

Design of RNA-Seq and Result Interpretation

Sanzhen Liu

Department of Plant Pathology
Kansas State University

@K-State IGF RNA-Seq Workshop

6/22/2017

Outline

Review of RNA-Seq procedure

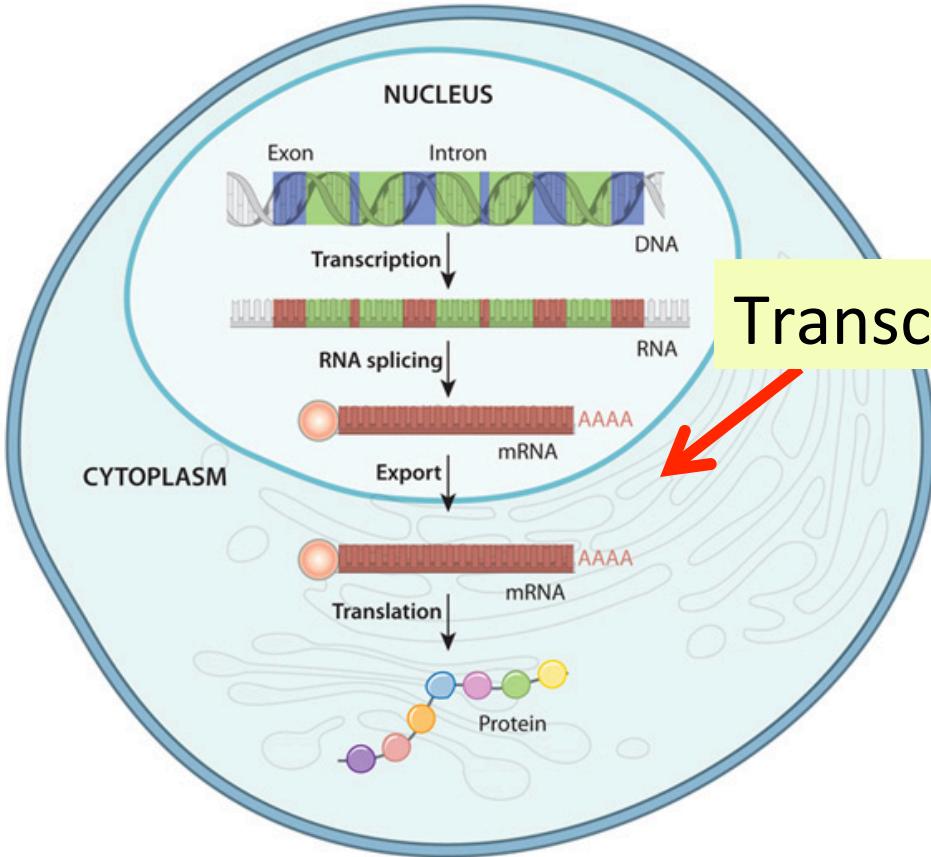
Design of DE experiments and results

- Experimental design
- Multiple test correction

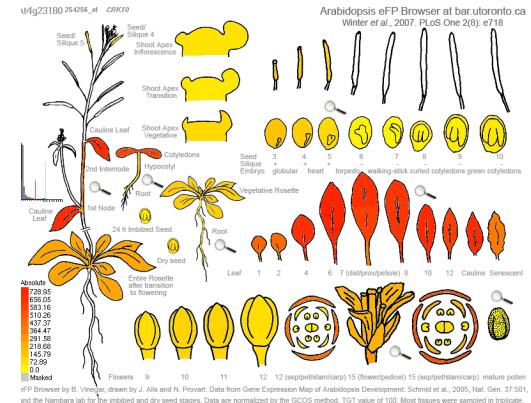
Other analyses

- Visualization
- GO term enrichment analysis

Gene expression



Transcripts



Expression profiles in different tissues



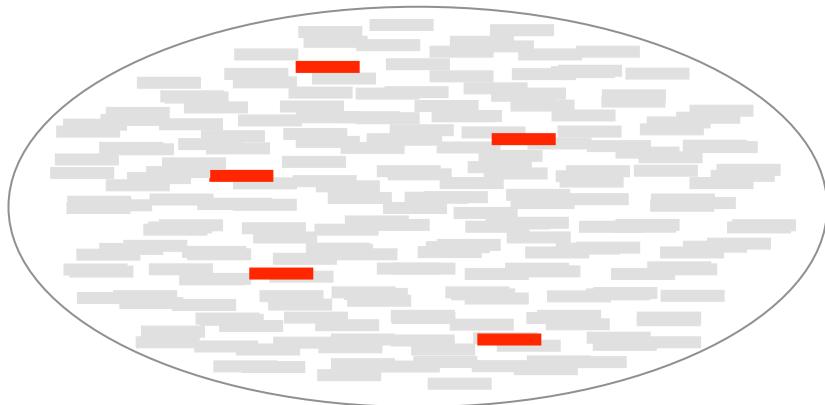
cragenomica.es

DNA to protein in eukaryote

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

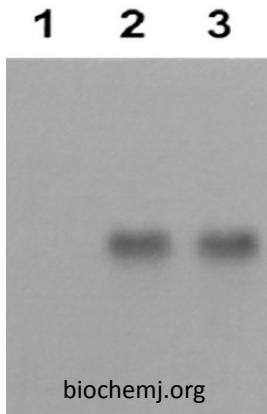
Response to biotic stress

Approaches for quantification of gene expression

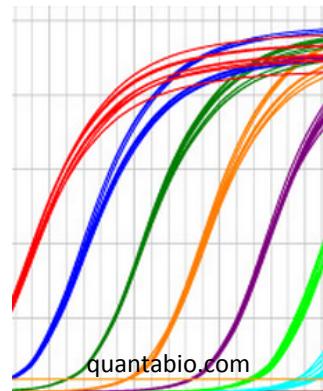


How can we measure the accumulative level of transcripts of **a given gene** in millions/billions of transcripts?

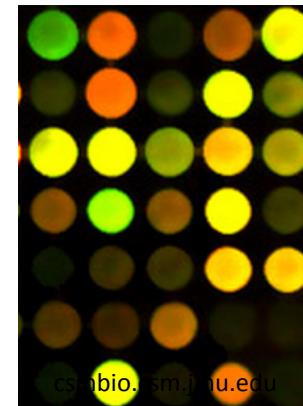
Northern blot



qRT-PCR

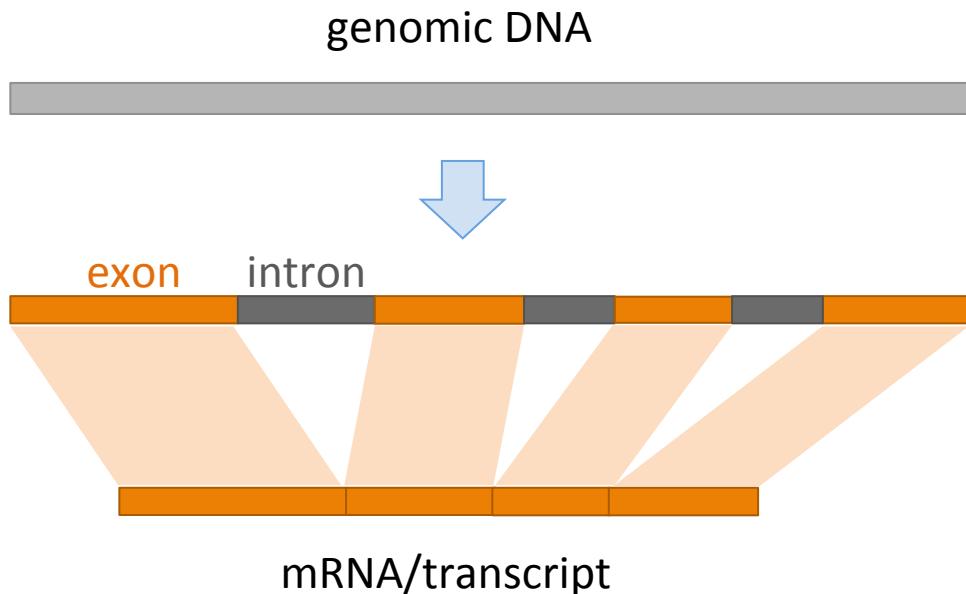


microarray



RNA-Seq

Rationale of RNA-Seq for differential expression (DE)

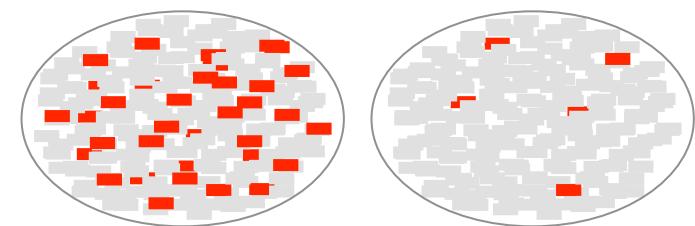


Essentially, RNA-Seq is designed to measure mRNA accumulation levels of genes by

- 1) recognizing mRNA based on sequences**
- 2) quantifying mRNA of each gene**

Millions times of sampling to quantify each component (transcript) in tissue samples.

10 millions of transcripts in each



100 **gene of interest** 5

sequence **1,000** transcripts

0 0

sequence **1 million** transcripts

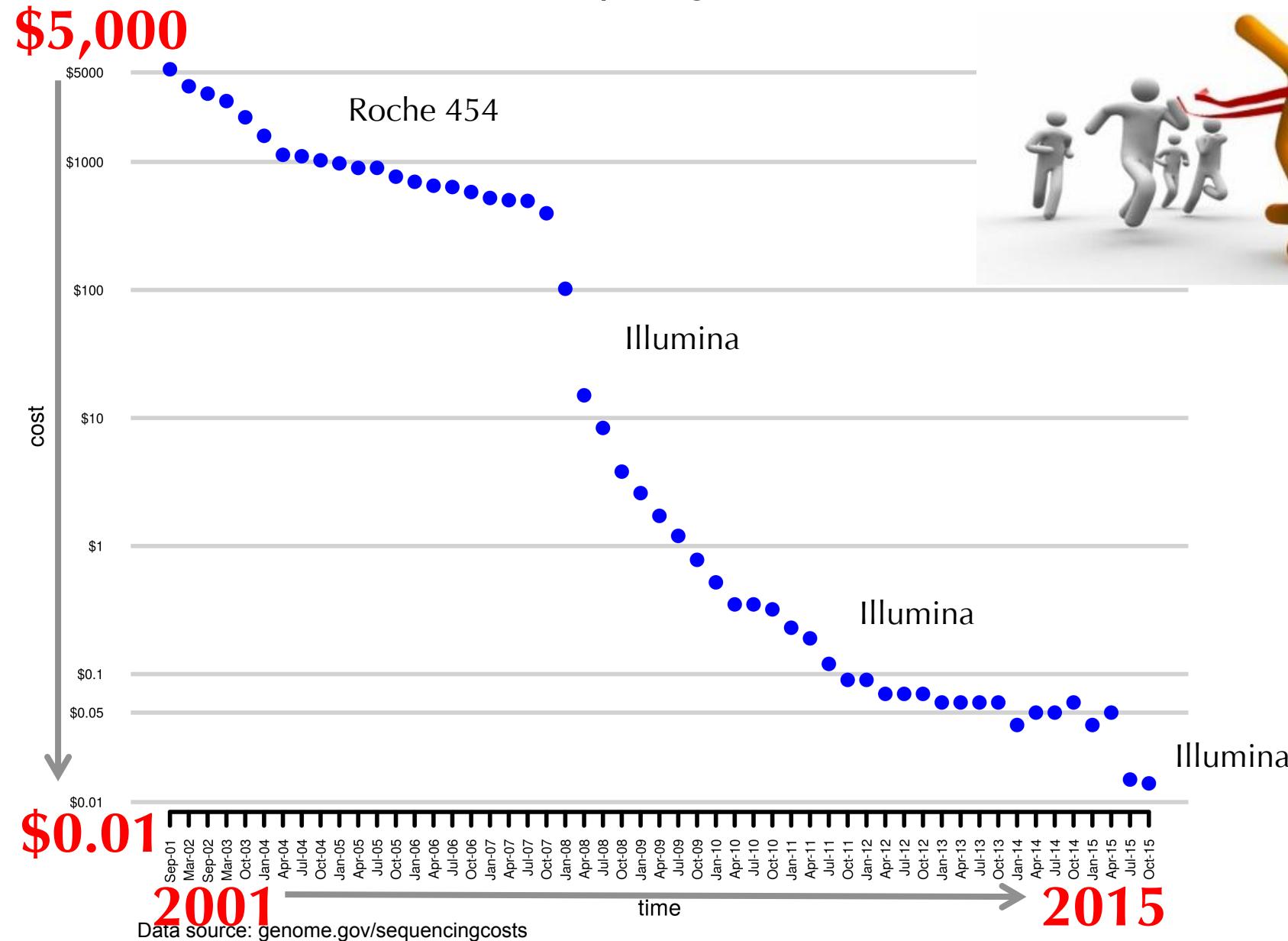
10 1



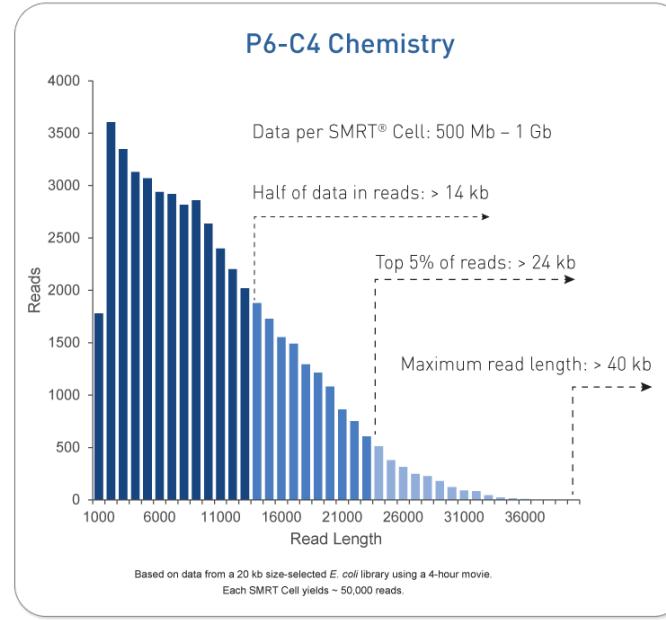
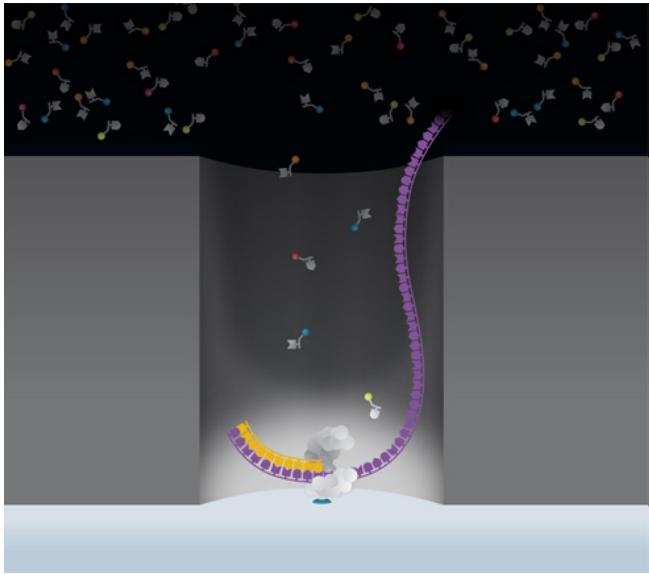
Differential expression (DE)?

Race of sequencing technologies

cost per megabase

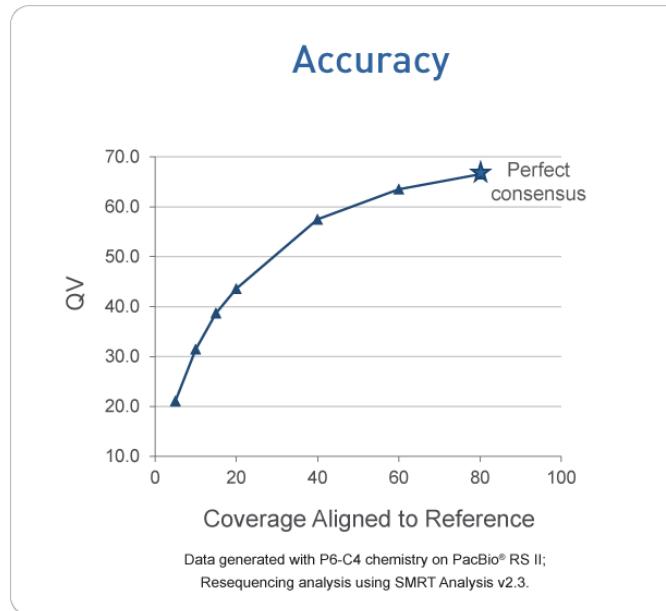


PacBio – Single Molecule Real Time (SMRT)



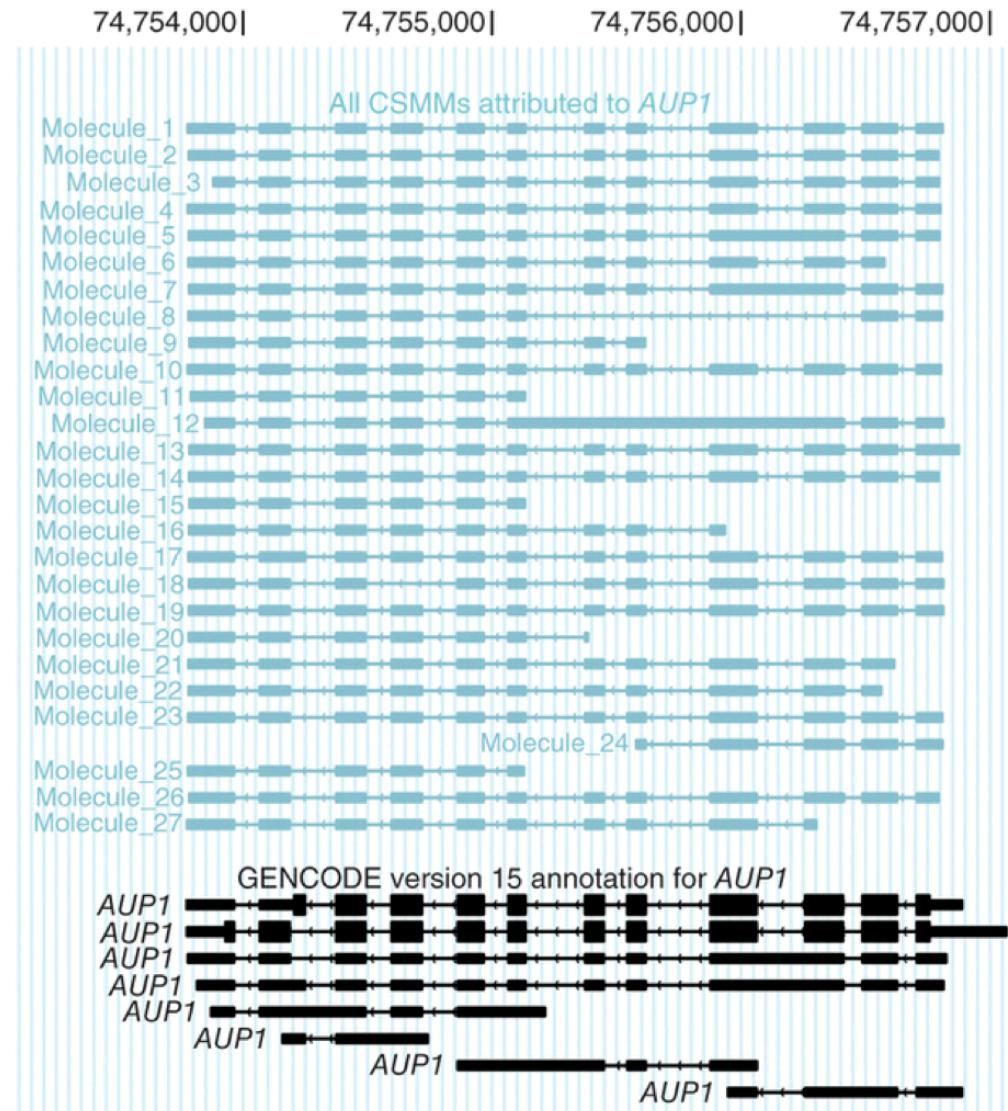
[PacBio tech video](#)

- Single molecule sequencing
- no amplifications required
- **up to 70+ kbp sequencing**
- Moderate sequencing throughput
- **high sequencing error rate (~15%, random, no-context-specific errors)**



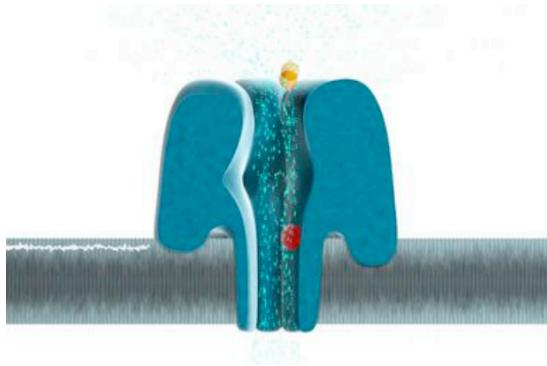
Long reads to sequence **full-length** cDNA

- The majority of reads represent all splice sites of original transcripts
- Isoforms can be monitored at a single-molecule level **without amplification** or fragmentation



Oxford Nanopore

A promising technology



As each nucleobase passes through the pore the current is affected and this change allows sequence to be read out.

- Single molecular sequencing
- No amplifications
- **Long reads (?kb)**
- **Error rate is high (~10-30%)**

MinION

1. USB disposable sequencer
2. Hundreds of Mb in several hours



Simple sample preparation
(Coming soon: automated sample preparation from Voltrax)



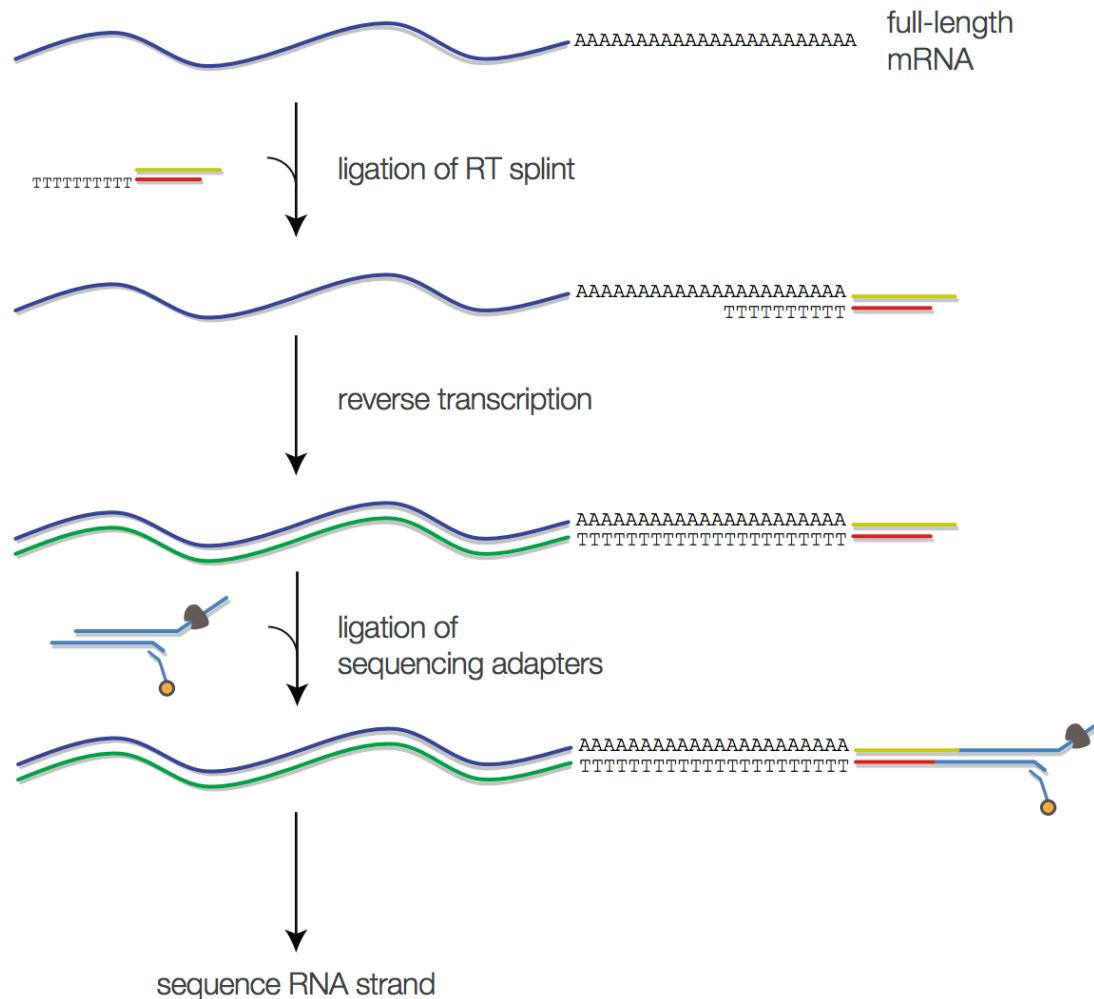
Pocket-sized MinION for analysis anywhere

PromethION (NEW)

1. “MinION cluster”?
2. High-throughput (1Tb output)



Nanopore – **direct** RNA sequencing



RNA-Seq procedure for DE analysis

1

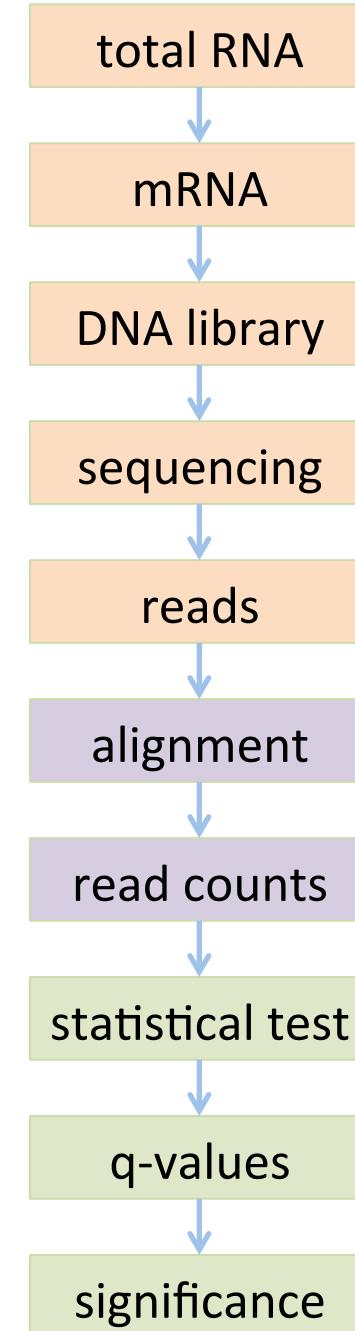
RNA to sequencing
reads

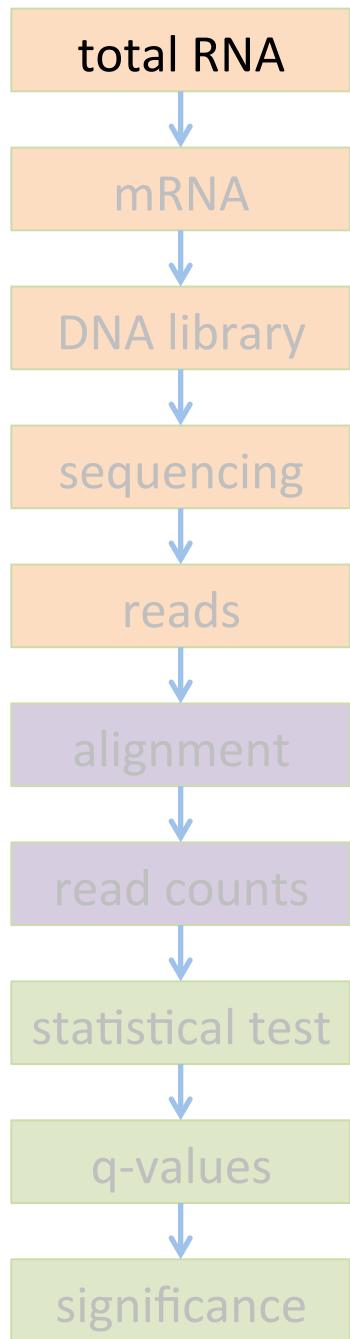
2

reads to read
counts per gene

3

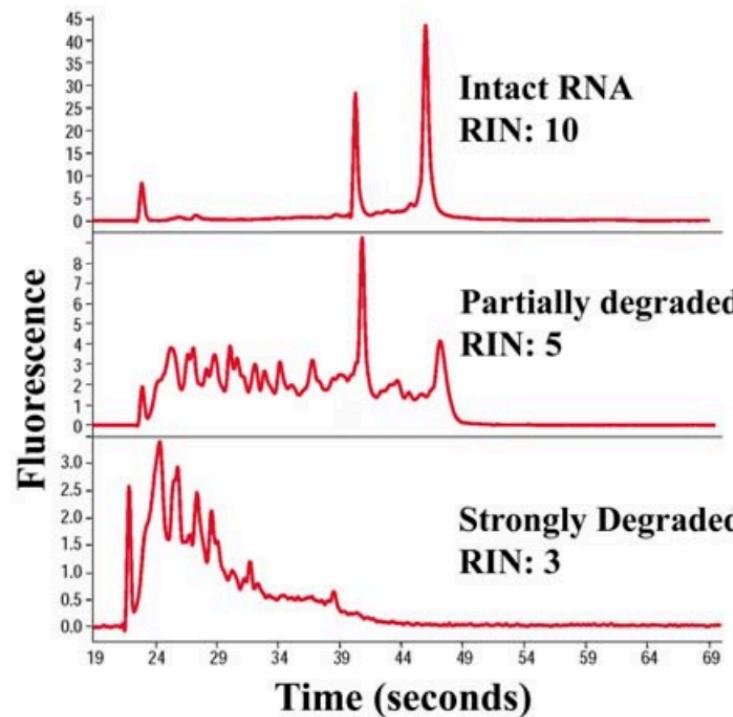
read counts to
significant genes

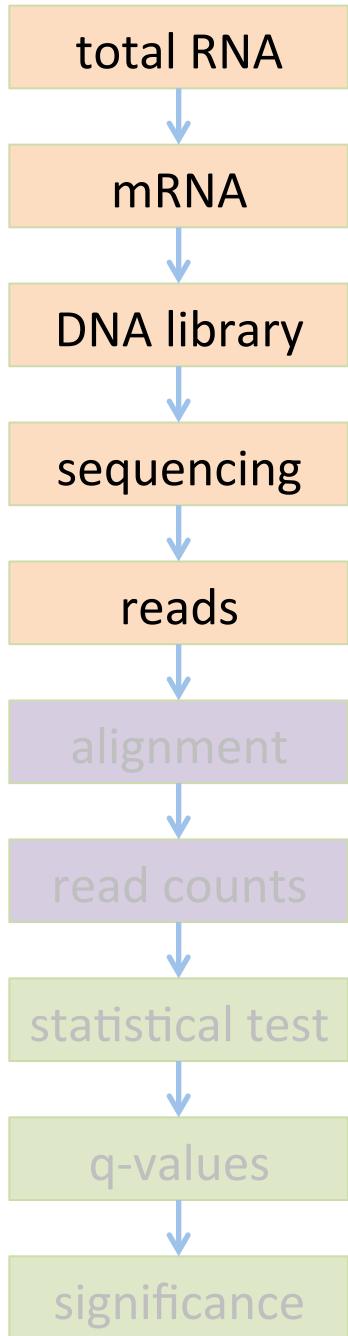




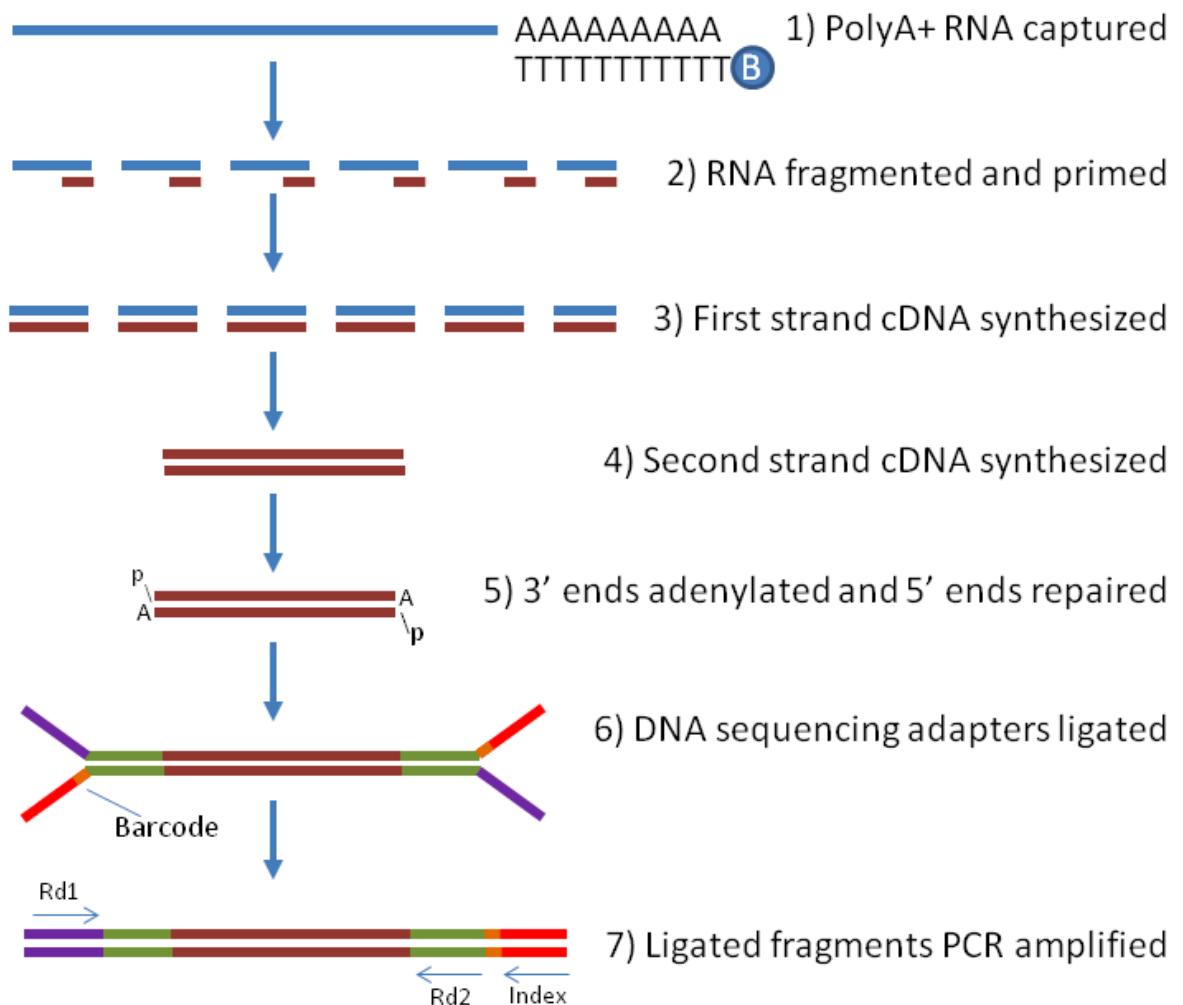
total RNA

RIN: RNA Integrity Number

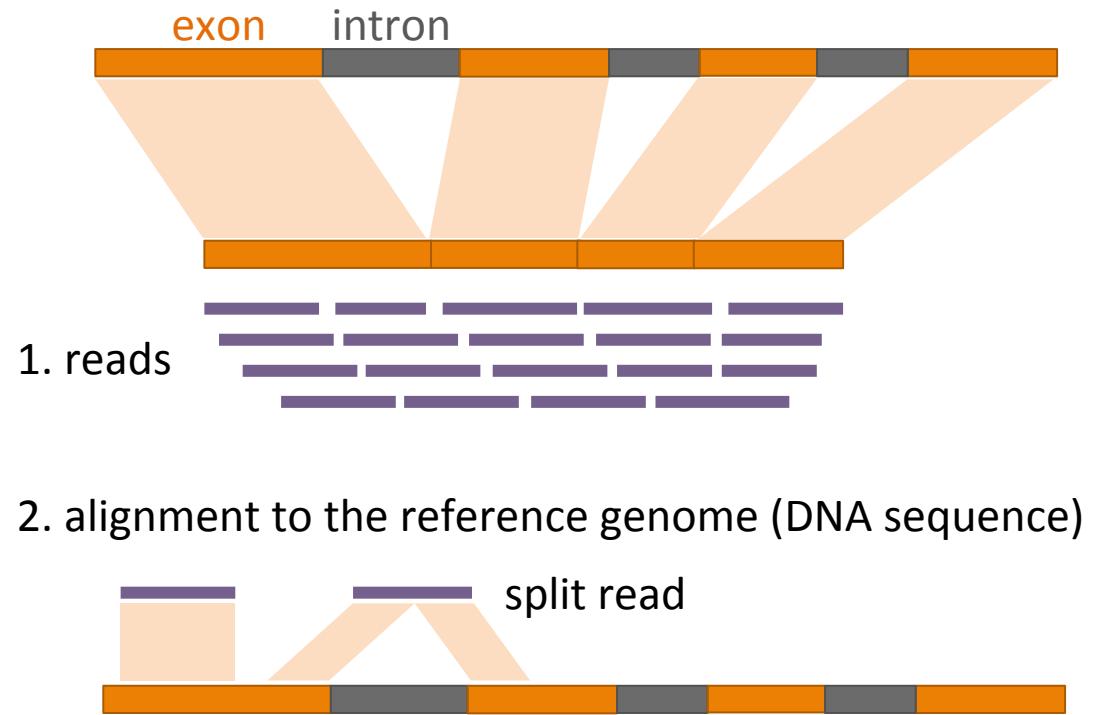
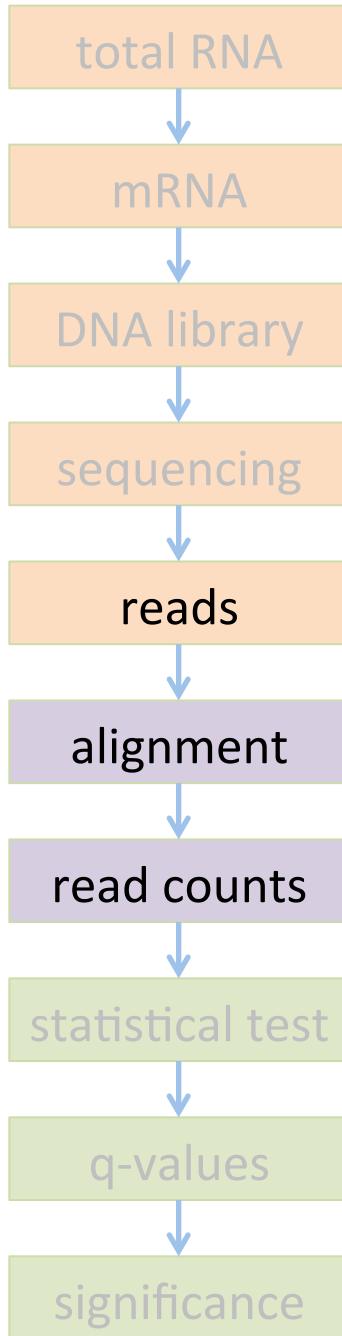




RNA to sequencing reads



Reads to read counts per gene



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

3. **read counts**

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

Reads to read counts per gene – commands required data

1. Reference genome (fasta file)
2. Reads (fastq files)
3. Gene information (gff file)

Reads to read counts per gene – commands **reference indexing**

```
STAR --runMode genomeGenerate \  
--genomeDir . \  
--genomeFastaFiles reference.fas \  
--sjdbGTFfile genes.gtf \  
--runThreadN 48
```

Reads to read counts per gene – commands alignment and read counting

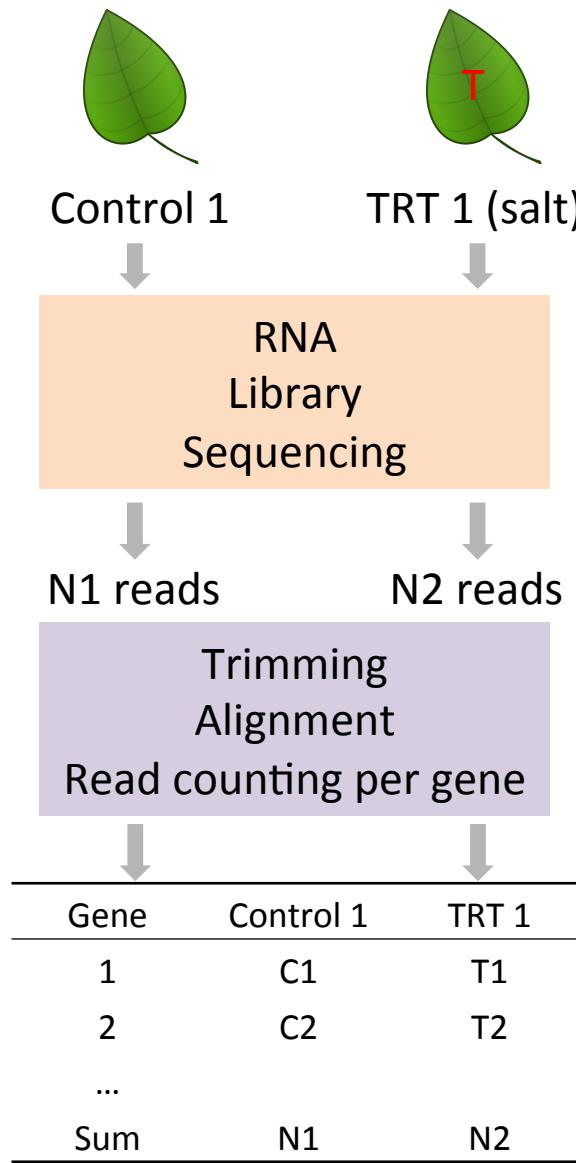
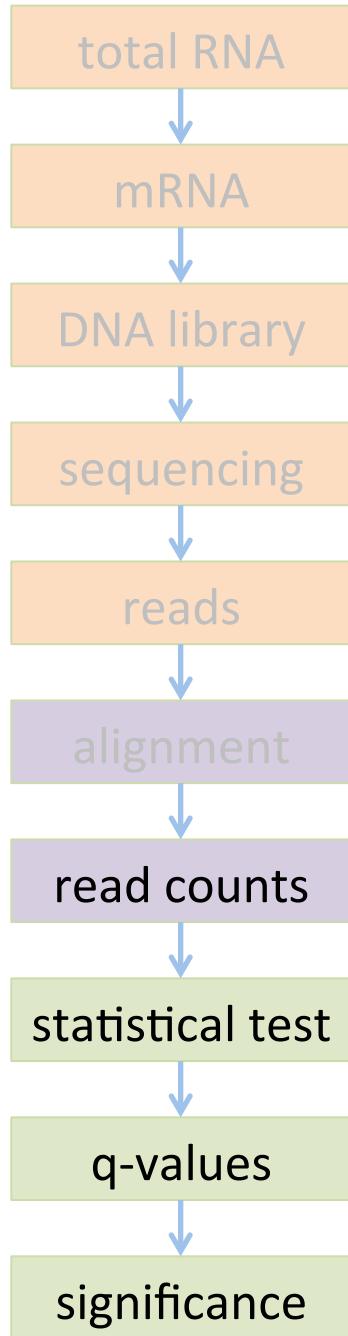
```
STAR --genomeDir reference.fas \
--readFilesIn read1.fq read2.fq \
--alignIntronMax 100000 \
--alignMatesGapMax 100000 \
--outFileNamePrefix output \
--outSAMattrIHstart 0 \
--outSAMmultNmax 1 \
--outSAMstrandField intronMotif \
--outFilterIntronMotifs RemoveNoncanonicalUnannotated \
--outSAMtype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

1. Reference genome
2. Reads
3. Gene information

Read counts (Raw) per gene

	Gene	sample 1	sample 2	sample 3
gene 1	483	6,075	5,934	3,370
gene 2	1,422	295	377	169
...

Read counts to significant genes



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 exists
 - **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 Cervera¹⁰, Andrew McPherson¹¹, Michał Wojciech Szcześniak¹², Daniel J. Gaffney³, Laura L. Elo¹³, Xuegong Zhang^{14,15} and Ali Mortazavi^{16,17*}

Outline

Review of RNA-Seq procedure

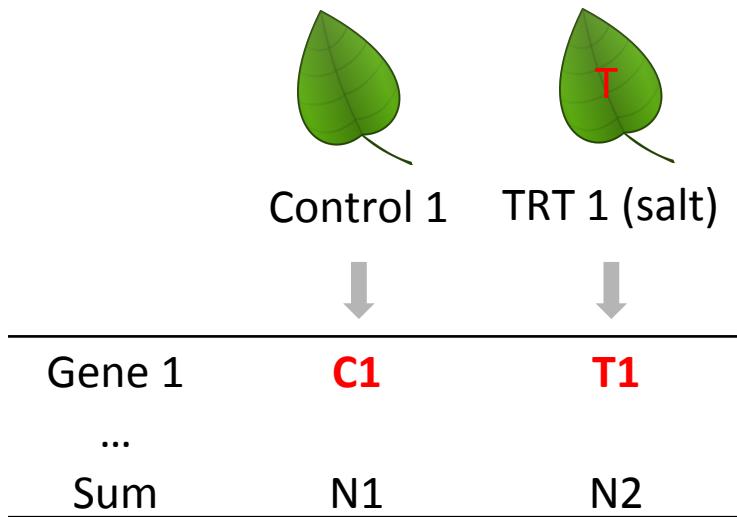
Design of DE experiments and results

- Experimental design
- Multiple test correction

Other analyses

- Visualization
- GO term enrichment analysis

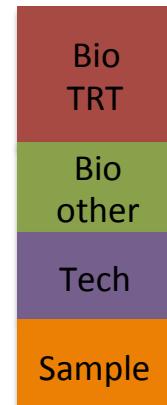
An RNA-Seq experiment – source of variance



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

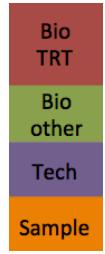
Randomly sample 10^7

First sampling	6
Second sampling	13
Third sampling	8

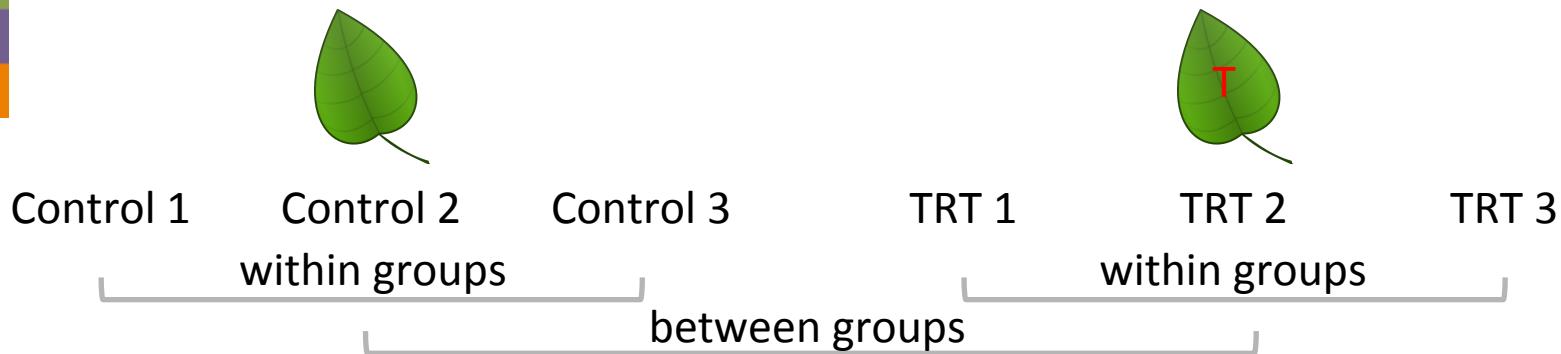
Randomly sample 10^8

First sampling	102
Second sampling	93
Third sampling	97

Sequence depth (sampling number) matters.



Technical replication



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

Technical
 replicate
 Tech
 Sample
 within
 groups

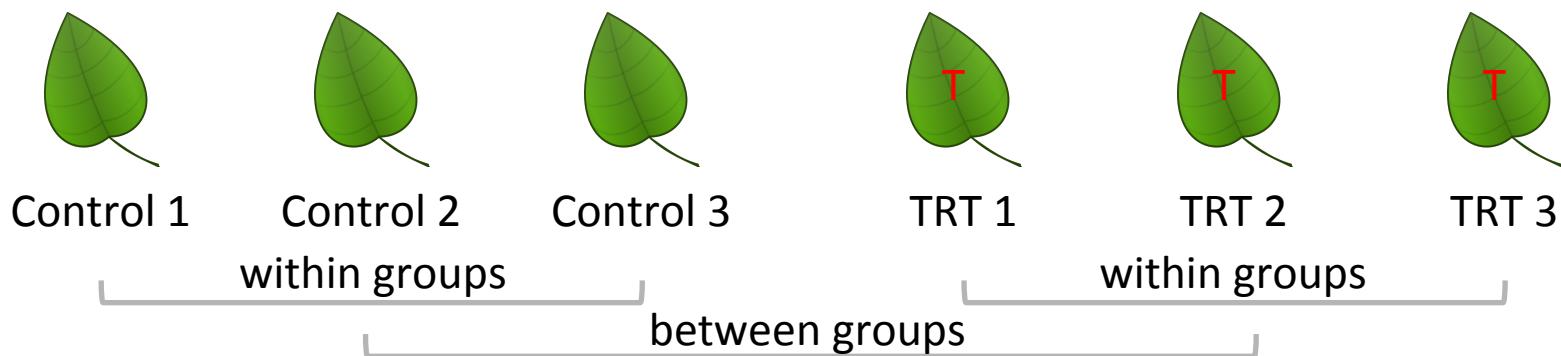
Bio
 TRT
 Bio
 other
 Tech
 Sample
 between
 groups

Compare to declare
 the significance

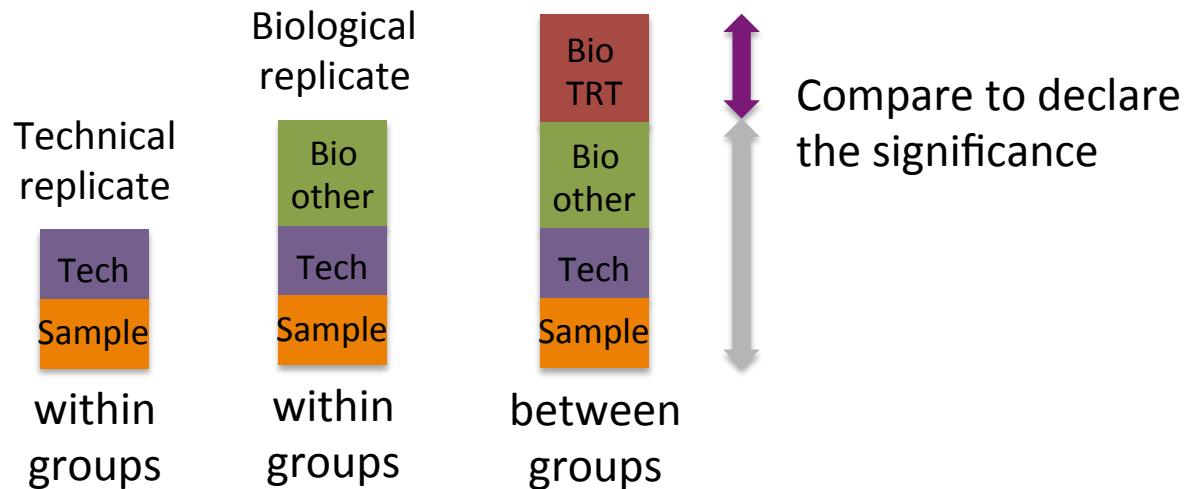
False power



Biological replication



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

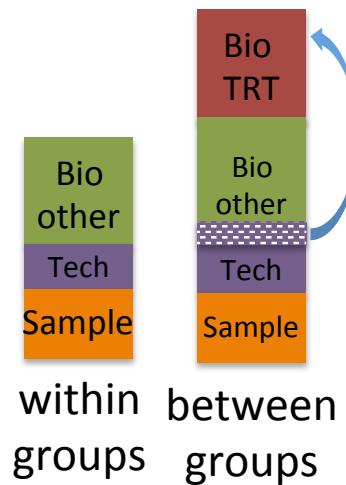


1. Use biological replication instead of technical replication unless you have your own interest.
2. More replicates increase the power to detect small treatment effect

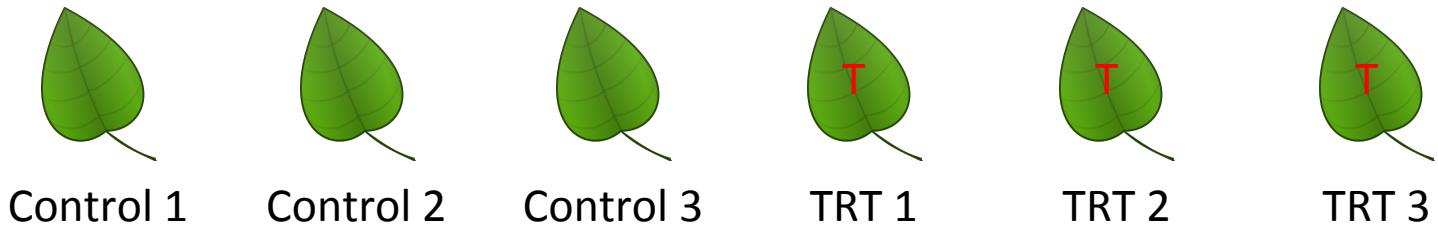
Question I

My lab conducted an RNA-Seq experiment to identify the DEs between two biological groups to examine a treatment of great interest. Each group has five biological replicates. I told my graduate student to perform the experiment of each group separately (then I don't need to worry that the samples from two groups are messed up).

Is this a sound experimental design? Why?



Comparison of read counts among different samples



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**

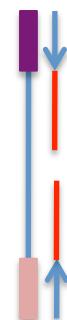


total reads: **10 millions** of total reads

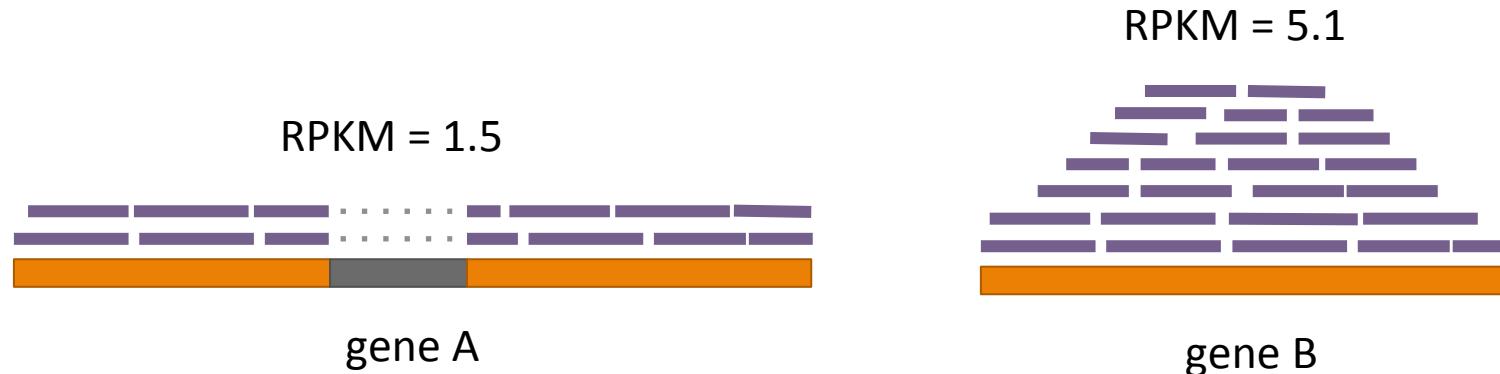
RPKM of X = ? = **3.6**

- **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 gene B has higher expression than 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

Outline

Review of RNA-Seq procedure

Design of DE experiments and results

- Experimental design
- Multiple test correction

Other analyses

- Visualization
- GO term enrichment analysis

DE result

DE Result		
GenelD	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 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:**

$p_1 = 0.8; p_2 = 0.1; p_3 = 0.3; p_4 = 0.5; \dots; p_{20} = 0.03$

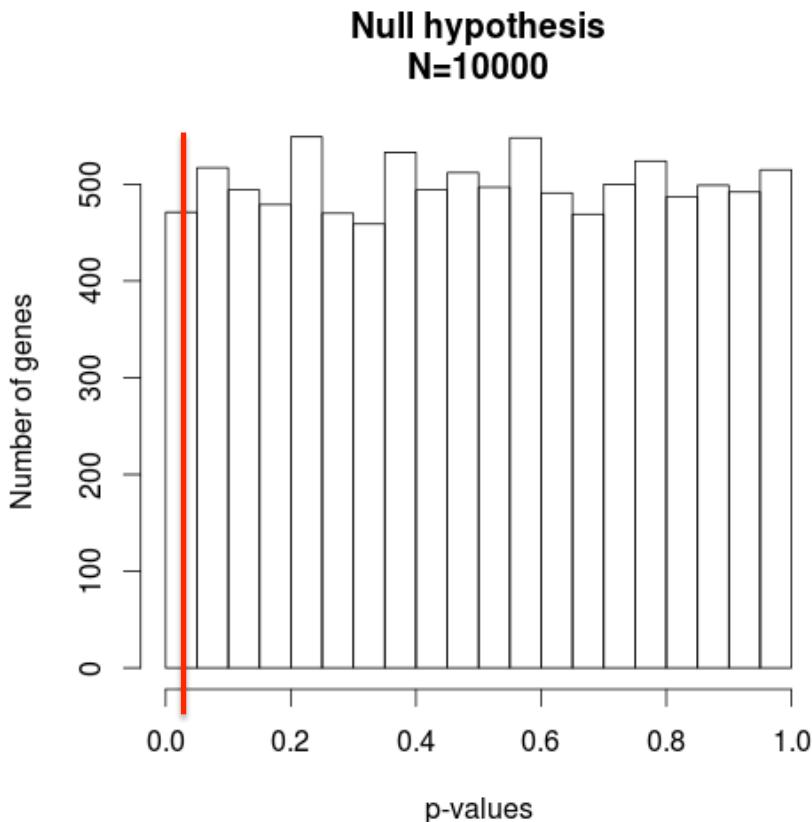
At the 5% significant level (P-value threshold = 0.05),
we will reject the null hypothesis for p_{20} .

Anything wrong here?

Multiple testing correction

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

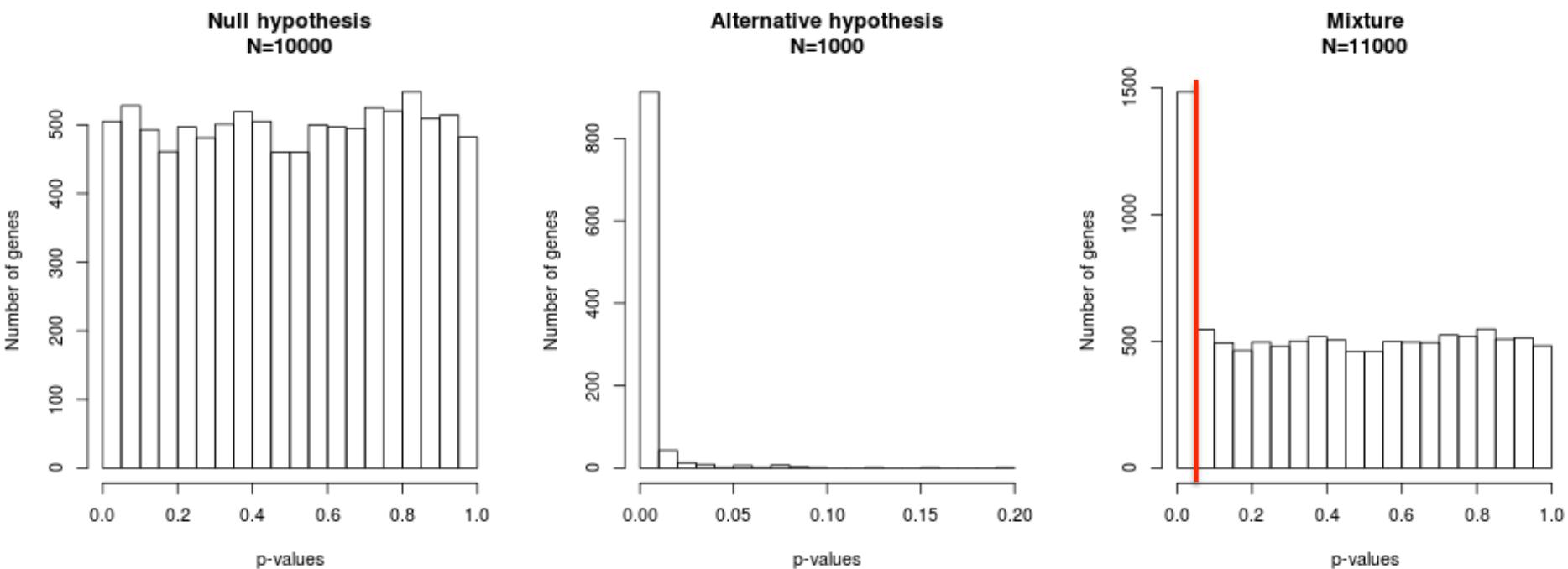
P-value distribution under the null hypothesis (e.g., no treatment effect)



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, a P-value is distributed uniformly from 0 to 1.

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



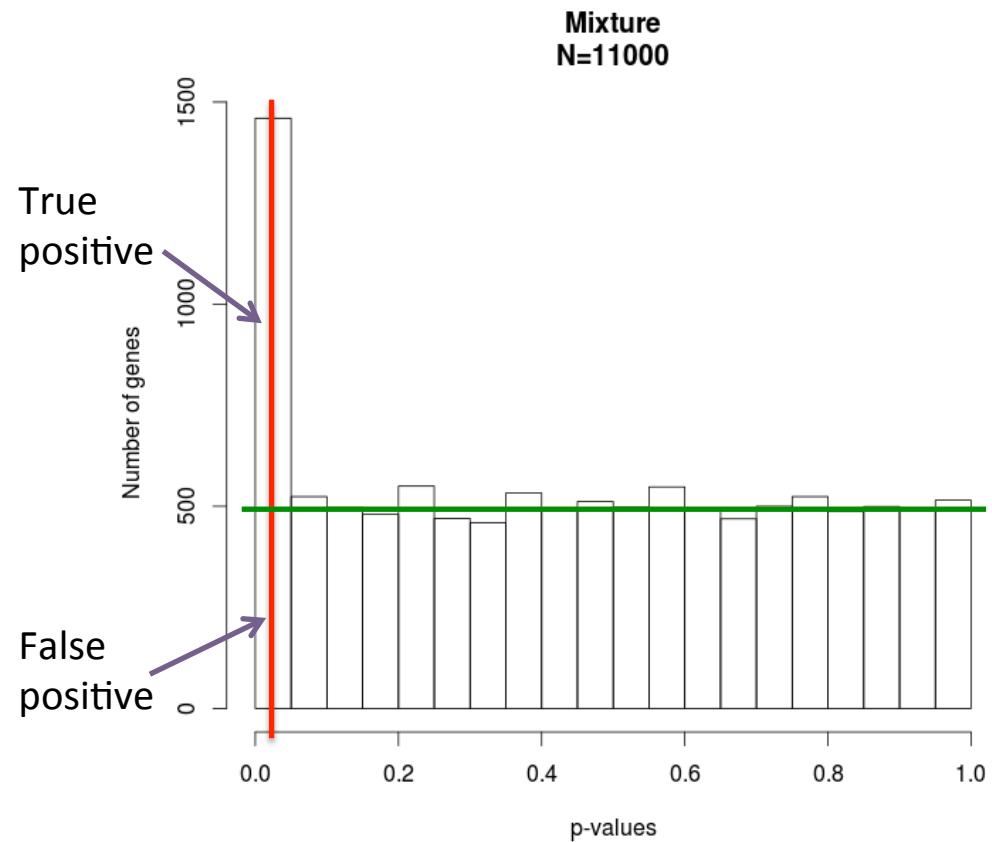
When the null hypothesis is true, a P-value is distributed uniformly.

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

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

Total number of tests: m = 20

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?

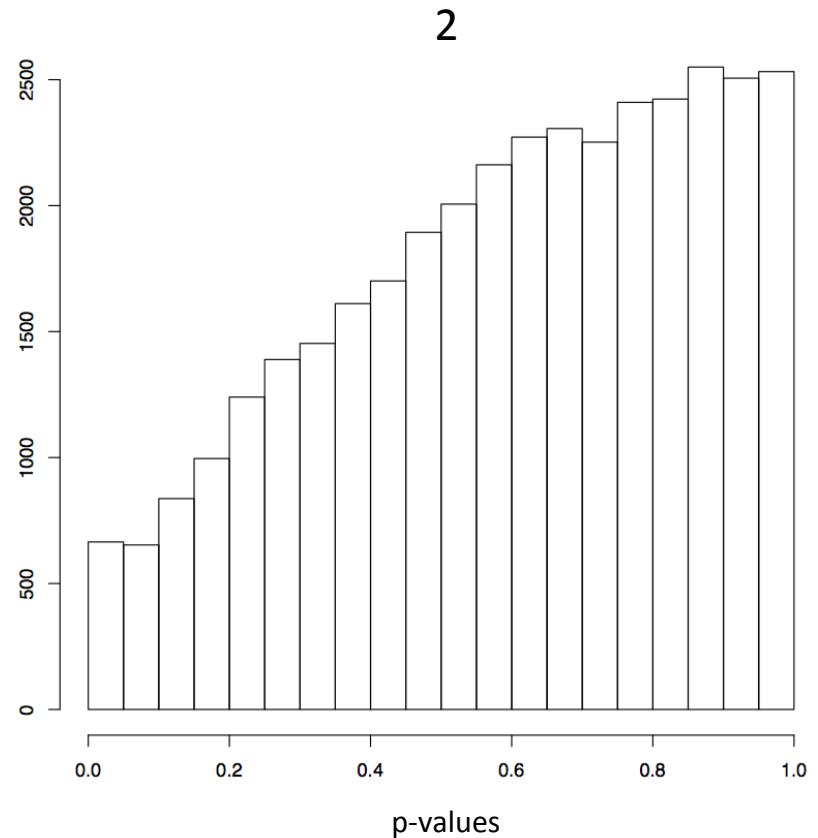
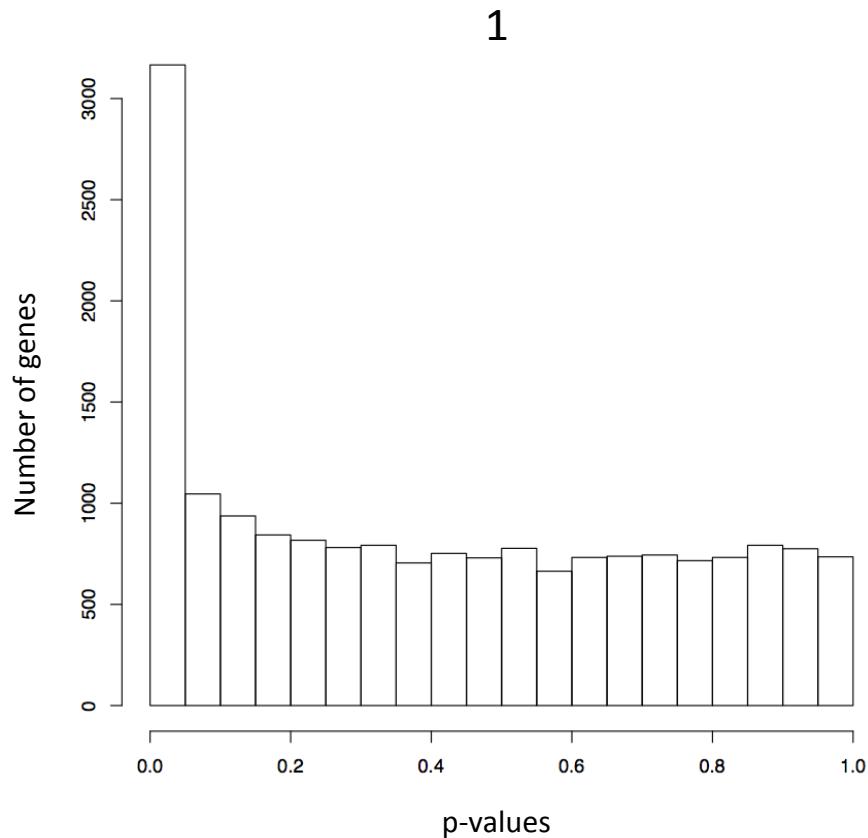
1. I am 95% confident that 500 genes are DE.
2. The 5% genes (25 genes) in the set are expected to be false DE genes.

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%.

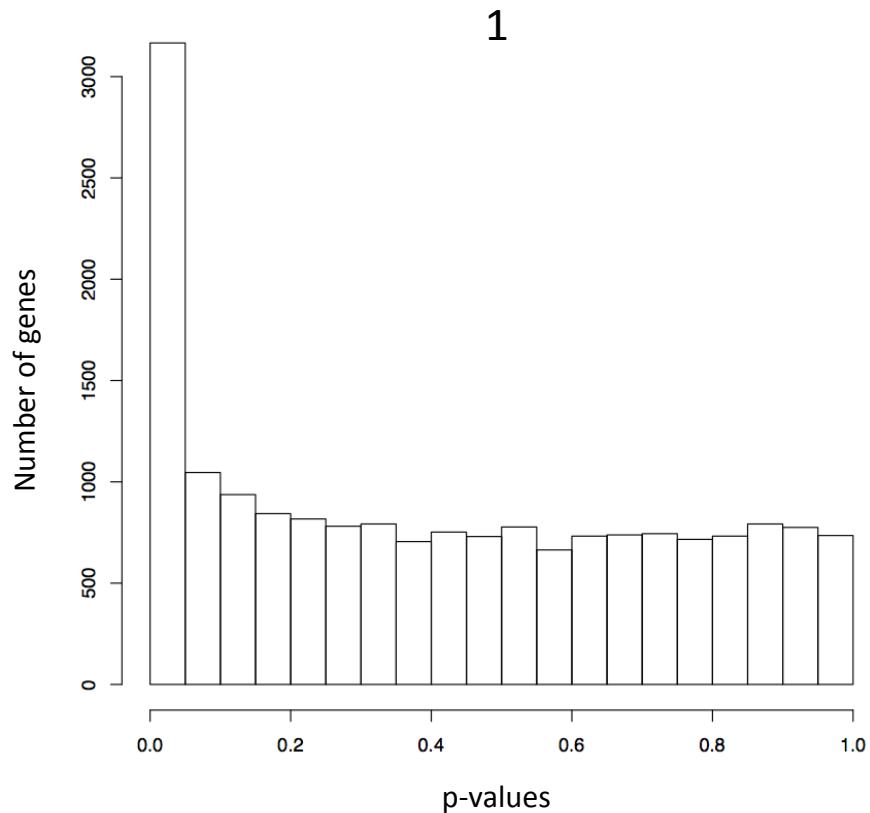
	True null hypothesis (H_0)	False null hypothesis (H_1)	Total
Rejected (Declared significance)	10	90	100

P-value histograms from real studies

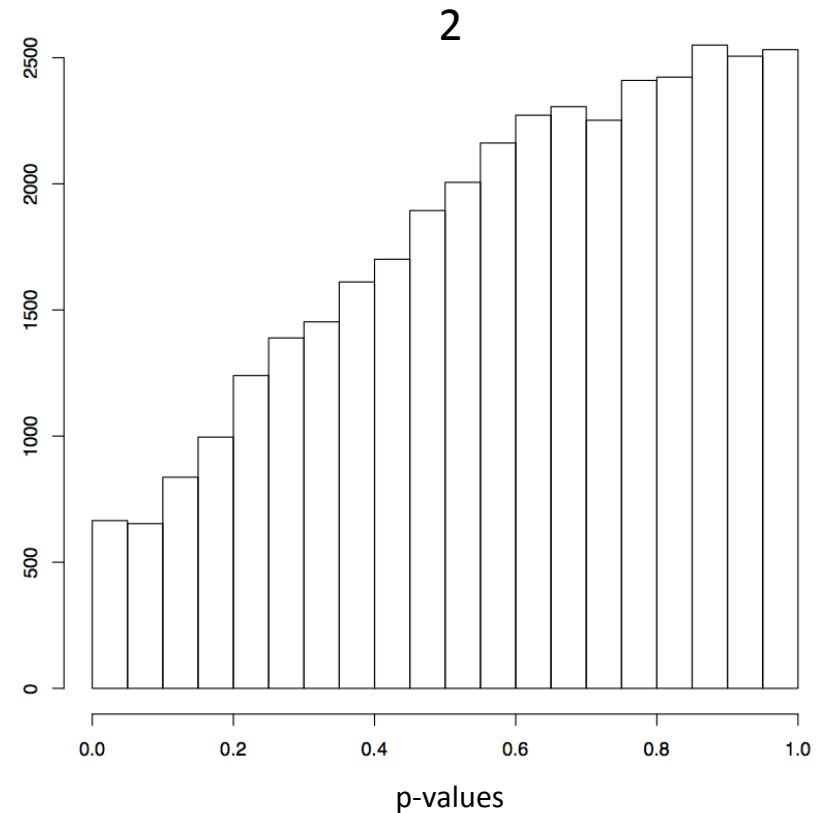


If you perform an RNA-Seq experiment, which one would you like to obtain?

P-value histograms from real studies



DE = 1,370, FDR=5%



DE = 0, FDR=20%

Outline

Review of RNA-Seq procedure

Design of DE experiments and results

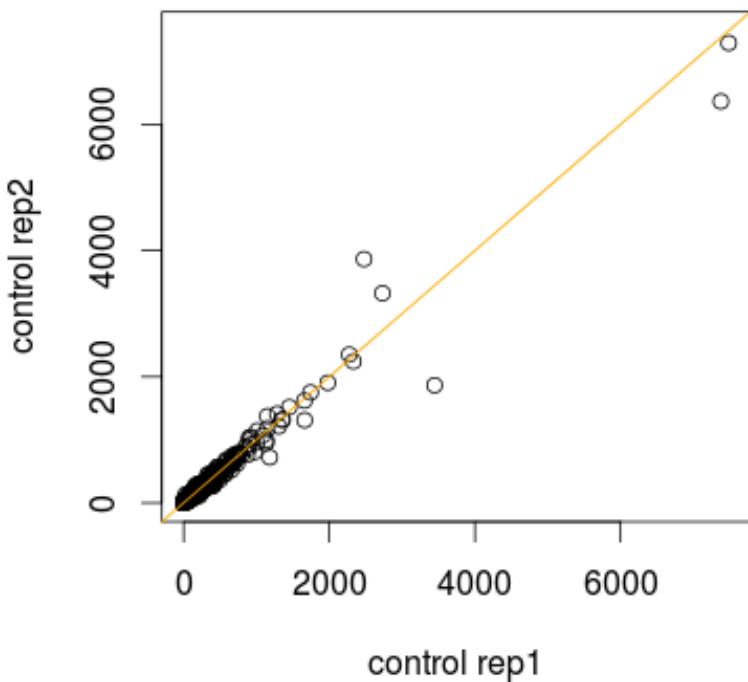
- Experimental design
- Multiple test correction

Other analyses

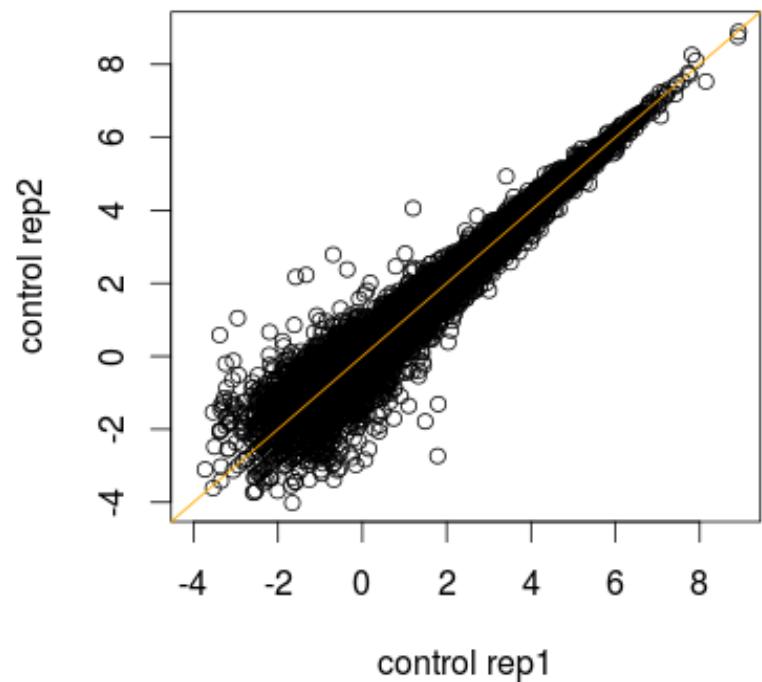
- Data visualization
- GO term enrichment analysis

Scatter plot

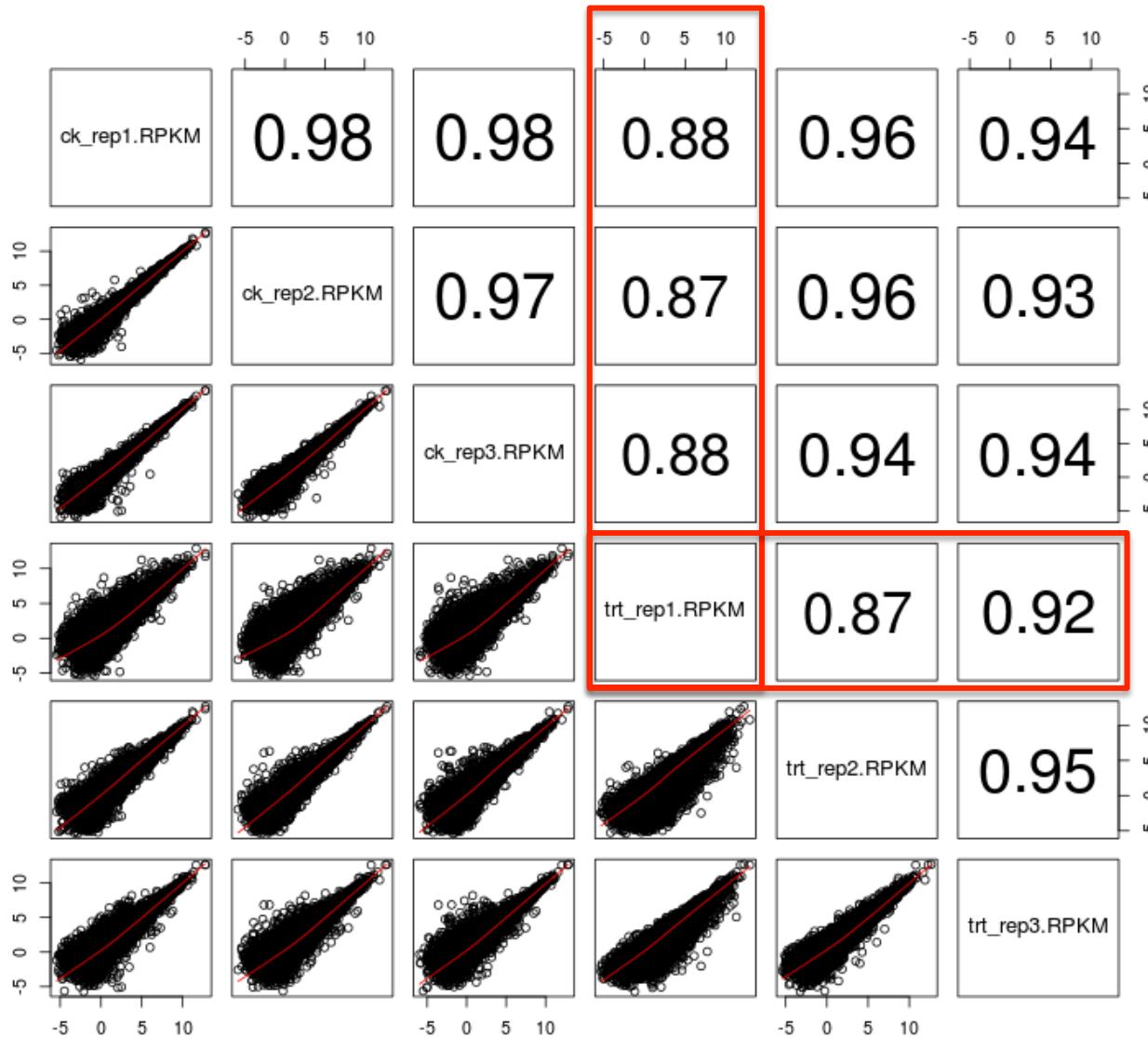
RPKM scatter plot



Log RPKM scatter plot



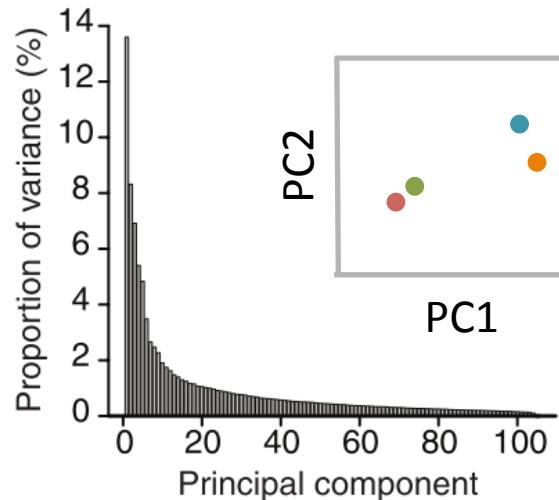
Pair-wise scatter plot



Principal Component Analysis (PCA)

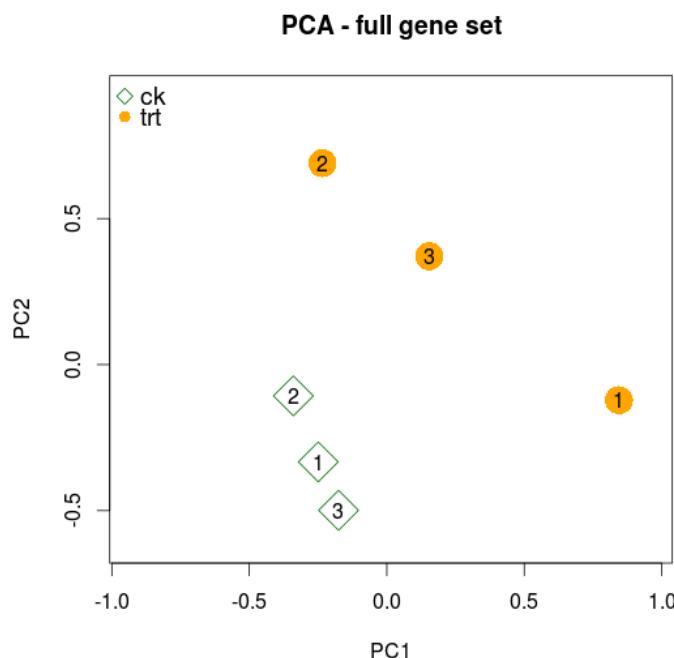
PCA is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set.

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



GenelD	Control			Treatment		
	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

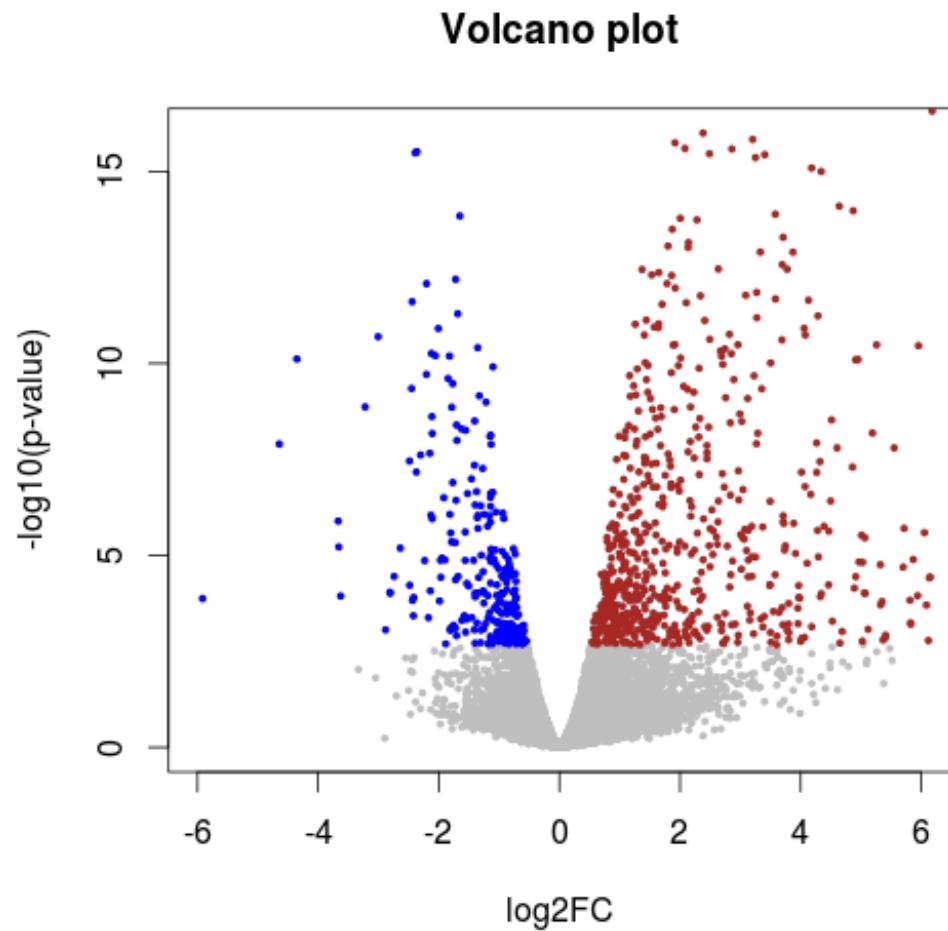
Normalized and standardized data



Volcano plot



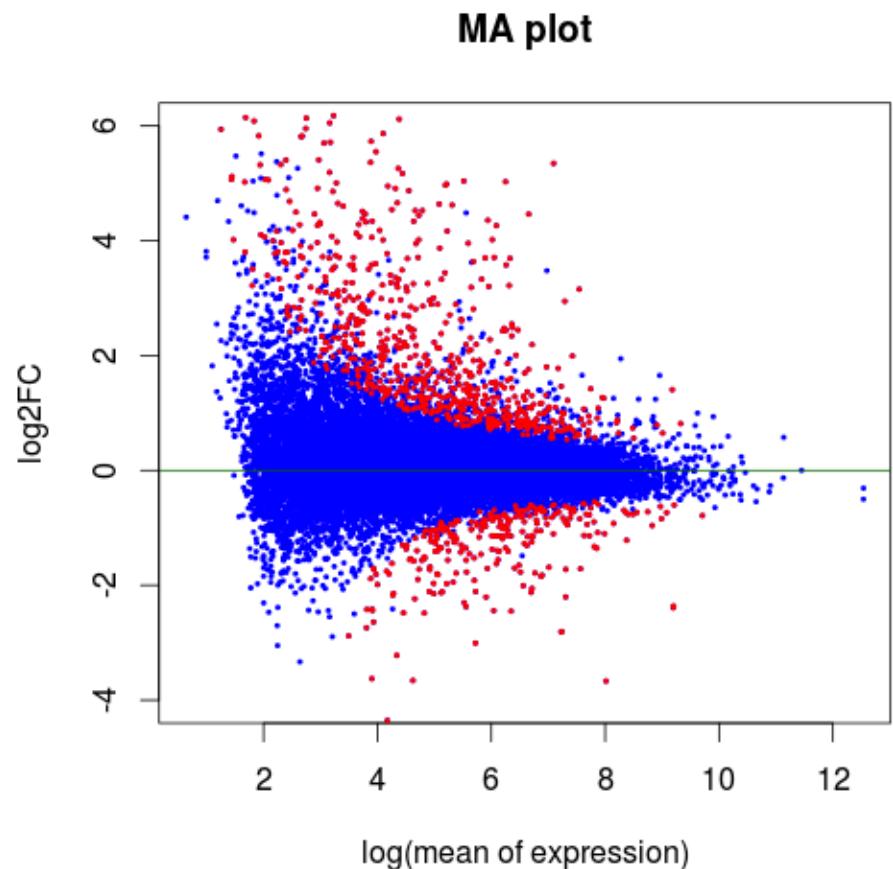
DE Result			
GenID	Log2FC	p-value	-log10(pvalue)
1	-0.40	0.037	1.43
2	0.03	0.916	0.04
3	-0.89	2.42E-05	4.62
4	0.30	0.130	0.89
5	-0.36	0.140	0.85
6	-0.07	0.811	0.09
...			



MA plot

M (log ratios) and A (mean average)

GenID	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

Outline

Review of RNA-Seq procedure

Design of DE experiments and results

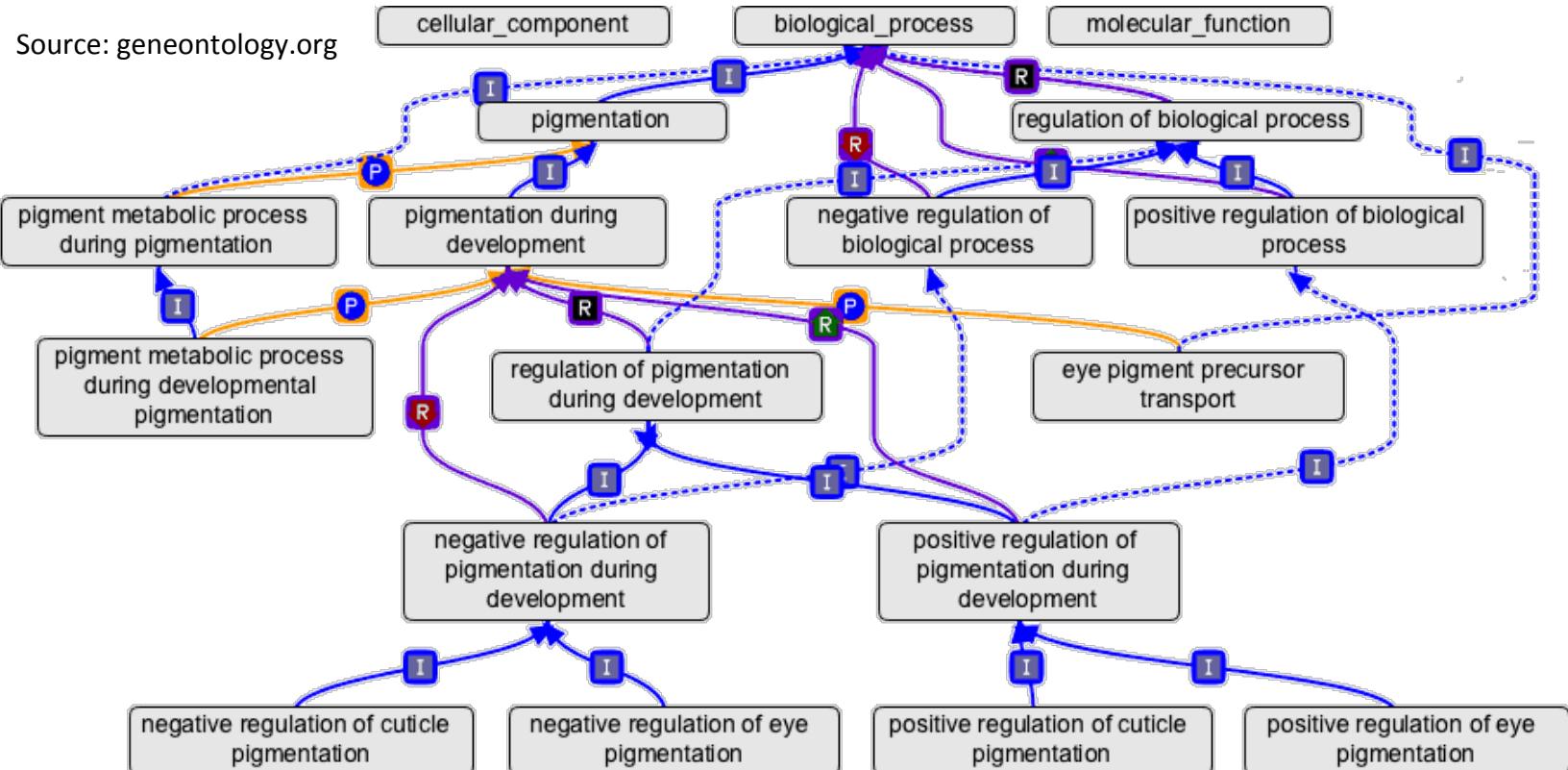
- Experimental design
- Multiple test correction

Other analyses

- Visualization
- GO term 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.



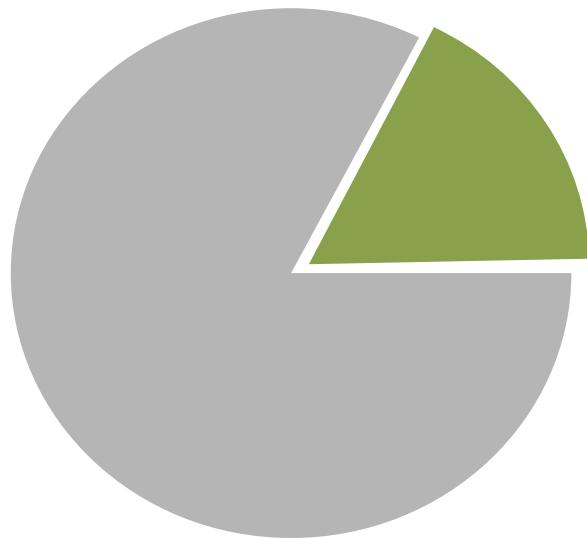
Three domains, three roots

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

Edge: term-term connection

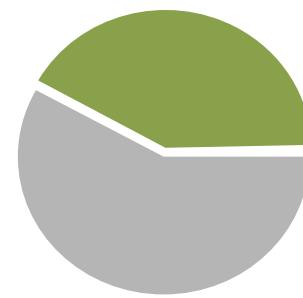
Each GO term can be traced back to a root

Category enrichment



All genes

Green enriched?



significant
gene set

GO enrichment 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
...	...

Fisher's Exact Test:
p-value = 2.518e-06



GRMZM2G001475	1
GRMZM2G002652	2
GRMZM2G006480	3
...	...
GRMZM5G868038	40

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

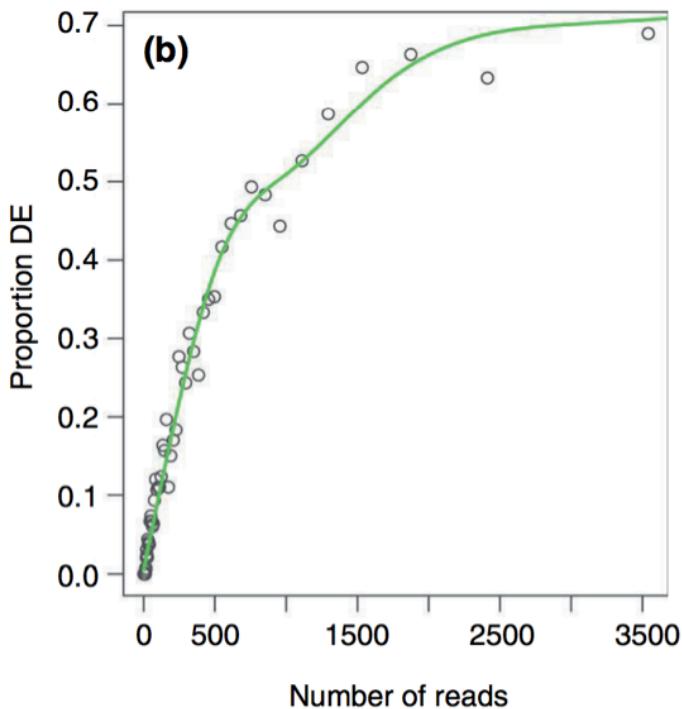
	GO:0006519	Others
Significant	5	210
Not significant	35	39416

Name
Ontology
Definition

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

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	Significant?	Read counts	Proportion
GRMZM2G001475	no	224	0.16
GRMZM2G002652	no	51	0.05
GRMZM2G006480	yes	536	0.38
...
GRMZM5G868038	no	0	0

3. Weighted sampling to perform enrichment test

GO:0006519	# DE
Obs (from the DE analysis)	5
1 st weighted sampling	1
2 nd weighted sampling	0
3 rd weighted sampling	2
...	...

→ p-value

Summary

- Biological replication rather than technical replication is 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.
- A proper GO enrichment test needs to be used.
- R is an excellent tool to visualize the data.
- Many bioinformatics pipelines and statistical methods have been developed. Methods and parameters need to be carefully selected.