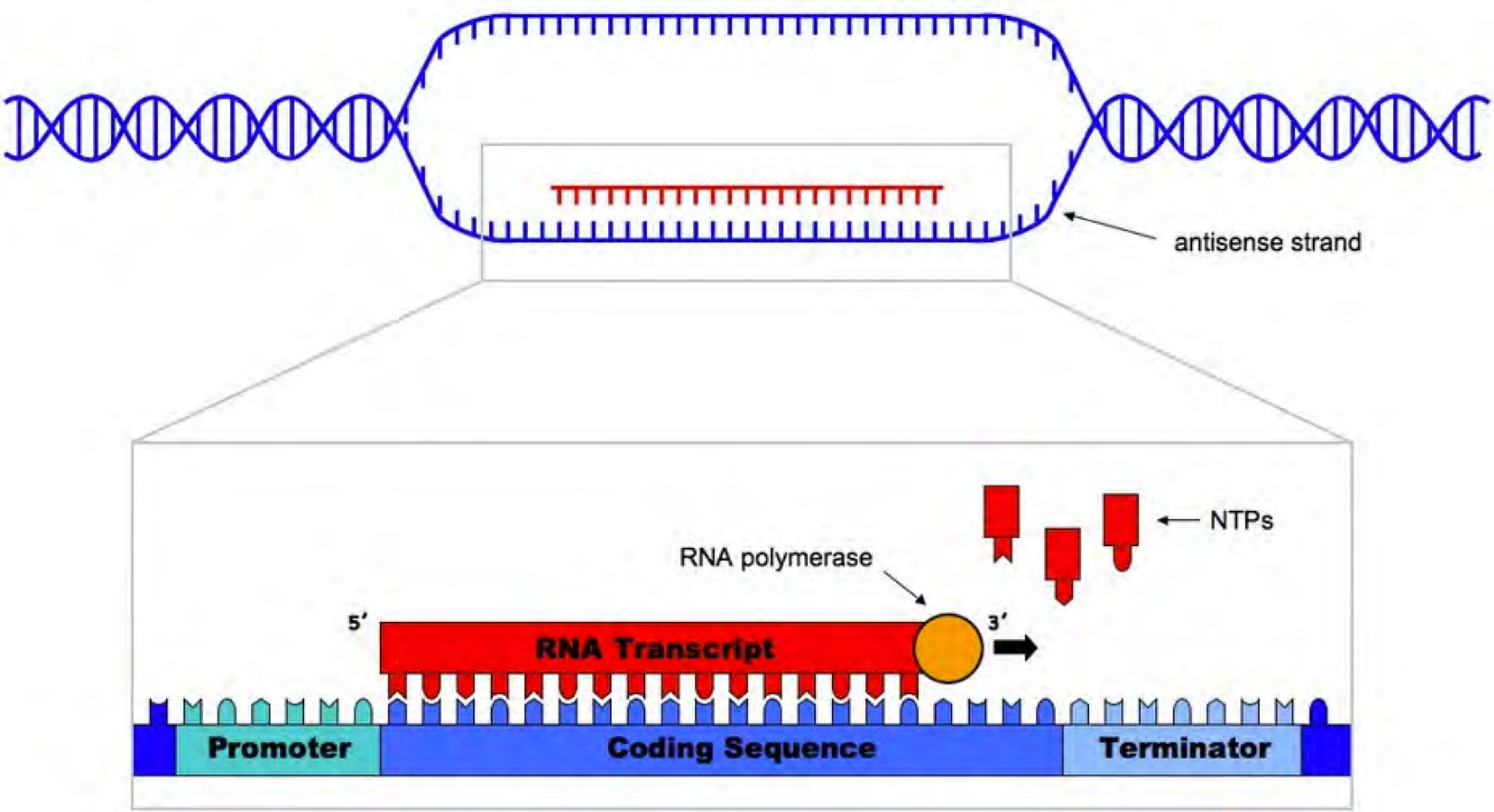
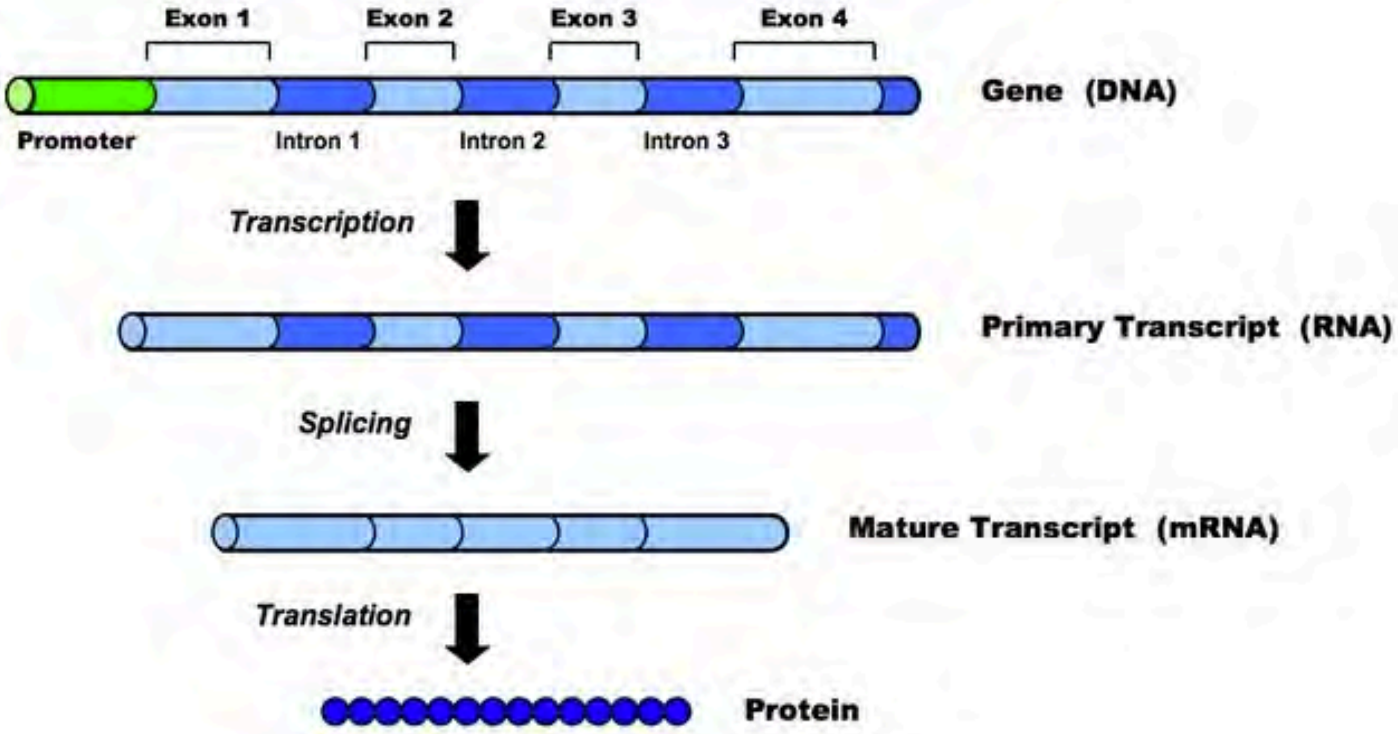
RNA-seq Data: Lecture 12

Amit Ghosh

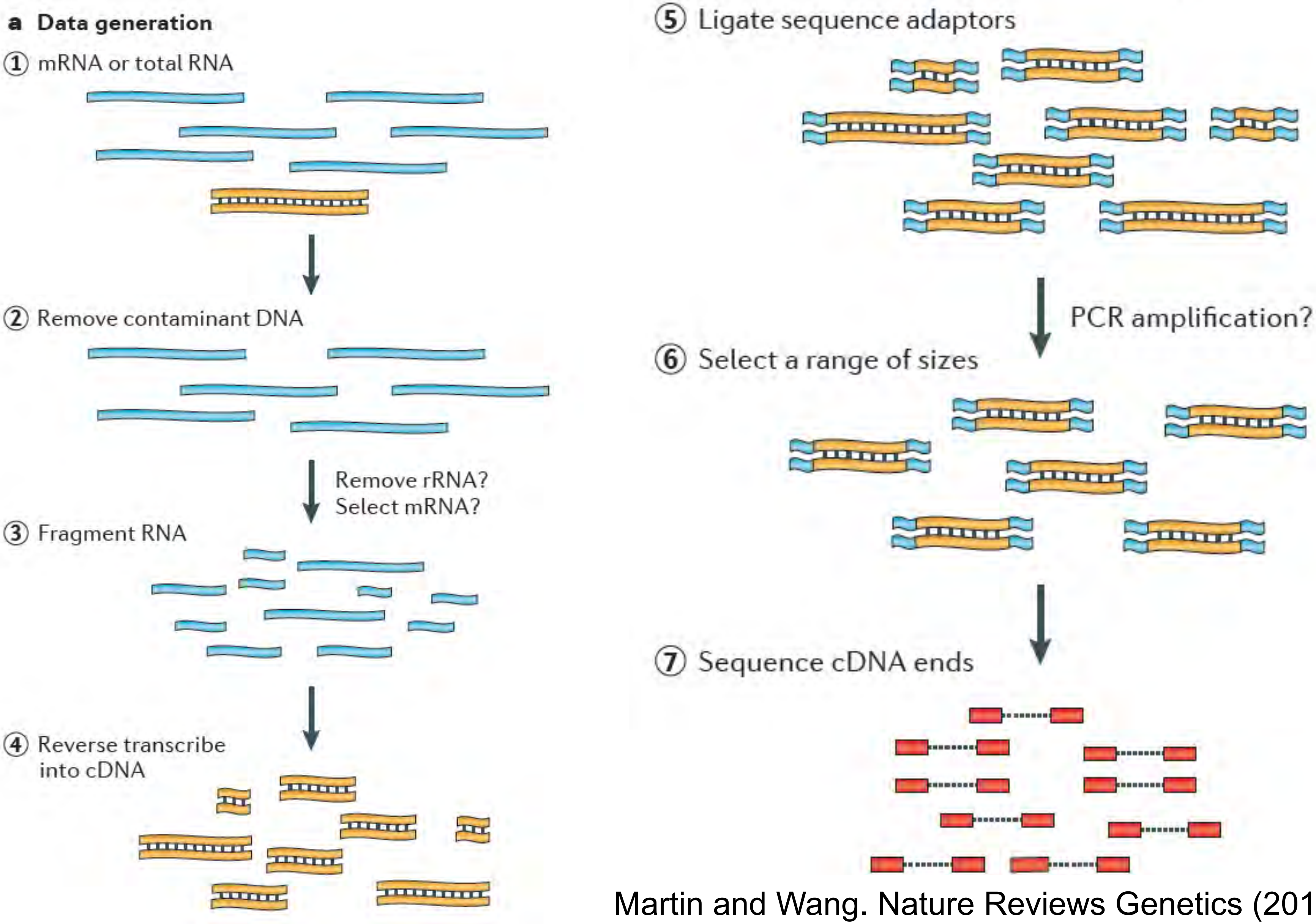
IIT Kharagpur



Transcription and Splicing



RNA-seq Protocol



a Data generation

① mRNA or total RNA

② Remove contaminant DNA

Remove rRNA?
Select mRNA?

③ Fragment RNA

④ Reverse transcribe
into cDNA

⑤ Ligate sequence adaptors

PCR amplification?

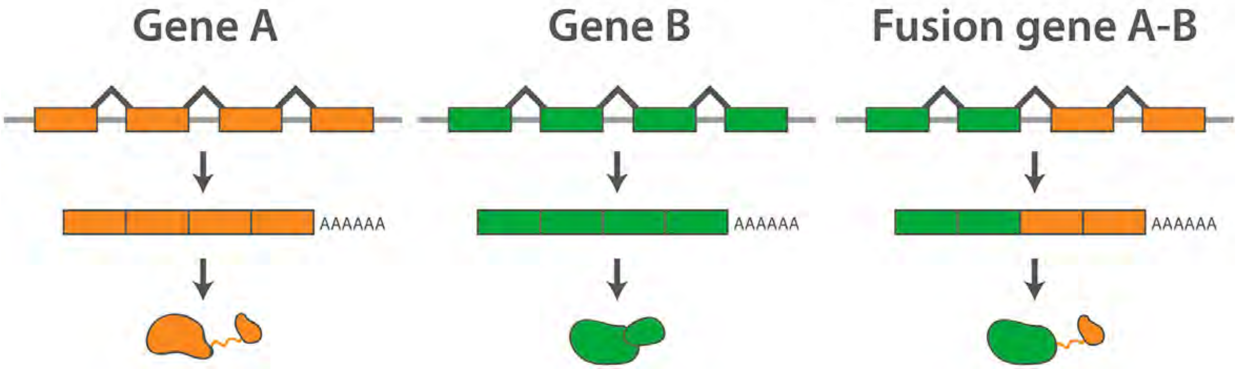
⑥ Select a range of sizes

⑦ Sequence cDNA ends

Martin and Wang. Nature Reviews Genetics (2011)

RNA-seq Applications

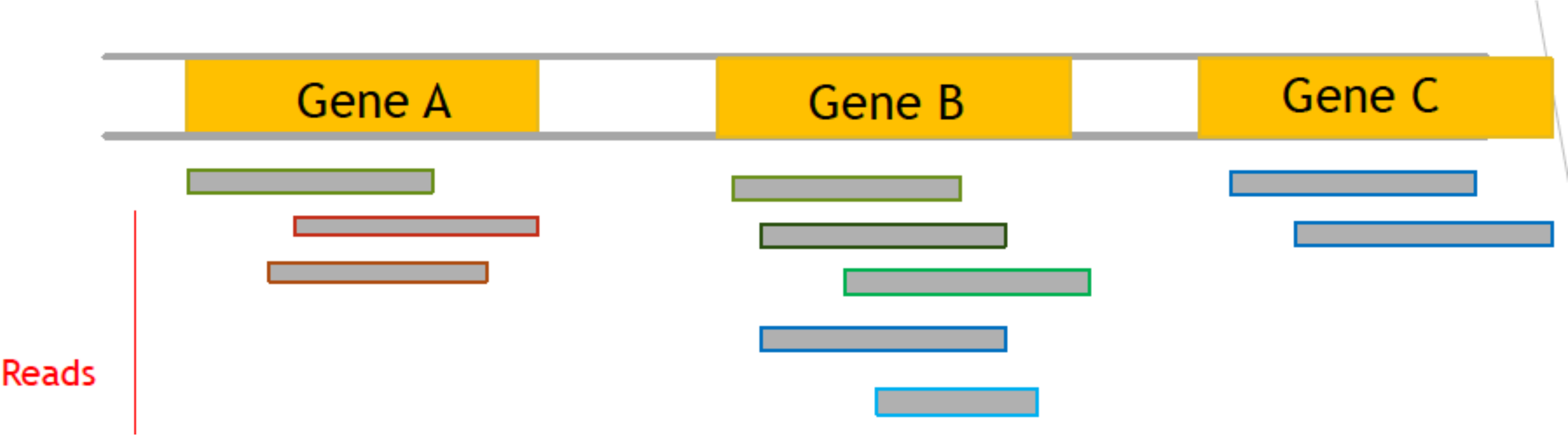
- Examine the expression of all the gene in specific conditions:
  a. Developmental stages
  b. Different tissues
  c. Normal vs. disease
  d. Drug treatment
  e. Gene perturbation



Latysheva *et a*l Mol Cell (2016)

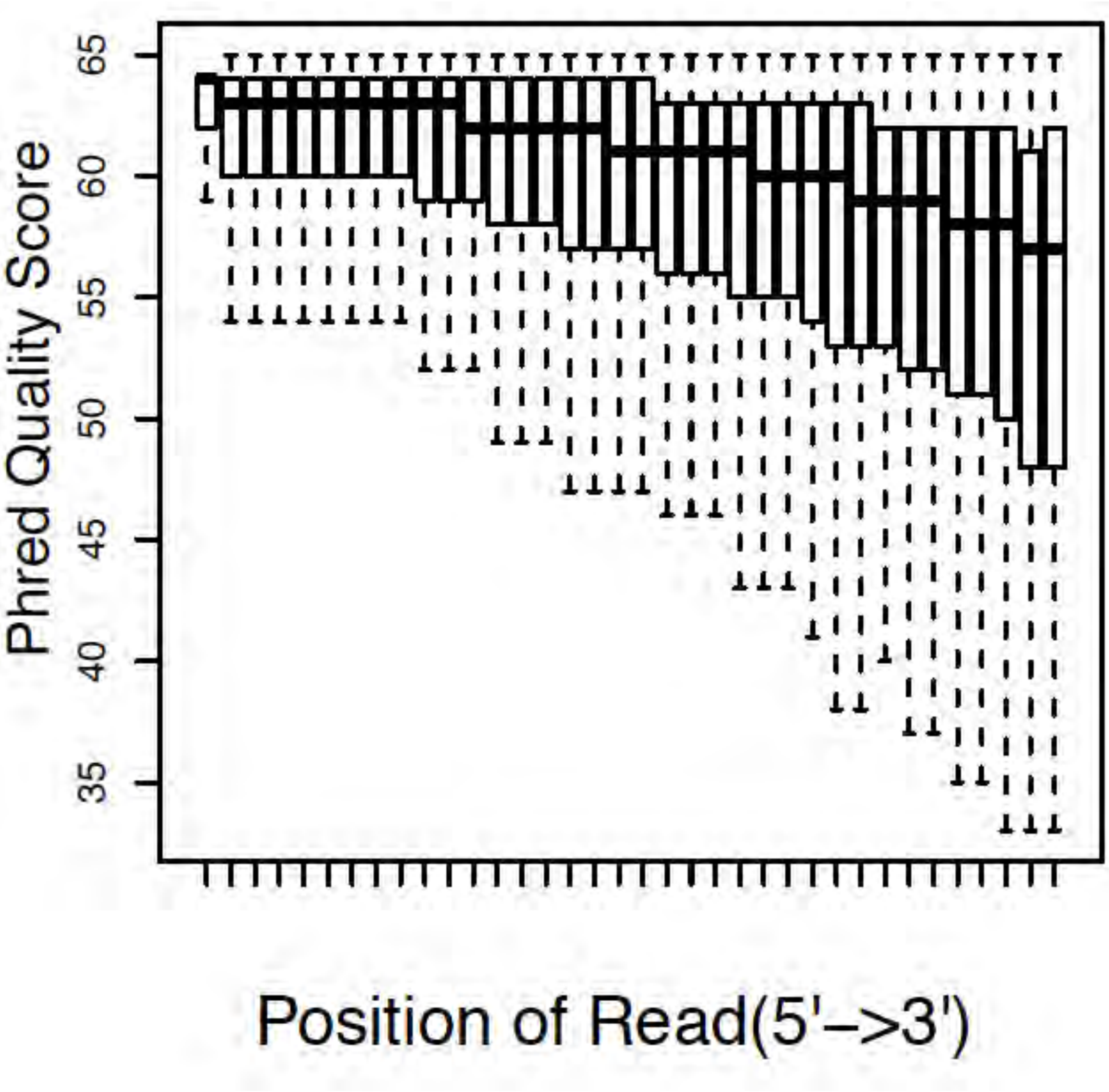
- Find novel genes or transcripts
- Find gene mutations or gene fusions
- Do not have to know the genome sequence or predict genes
- Digital representation gene expression

Mapping for transcriptome sequencing



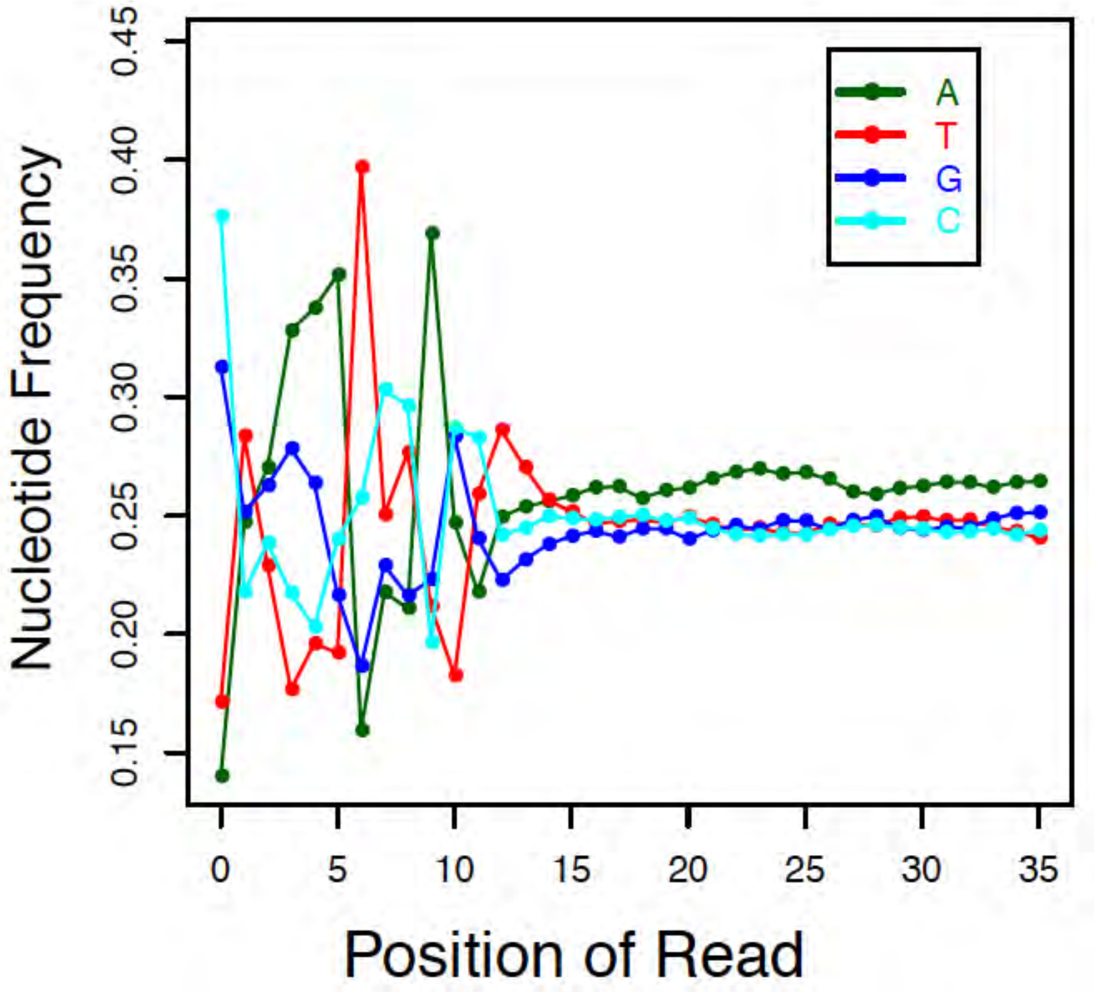
RNA-seq QC (RSeQC): FASTQC Read Quality

- Overall mappability reads
- Ideally greater than 50%
- Higher the better



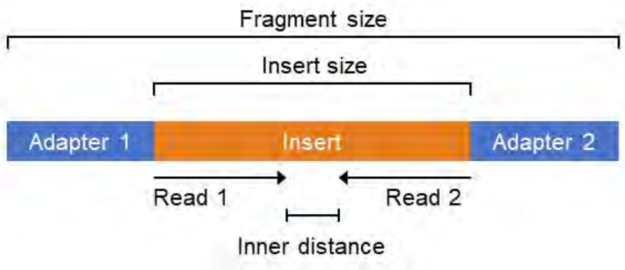
Wang *et a*l Bioinformatics (2012)

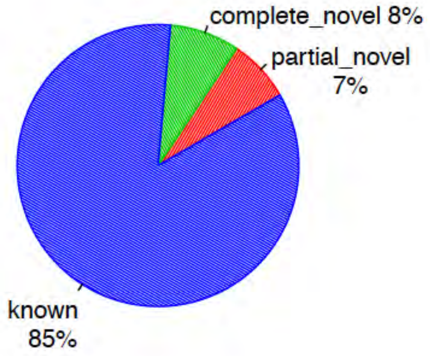
RSeQC: Nucleotide Compositions



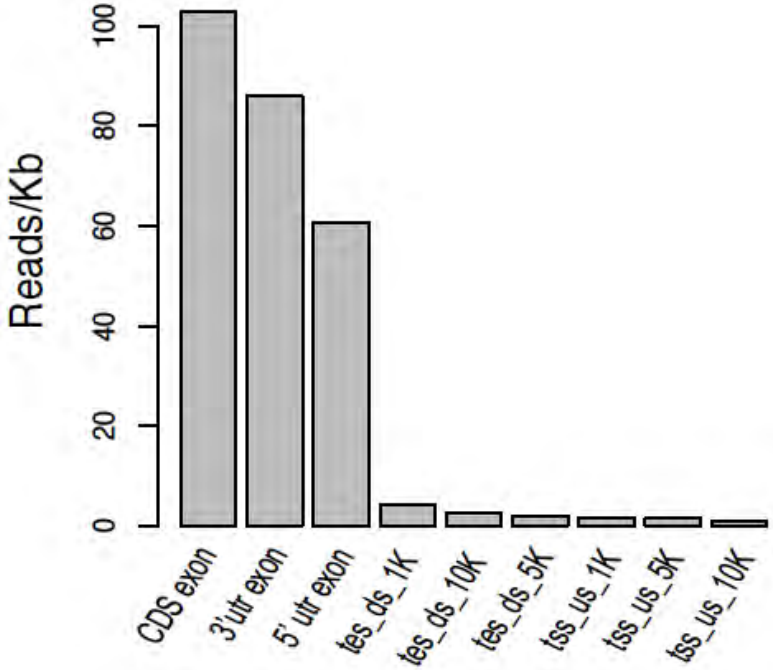
- Trim first few bases for every read

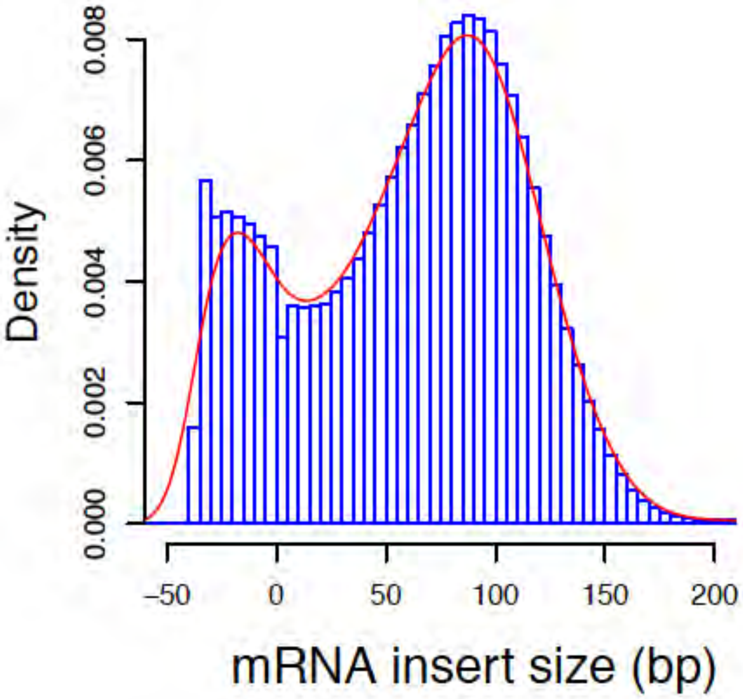
RSeQC: Insert Size and Read Distributions



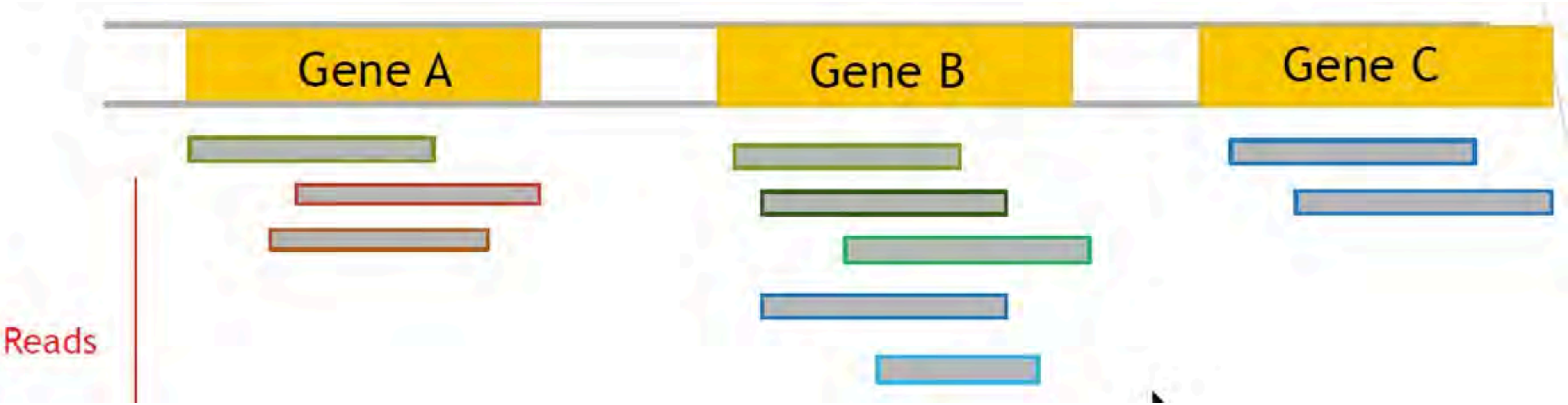


Splice junstion annotation





Mapping of Transcriptome sequencing



Transcriptome sequencing

- Comparison between treatment vs control group
- Diseased vs healthy
- Cancer vs Normal

Relative expression level: diseased vs healthy

- The Total coverage may vary across samples

Relative expression level of gene A=

$$\frac{\text{(Coverage in diseased sample in gene A)/(Total no of reads in diseased sample)}}{\text{(Coverage in healthy sample in gene A)/(Total no of reads in healthy sample)}}$$

Fold-change in expression level of gene A =

$\log_2$(relative expression level of gene A)

Fold-change in expression level

Region	Coverage in diseased	Coverage in healthy
Gene A	100	50
Gene B	100	200
Gene C	50	100
Gene D	1000	1500

Can we say gene D is more highly expressed than gene B?

Fold-change in expression level

No! Coverage also depends on gene length

Region	Coverage in diseased	Coverage in healthy
Gene A (1kb)	100	50
Gene B (1kb)	100	200
Gene C (2kb)	50	100
Gene D (10kb)	1000	1500

We need to consider gene lengths as well in addition to total coverage

RNA-seq data normalization

Region	Coverage in diseased	Coverage in healthy
Gene A (1kb)	100	50
Gene B (1kb)	100	200
Gene C (2kb)	50	100
Gene D (10kb)	1000	1500
Total	1250	1850

RPKM/FPKM normalization

- Reads per kilobase per million mapped reads
  (for single-end sequencing data)

- Fragments per kilobase per million mapped reads
  (for paired-end sequencing data)

- First normalize by total no. of reads then by gene length (in kb)

- Normalized count = (No. of reads mapping to gene A x 10^6)/
  (total no. of reads x gene length)

RPKM/FPKM: Step 1

Region	Coverage in diseased	Coverage in healthy
Gene A (1kb)	100/1250	50/1850
Gene B (1kb)	100/1250	200/1850
Gene C (2kb)	50/1250	100/1850
Gene D (10kb)	1000/1250	1500/1850
Total	1250	1850

RPKM/FPKM: Step 2

Region	Coverage in diseased	Coverage in healthy
Gene A (1kb)	100/(1250x1)	50/(1850x1)
Gene B (1kb)	100/(1250x1)	200/(1850x1)
Gene C (2kb)	50/(1250x2)	100/(1850x2)
Gene D (10kb)	1000/(1250x10)	1500/(1850x10)
Total	0.28	0.25

RPKM/FPKM: Step 2

Region	Coverage in diseased	Coverage in healthy
Gene A (1kb)	0.08	0.02
Gene B (1kb)	0.08	0.10
Gene C (2kb)	0.04	0.05
Gene D (10kb)	0.08	0.08
Total	0.28	0.25

RPKM/FPKM normalization

What is the issue?

TPM normalization

Transcripts per million mapped reads

(both for single-end sequencing and paired-end sequencing data)

First normalize by gene length then by total no. of reads

TPM: Step 1

Region	Coverage in diseased	Coverage in healthy
Gene A (1kb)	100/1	50/1
Gene B (1kb)	100/1	200/1
Gene C (2kb)	50/2	100/2
Gene D (10kb)	1000/10	1500/10
Total	325	450

TPM: Step 2

Region	Coverage in diseased	Coverage in healthy
Gene A (1kb)	100/(1x325)	50/(1x450)
Gene B (1kb)	100/(1x325)	200/(1x450)
Gene C (2kb)	50/(2x325)	100/(2x450)
Gene D (10kb)	1000/(10x325)	1500/(10x450)
Total	0.99	0.99

TPM: Step 2

Region	Coverage in diseased	Coverage in healthy
Gene A (1kb)	0.307	0.11
Gene B (1kb)	0.307	0.44
Gene C (2kb)	0.076	0.11
Gene D (10kb)	0.307	0.33
Total	0.99	0.99

Differential Gene Expression

Linear Model for Differential Expression

$Y_{ijk} = \mu_j + \alpha_{ij} + \text{error}_{ijk}$

separate model for gene i

k is a specific sample

μ_j is the mean expression for gene i over all the samples

α_{ij} is the deviation of the mean of the i^{th} condition

When comparing the samples in condition 1 (treatment) with condition 2 (control), we care whether:

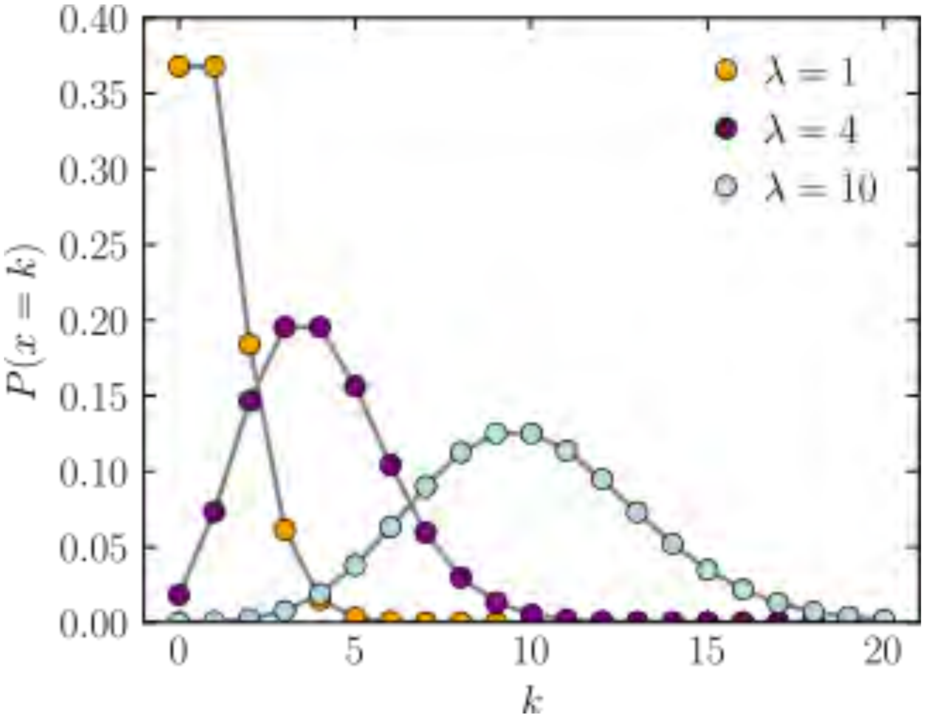
$\alpha_{\text{treatment,gene\_i}} - \alpha_{\text{control,gene\_i}} = 0$ (null hypothesis)

T tests

Sequencing Read Distribution

- The number of patients arriving in an emergency room between 10 and 11PM

- No of reads mapped to a gene with 3KB effective length



- Poisson distribution

$$P(k \text{ events in interval}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- λ average events per interval
- K events in an interval
- Variance = mean = λ

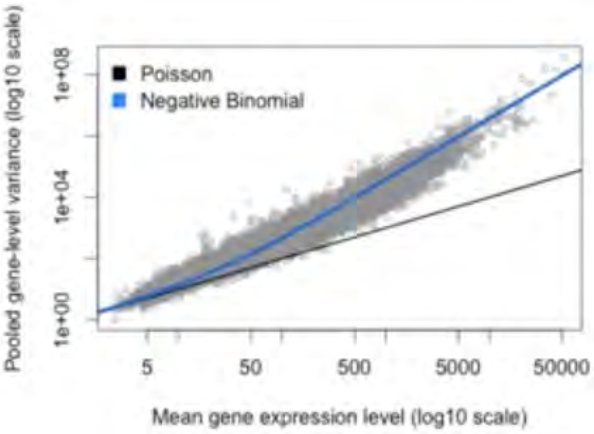
School of Energy Science & Engineering

Sequencing Read Distribution

- In reality, sequencing data is over-dispersed
  - (Mean < Variance)

Mean: $\frac{pr}{1-p}$

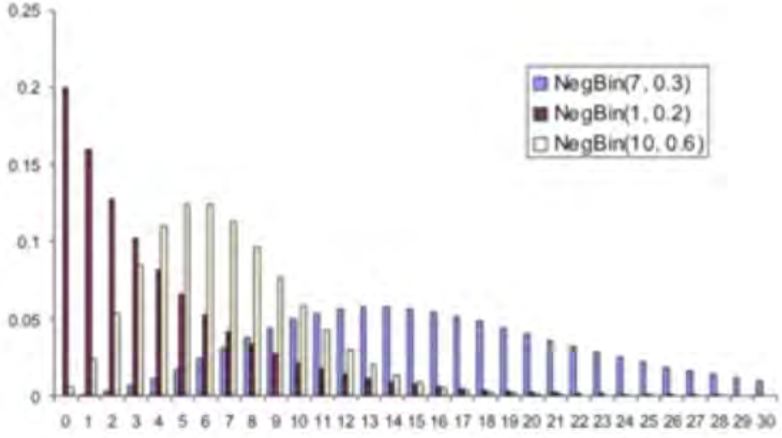
Variance: $\frac{pr}{(1-p)^2}$



- Negative bionomial
  - NB(r,p)
  - If Pb(succ) is o, no of success before the first r no of failure,

- Probability mass function

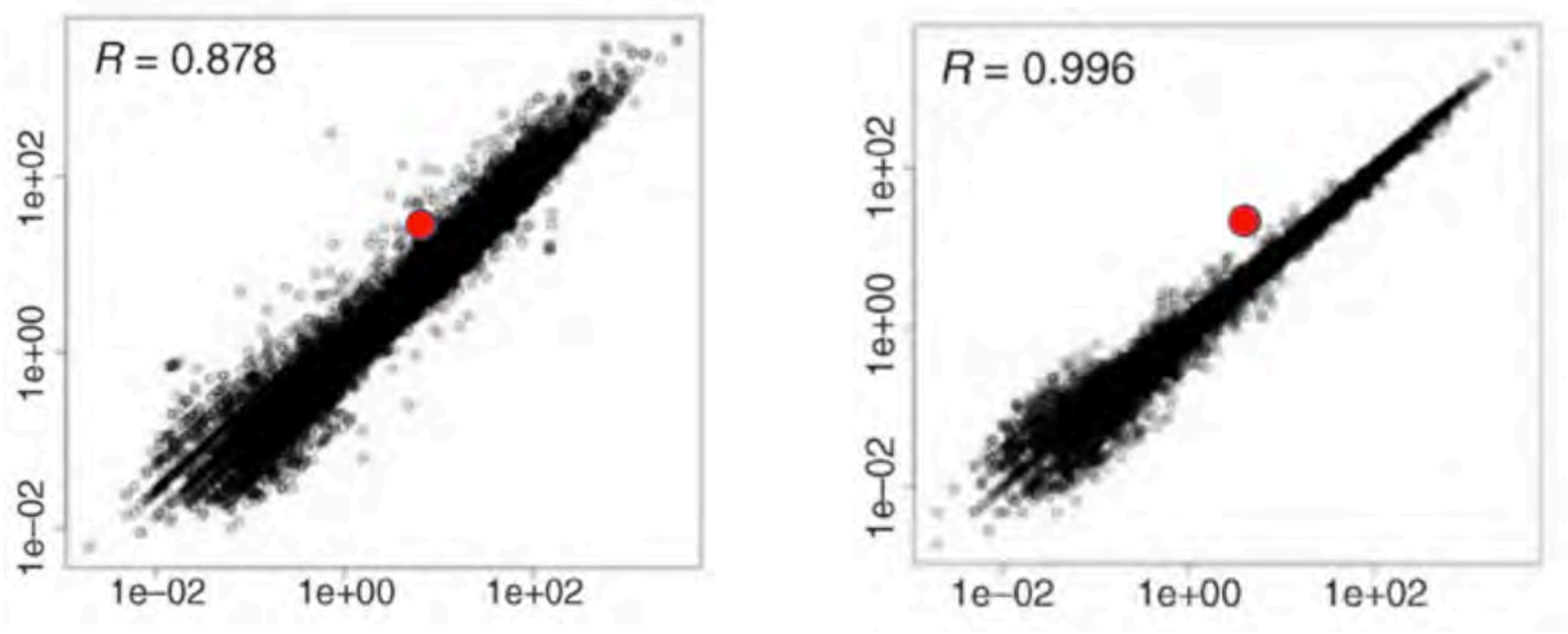
$$k \mapsto \binom{k+r-1}{k} \cdot (1-p)^r p^k,$$



Model variance from Limited Replicates

- Problem with estimating variance when the sample size is small (2-3 replicates in each condition)

- By looking at other genes



DESeq2: Modeling Over Dispersion

raw count for gene i, sample j

normalization factor (depth, gene length, etc)

expr level of interest

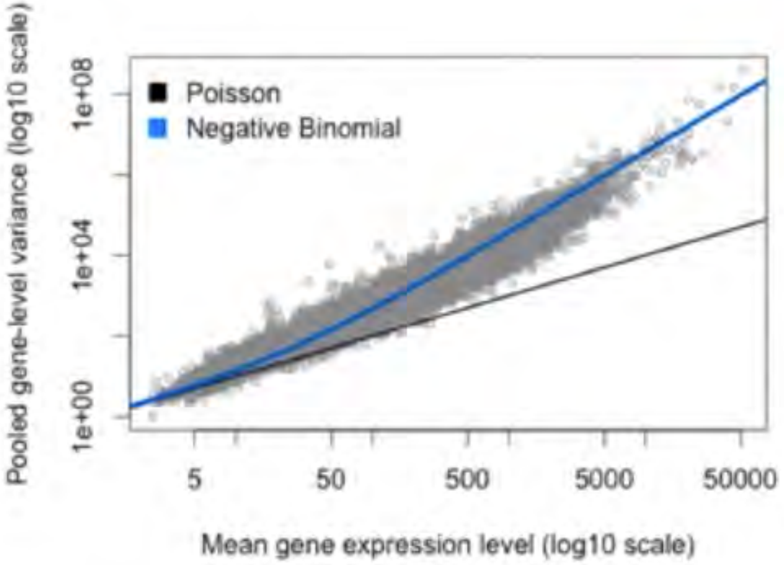
dispersion for gene i

$$K_{ij} \sim \mathrm{NB}(s_{ij} q_{ij}, \alpha_i)$$

$$\mathrm{Var}(K_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2$$

Poisson from sampling fragments

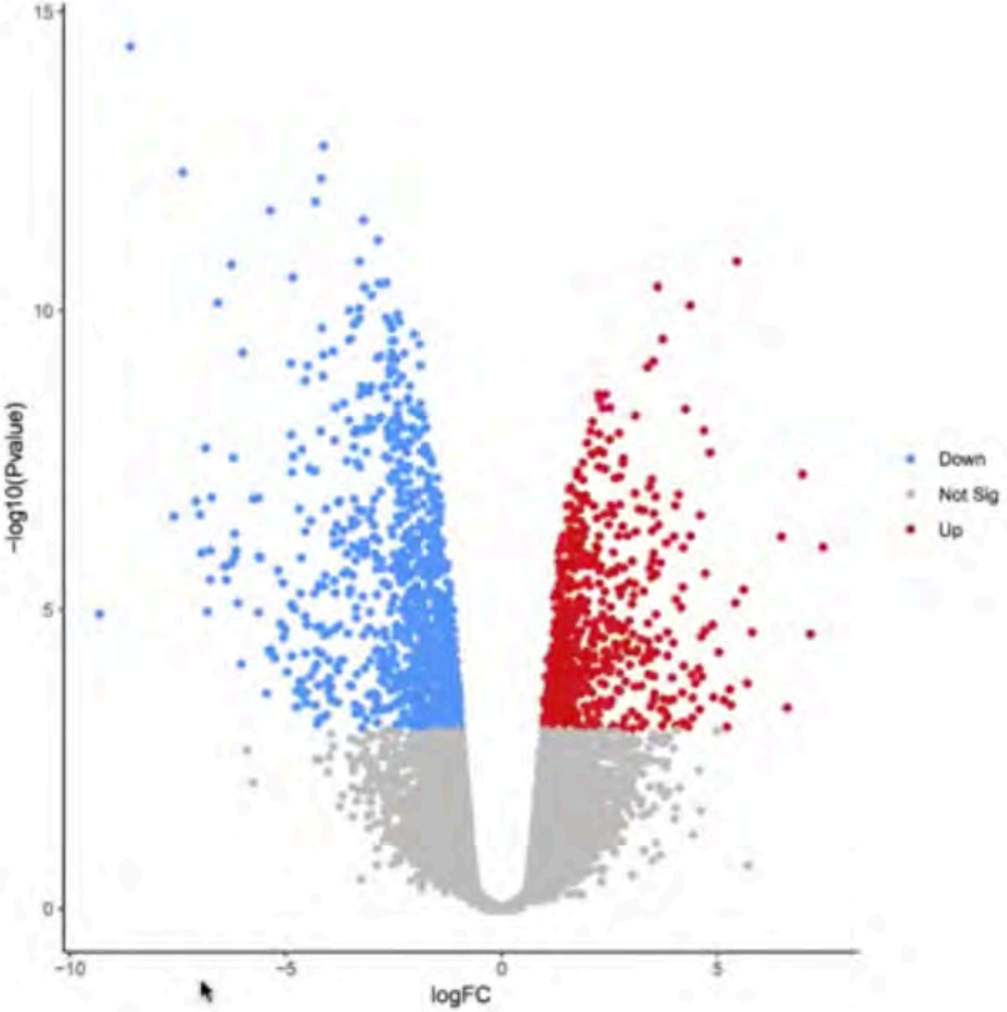
Extra variation due to biological variance



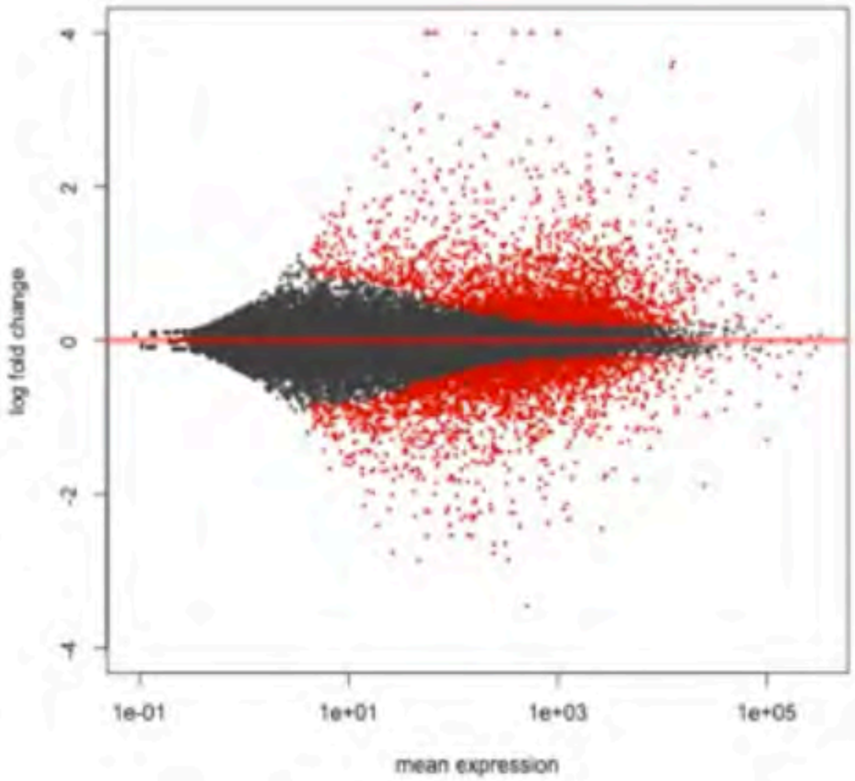
Dispersion for gene i is associated with its expression level estimated by borrowing info from all gene

Visualize Differential Expression

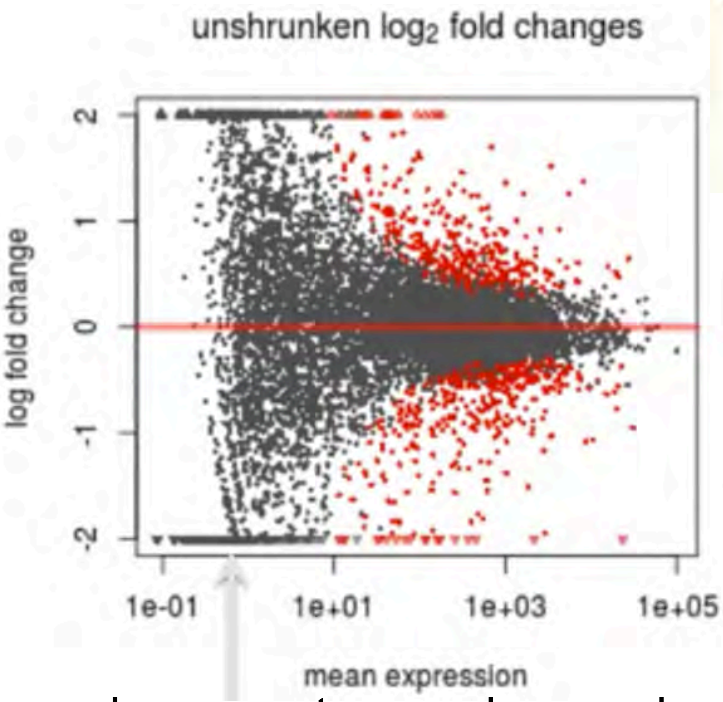
Volcano Plot

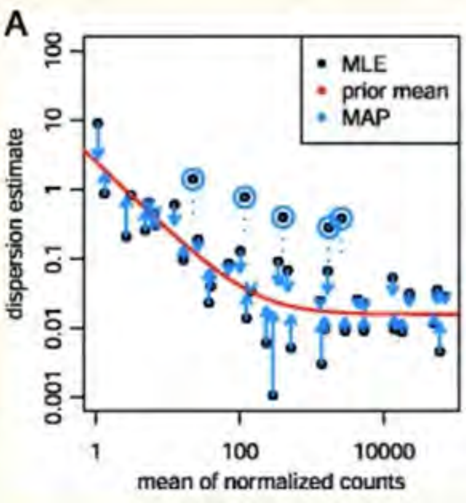


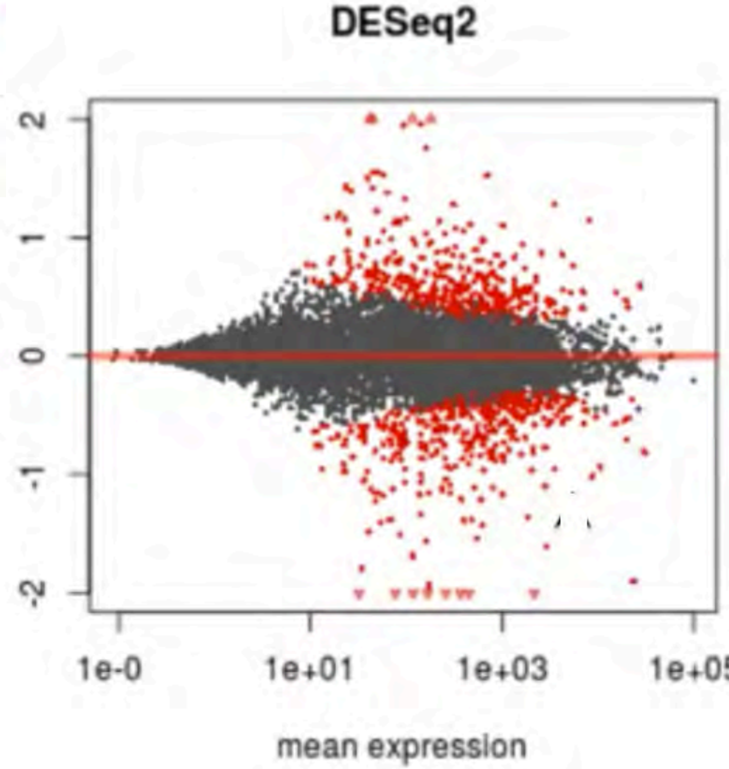
MA Plot



Fold Change with Variance Stabilization







Low count genes have noisy variance and Fold Change estimates

Low count genes need strong variance moderation

Testing for difference in gene expression

Diseased vs healthy samples or Condition 1 vs Condition 2

- Hypothesis testing: Null hypothesis (H_0)
- No difference in expression between diseased and healthy samples
- Statistical test among replicates of diseased vs healthy samples
- Level of significance (p-value): if $p<0.05$, then we can reject null hypothesis
- if $p>0.05$, then we do not have enough evidence to reject null
- Alternative hypothesis (H1):
- Expression of gene 'i' is different between diseased and healthy samples

Multiple hypothesis Testing

We test differential expression for every gene with p-value, e.g. 0.01

H_0: no difference in gene expression; H_1: difference in expression

Reject H_0: call something to be differential expressed

For 20,000 genes in the genome:

potentially 0.01 x 20K = 200 genes wrongly called

Family-wise error rate or False discovery rate

Family-Wise Error Rate

P(false rejection at most one hypothesis) < α

P(no false rejection < 1 - α

Reject H_0: call something to be differential expressed

Bonferroni correction: to control the family-wise error rate for testing

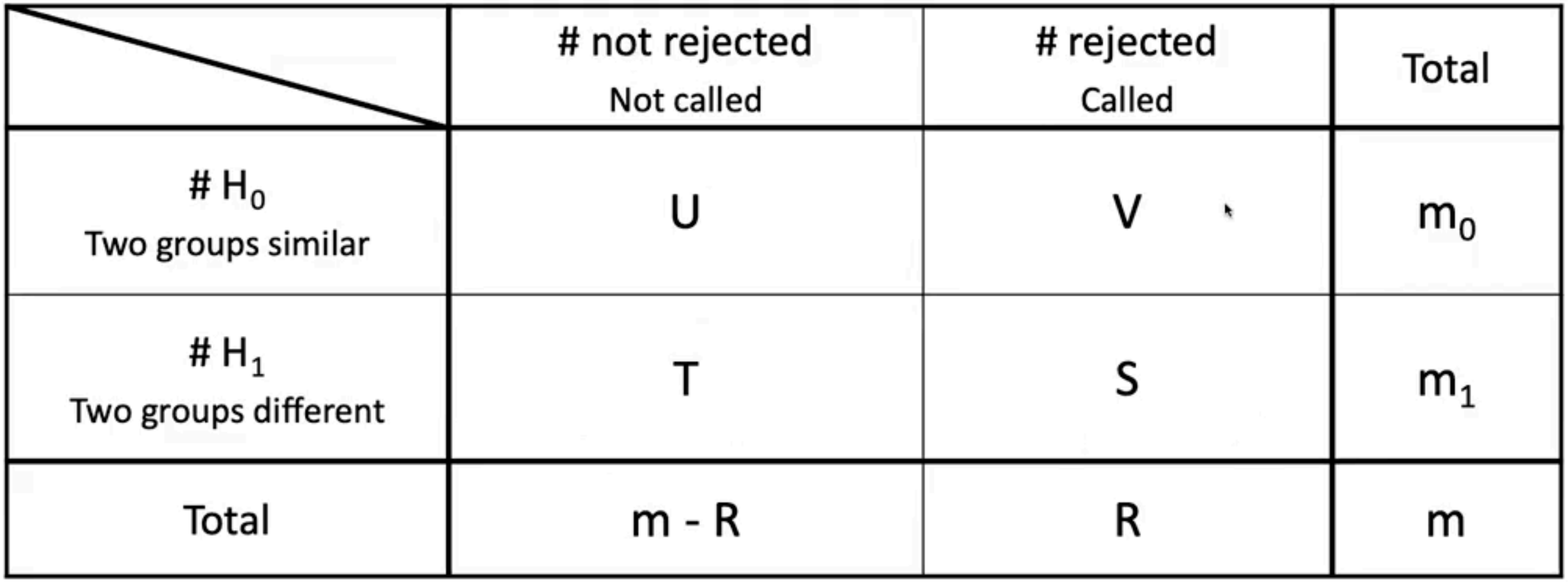
M hypothesis at level α, we need to control the false rejection rate for

each individual at α/m

For α is 0.05, for 20K genes prediction:

p-value cutoff is 0.05/20K = 2.5×10^{-6}

False Discovery Rate



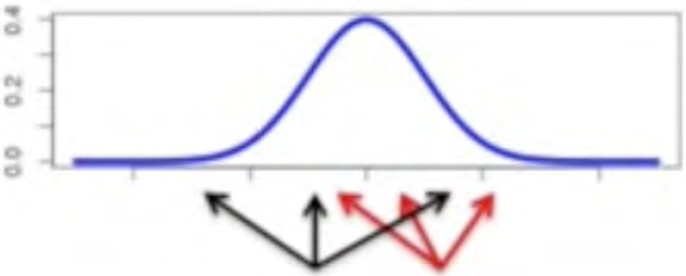
	# not rejected Not called	# rejected Called	Total
# H_0 Two groups similar	U	V	m_0
# H_1 Two groups different	T	S	m_1
Total	m - R	R	m

V: Type I errors, False Positives
T: Type II errors, False negatives
FDR = V/R

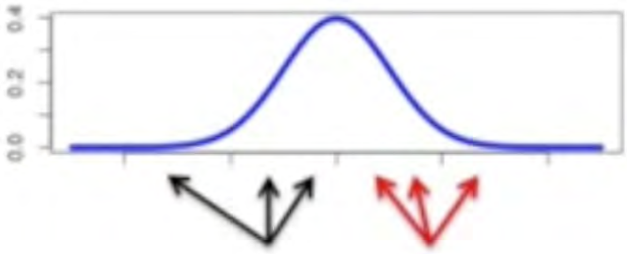
Measuring gene expression: Type I errors

Normally false positive are

95% of the time the samples will overlap



5% of the time they don’t



For 20,000 genes in the genome:

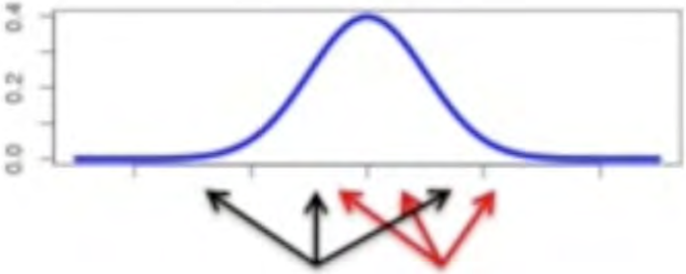
If we take two healthy samples and compared all 20,000 genes

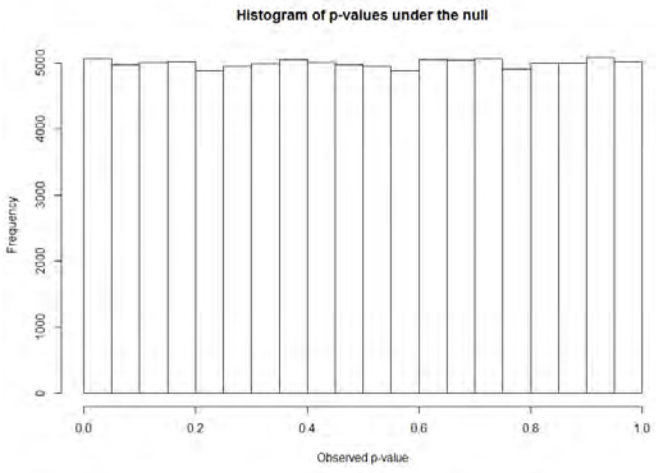
5% of 20,000 = 1000 false positives

Type I errors

p-values generated by testing samples from the same distribution

When samples come from same distribution the p-values are uniformly distributed



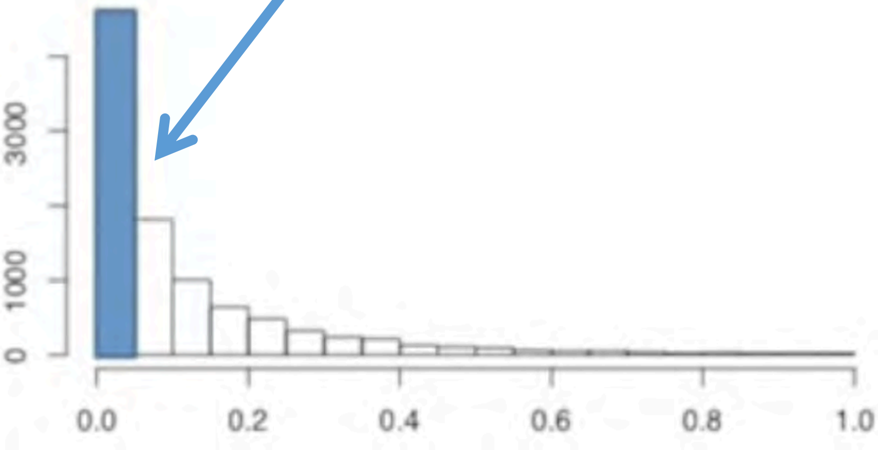


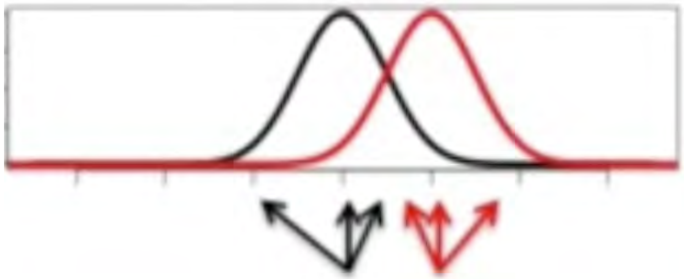
A histogram of 20,000 p-values generated by testing samples from the same distribution

p-values generated by testing samples from the two different distribution

Most of the p-values < 0.05 when the samples are not overlapped

The no of p-values

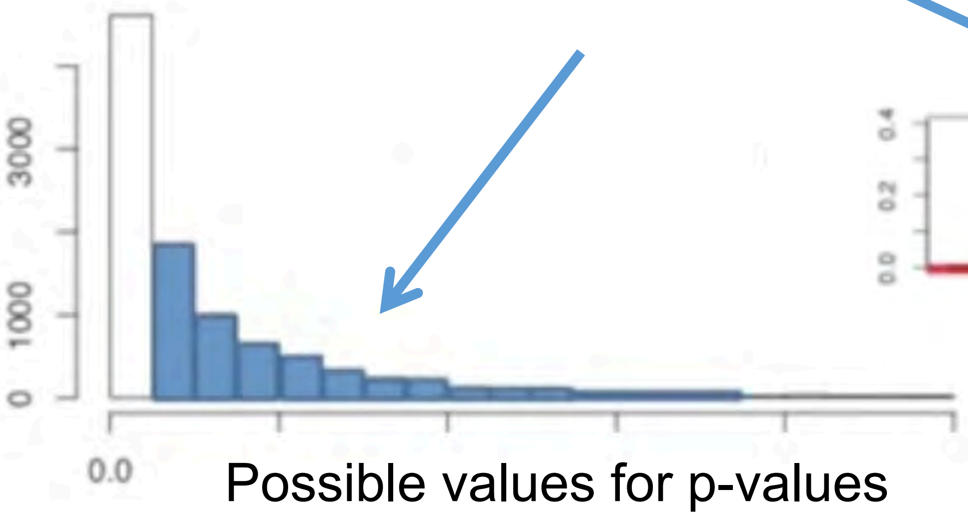




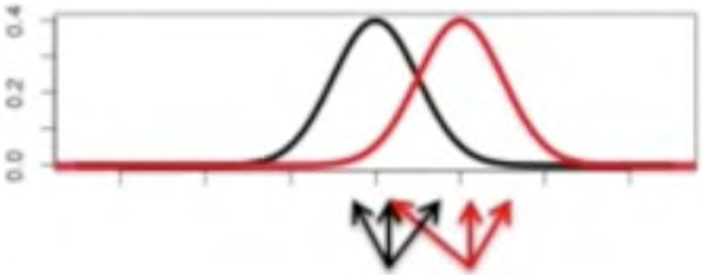
Now lets look at how p-values from two different distribution

The p-values > 0.05 are the false negatives from when the samples are overlapped

The no of p-values



Possible values for p-values



Example

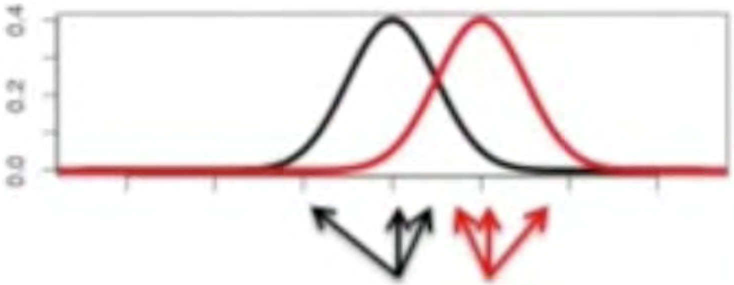
Now lets look at how p-values are distributed when they come from two different distribution

Now, imagine we are doing experiment where we are testing all the active genes in neuronal cells

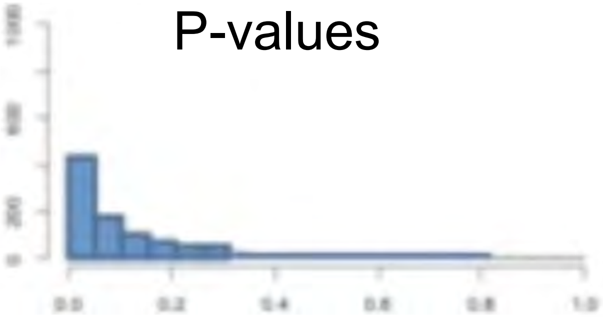
One set of neuronal cell is treated with a drug, the other is not

The histogram of p-values we obtain from all 20,000 genes is the sum of the two separate histograms

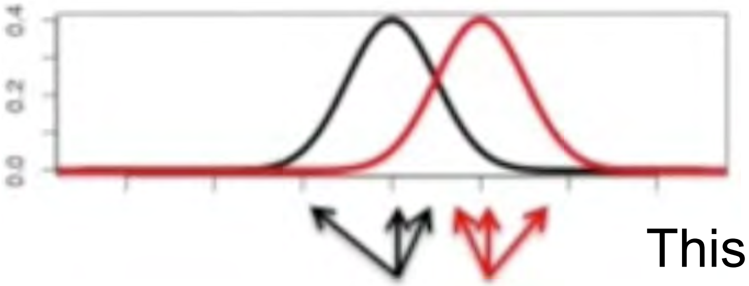
The drug might affect 2,000 genes

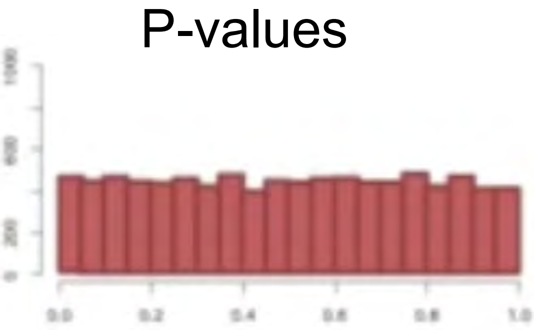


The measurements of these will come from two different distributions


P-values

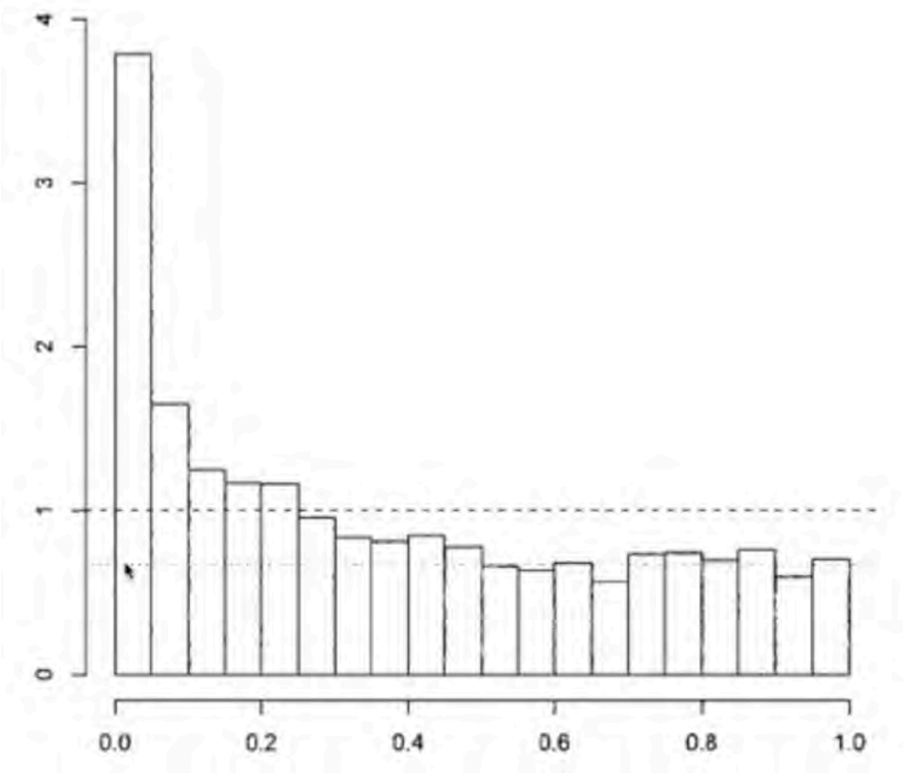
The drug might affect 18,000 genes might not affected by drugs

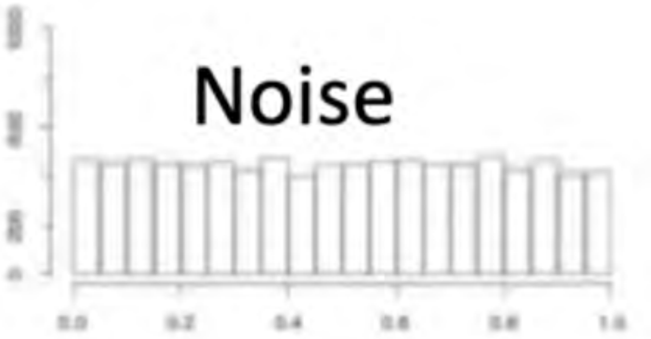


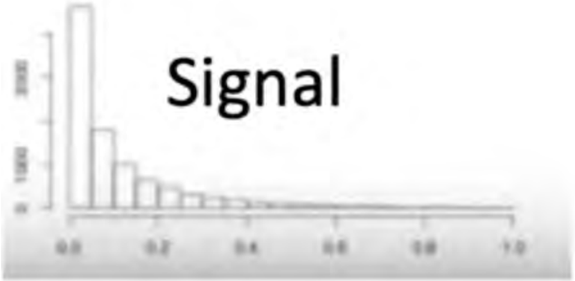

P-values

This means the measurement for most of the genes will come from the same distribution

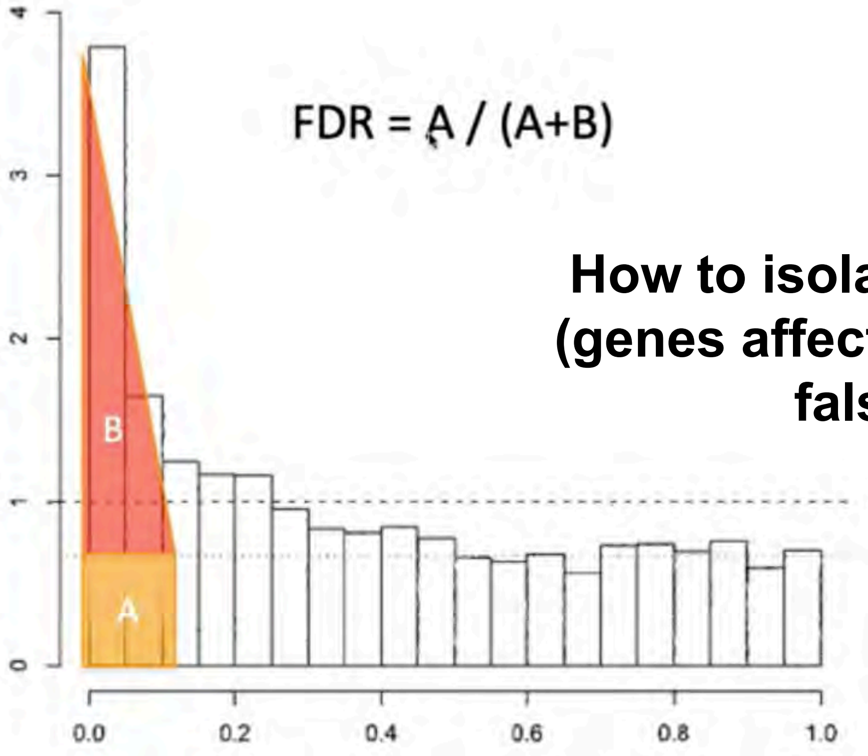
The histogram of p-values we obtain from all 20,000 genes is the sum of the two separate histograms

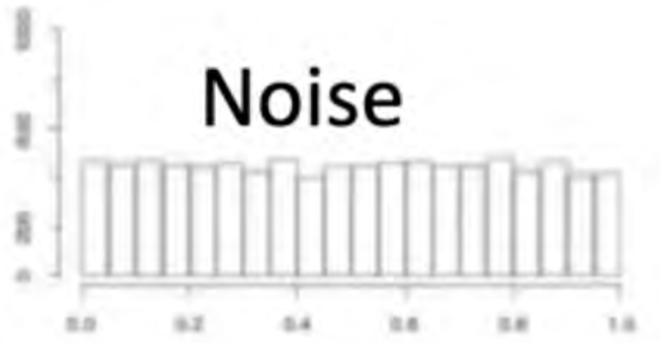




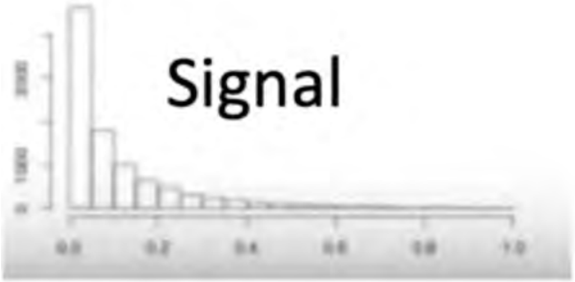


The histogram of p-values we obtain from all 20,000 genes is the sum of the two separate histograms





How to isolate the true positives (genes affected by drug) from the false positives



The False Discovery Rate(FDR) can control the number of false positive

The Benjamini-Hochberg method

It adjusts p-values in a way that limits the number of false positives that are reported as significant

Adjusted p-values means that it makes them larger

For example, before the FDR correction, your p-value might be 0.04 (significant)

After the FDR correction, your p-value be 0.06 (no longer significant)

False Discovery Rate

- Benjamini-Hochberg method: FDR, adjusted p-values
- Every p-values has its corresponding FDR (larger than p-values)
- Common FDR threshold: 1%, 5%, 10%, also sometimes filtered by fold change (1.2, 1.5, 2 fold change) to estimate signal/noise of hits
- Algorithm such as DESeq2 will ignore genes with too low expression to reduce the total number of hypotheses to test, in order to be more sensitive in finding real differentially expressed genes
- For expression, most people are comfortable with few hundred differentially expressed genes

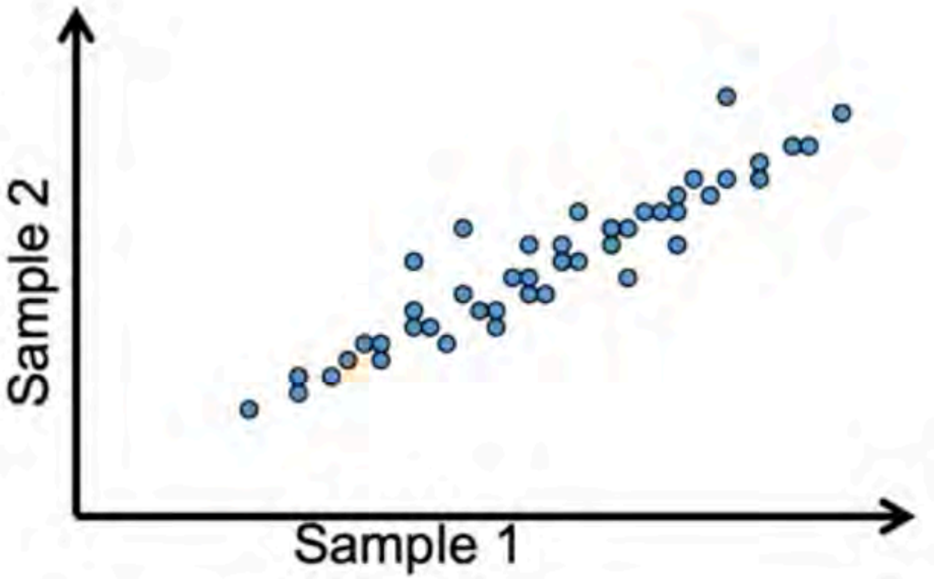
Normalization and Batch Effect Removal

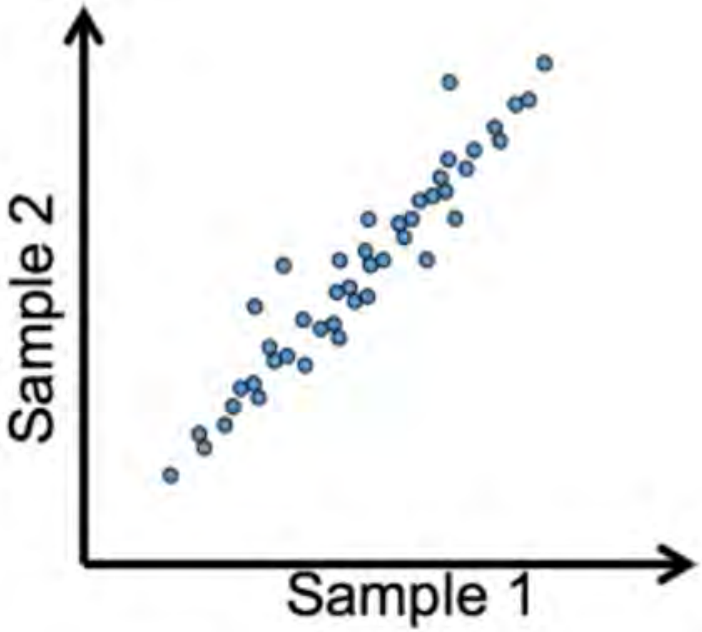
Median Scaling

Linear scaling

Ensure the different arrays have the same median value and same dynamic range

$$X' = (X-c_1)*c_2$$





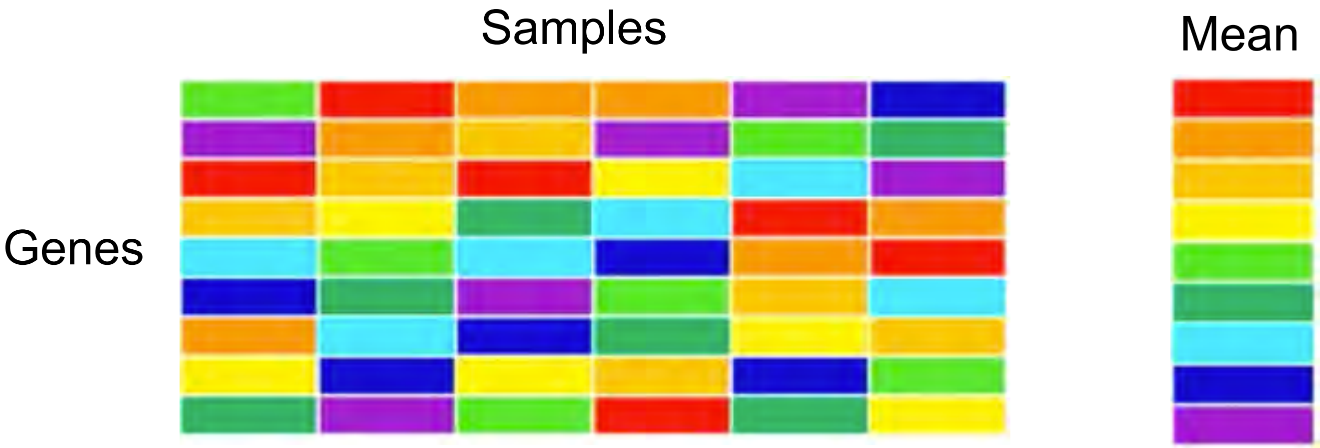
Quantile Normalization

- Imagine we have some data from a microarray experiment
- Microarrays measure how active genes are in a sample
- They do that by measuring the intensity of different colors of light
- If you have better light bulb for one experiment, every measurement might be brighter than every measurement from another experiment

Quantile Normalization

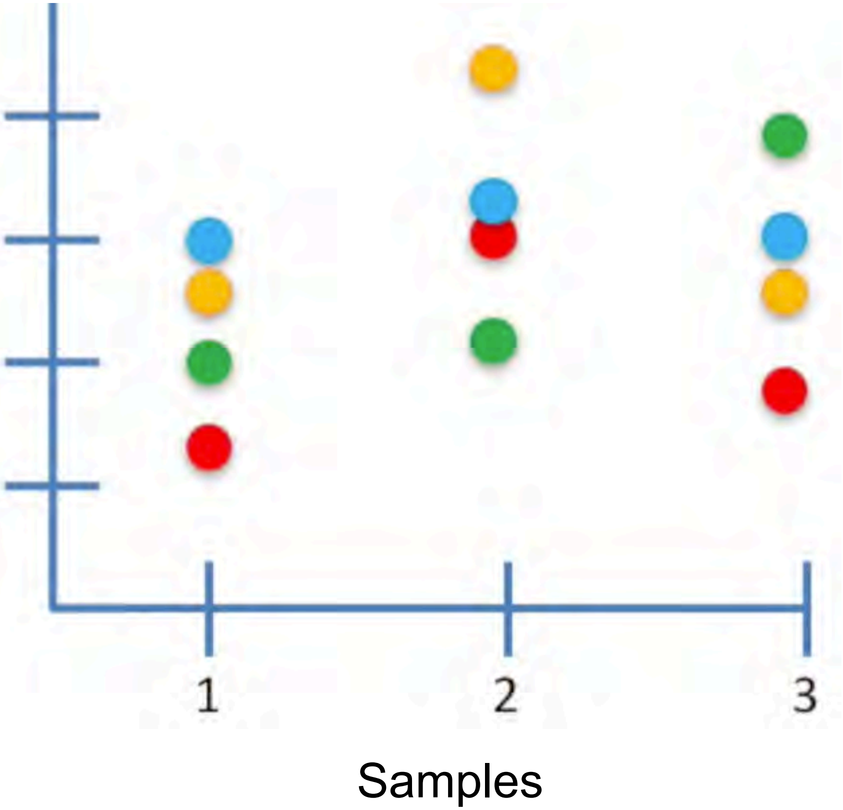
- Calculate mean for each quantile and reassign each probe by the mean quantile

- No experiment retain original value, but all experiments have exact same distribution

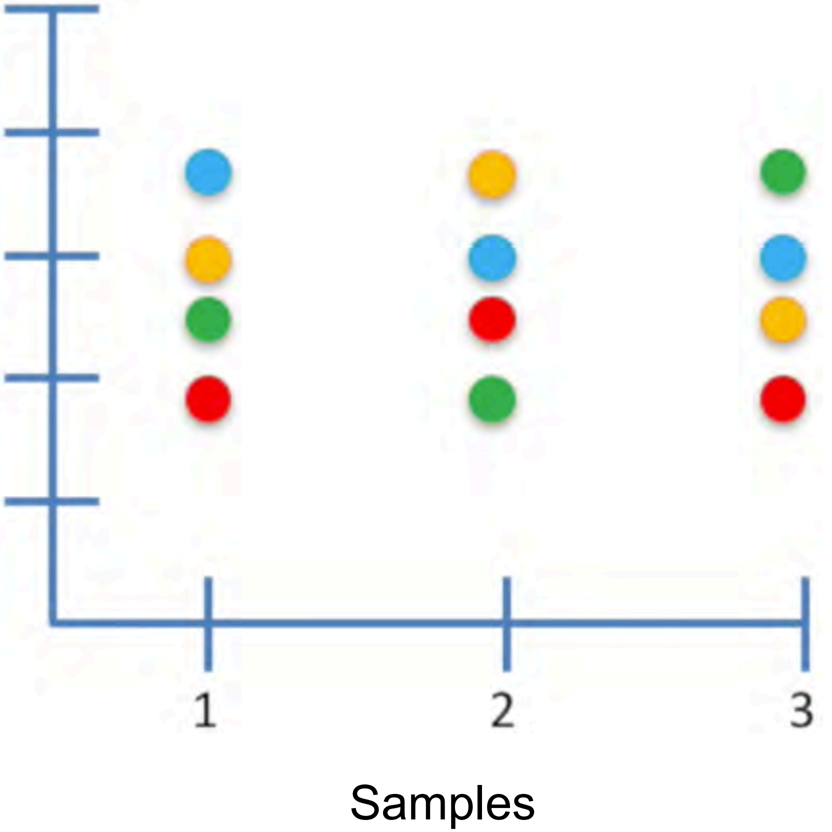


Quantile Normalization

Raw Gene Expression Data



Quantile Normalized Data

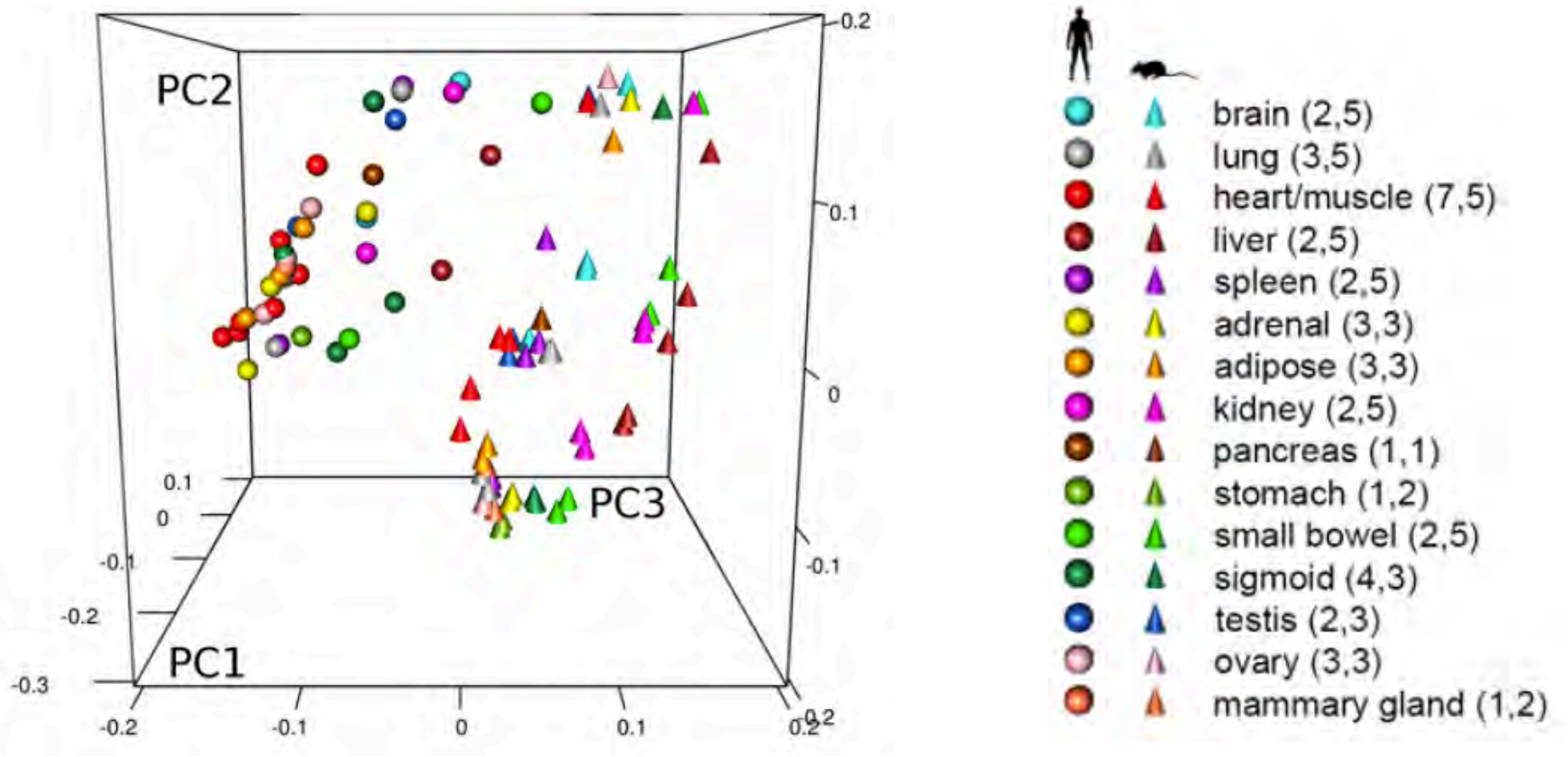


Motivation

- Batch Effect: Non-biological variation
- Caused by differences:
  - Different day/months of the experiments
  - Different reagents (enzymes, buffers)
  - Different mice (from different companies)
  - Different sequencers
  - Lab protocol of experimenter

Batch Effect Example

- Striking finding in 2014: Human heart is more similar with human than mouse brain



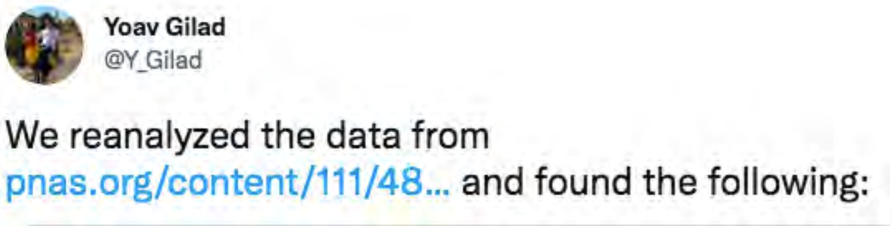
Lin *et a*l PNAS (2014)

Batch Effect Example

- Batch Effect: Non-biological variation

- Caused by differences:
  - Different day/months of the experiments
  - Different reagents (enzymes, buffers)
  - Different mice (from different companies)
  - Different sequencers
  - Lab protocol of experimenter

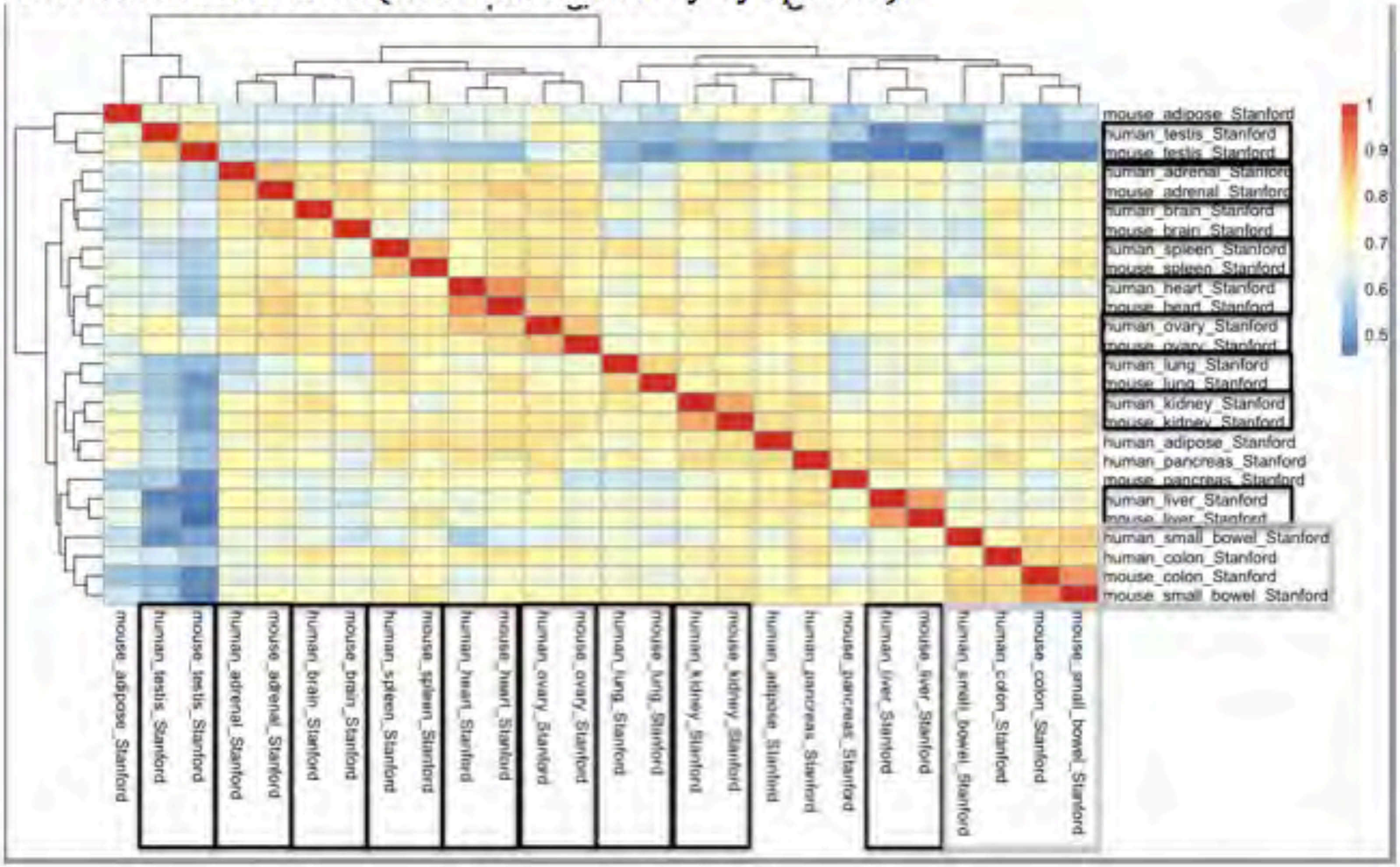
Batch Effect Example



Yoav Gilad
@Y_Gilad

We reanalyzed the data from pnas.org/content/111/48... and found the following:

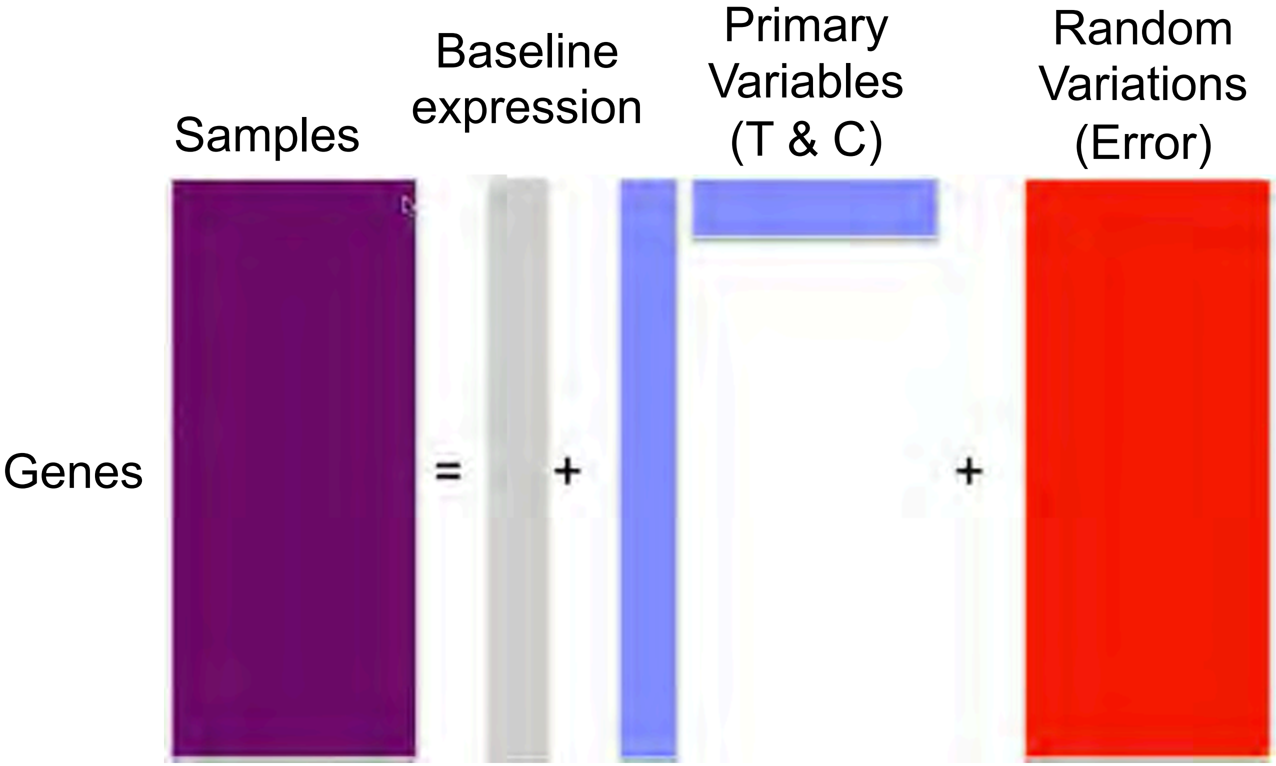
After batch correction (data cluster mostly by species):



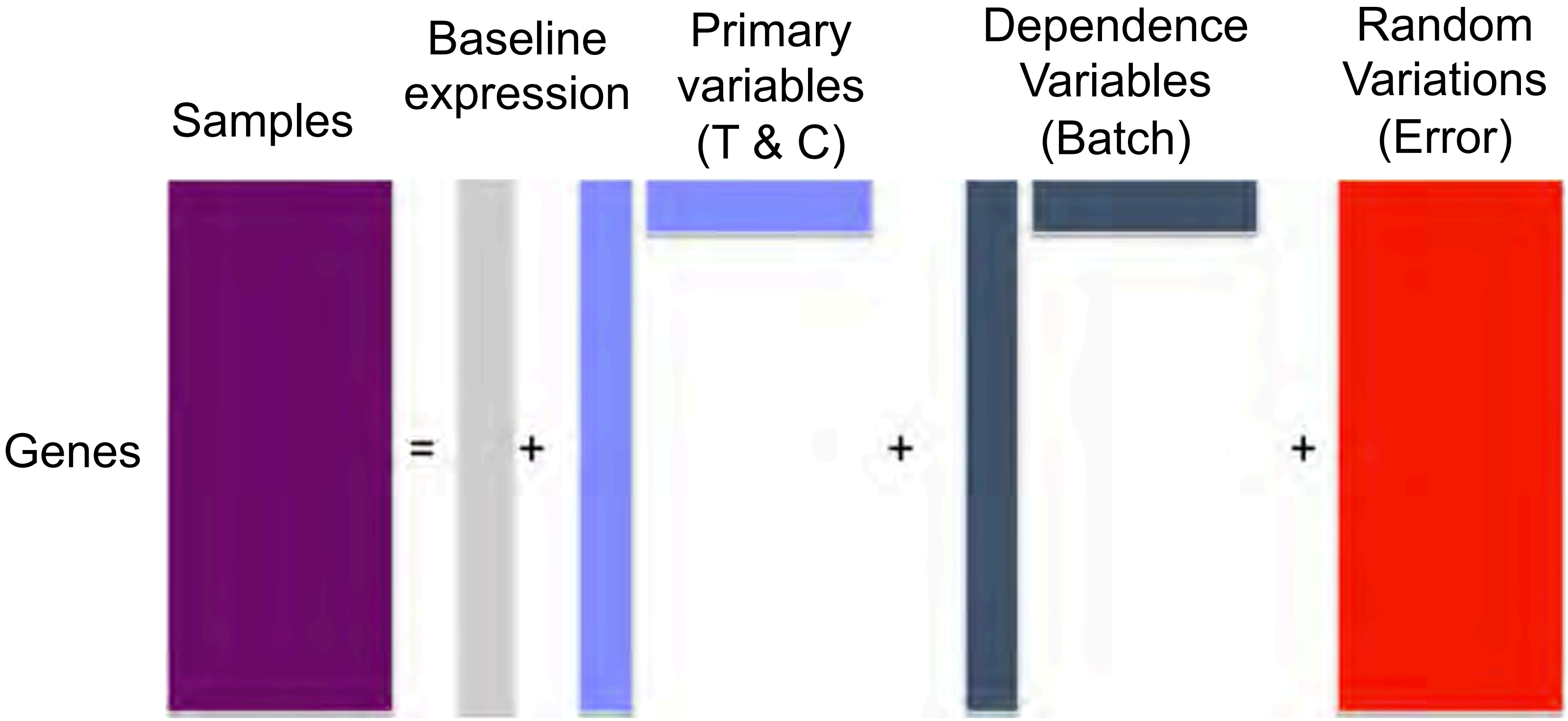
Importance of Experimental Design

- Better technology (e.g. sequencing vs microarrays) never eliminate the needs for good experimental design
- Try to be consistent and process all samples at the same time
- Record run date, labs, personnel, environmental variables, etc.
- When possible, balance groups of interests, at least include some controls in each batch
- Avoid perfect confounding when batch and group are perfectly correlated e.g.
- Treatments in one batch, control in another
- Treatment 1 in batch 1, Treatment 2 in batch 2

Decomposing Data Heterogeneity



Decomposing Data Heterogeneity


Samples
Baseline expression
Primary variables (T & C)
Dependence Variables (Batch)
Random Variations (Error)
Genes
=
+
+
+

Intuitively, consider batch a some kind of treatment effect

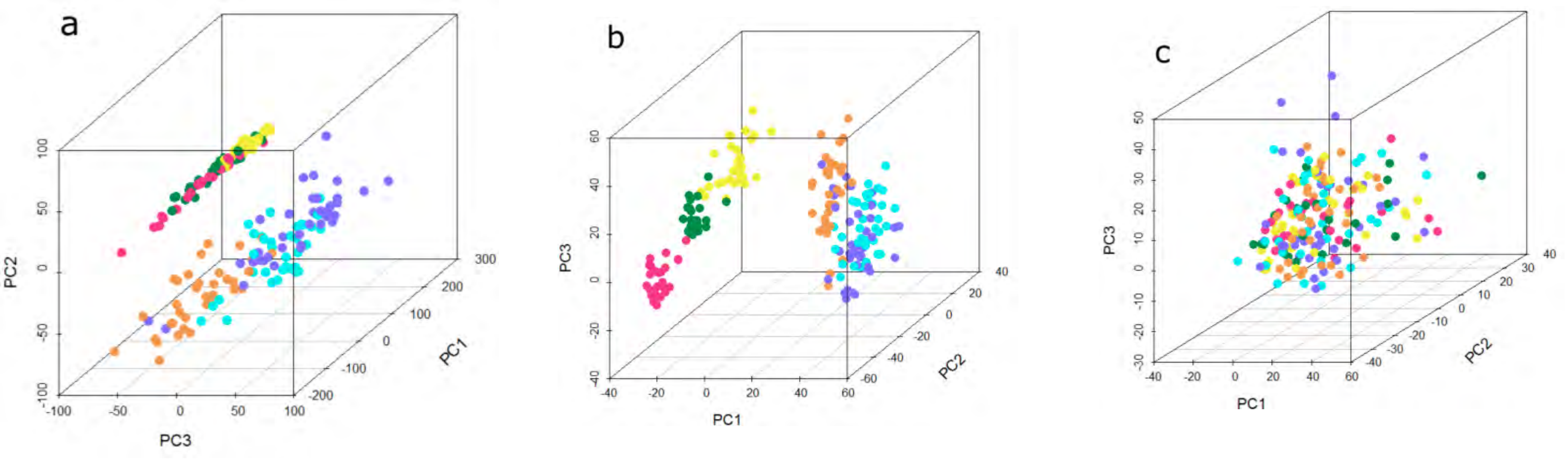
PCA for Batch Effect Detection

- PCA can be used to visualize and identify batch effect
- Obvious batch effect: early PC's separate samples by batch

Un-normalized

Quantile Normalization

COMBAT



Brezina *et a*l Microarrays (2015)

Identify and Correct Batch Effect

- Identify:
- Cluster/visualize the samples on log(TPM)
- Revome:
  - Combat (Johnson *et al* Biostatistics 2007) for simple batches
  - LIMA for complex (e.g. nested) batches
  - On log(TPM) values
- Check again:

  Clustering again on the batch corrected log(TPM)
- DESeq2 and consider batch as another variable (condition, simple batch) without touching the original expression index files