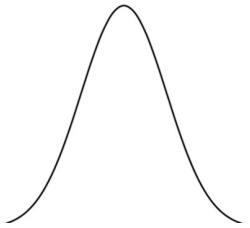
GO:0006519

GRMZM2G001475	no
GRMZM2G002652	no
GRMZM2G006480	no
…	…
GRMZM5G868038	no
GRMZM2G001475	no

0 DE and 215 non-DE

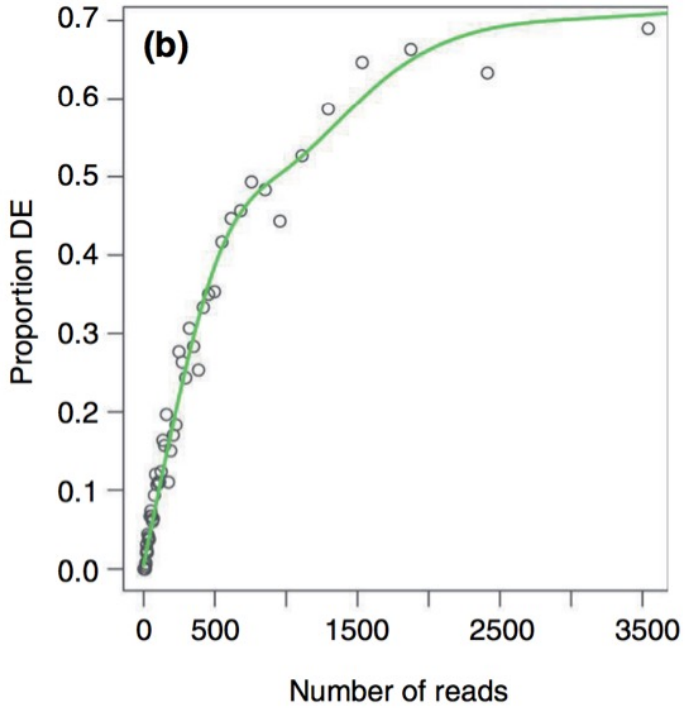
…

GO:0006519	# DE
1st sampling	1
2nd sampling	0
3rd sampling	2
…	…



GOSeq

Not all genes are equally likely to be selected as DEs.



Young MD, et al., (2010). Genome Biology, 11: R14.

1. The likelihood of DE as a function of number of reads is quantified through fitting a monotonic function to "proportion of DE" versus "number of reads".

2. The function is incorporated into the enrichment statistical test

Gene	Read counts	Proportion
GRMZM2G001475	224	0.16
GRMZM2G002652	51	0.05
GRMZM2G006480	**536**	**0.38**
…	…	…
GRMZM5G868038	0	0

3. Weighted sampling to perform enrichment test

GO:0006519	# DE
1st weighted sampling	1
2nd weighted sampling	0
3rd weighted sampling	2
…	…

→ p-value

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 **parameters** in each method need to be carefully selected.

REFERENCES

1. Benjamini Y, et al. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57:289-300.
2. Conesa A, et al. 2016. A survey of best practices for RNA-seq data analysis. Genome Biol 17:13.
3. Love MI, et al. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.
4. Robinson MD, et al. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140.