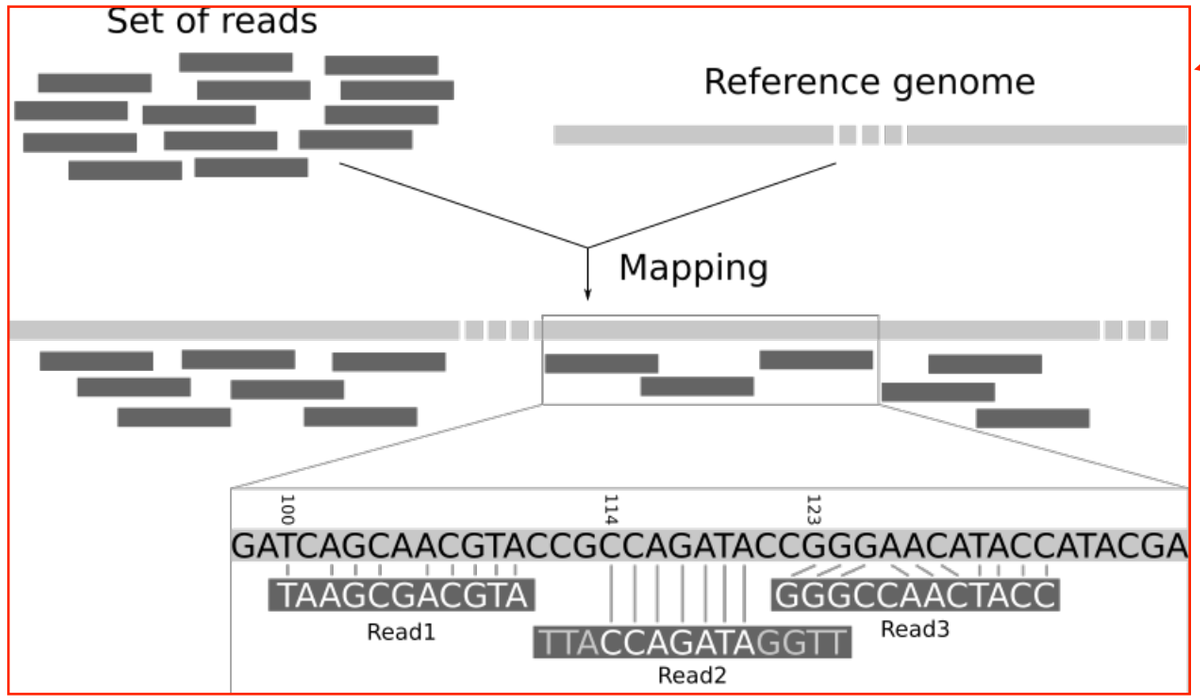
Basic concepts in Differential expression analysis

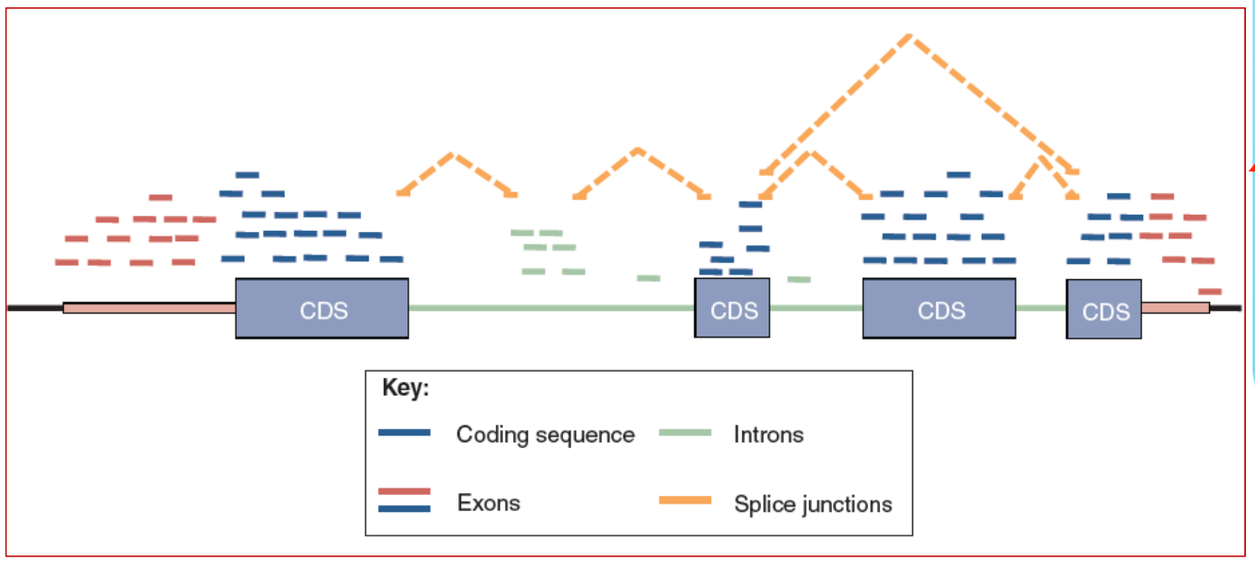
Mengjie Chen

Jesse is a first-year graduate student in GGSB, who has joined Gustav lab one week ago for a lab rotation. Gustav lab use HepG2 cell line to study the function of gene METTL3. Jesse was given an RNA-seq dataset with raw sequencing data from 10 METTL3 WT and KO samples. The goal is to compare these two conditions and find differentially expressed genes. Jesse has no clue about where to start. He talked to his postdoc supervisor, Marie, to get more information.

How do people process RNA-seq data nowadays?

There are implemented bioinformatics pipelines for RNA-seq. You can select one and run it on your samples. A pipeline is constructed by a sequence of different software and computational tools. For RNA-seq, we need to perform alignment, quality control, quantification and etc. Each step involves a choice of a method from many existing alternatives. My suggestion will be to use the pipeline maintained by our own lab, since we have tested it intensively and tuned some parameters. We also have samples processed by the same pipeline to benchmark with if necessary. Here is our pipeline.





Quality control

Inputs

Raw sequence data | Annotations | Reference genome

Pre-processing

Trimming | Indexing

Alignment

Tophat | Star | Others

Transcript assembly

Reference only | Reference guided: Cufflinks → Cuffmerge | *de novo*: Cufflinks → Cuffmerge

Abundance estimation

Count-based: Cuffquant | htseq -count

FPKM-based: Cufflinks

Expression analysis

Differential expression: CuffDiff | EdgeR

Alternative expression: CuffDiff | DEXSeq

Vizualization

Cumberbund | IGV | R

What are the next steps after alignment?

After alignment, we want to quantify gene expression levels for each sample and prepare for the cross-sample comparisons. The input of analysis will be several BAM files and the output will be a summary table about expression levels.

From BAM files to gene expression levels

A SAM file (.sam) is a tab-delimited text file that contains sequence alignment data.
A BAM file (.bam) is the binary version of a SAM file.

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                         CAGCGGCAT
```

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

http://samtools.github.io/hts-specs/SAMv1.pdf

From BAM files to gene expression levels

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001    99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002     0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003     0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001   147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

	# of mappable reads
Gene 1	10
Gene 2	50
Gene 3	37

	# of mappable reads
Exon 1	14
Exon 2	5
Exon 3	27

	# of mappable reads
Transcript 1	8
Transcript 2	15
Transcript 3	17

http://samtools.github.io/hts-specs/SAMv1.pdf

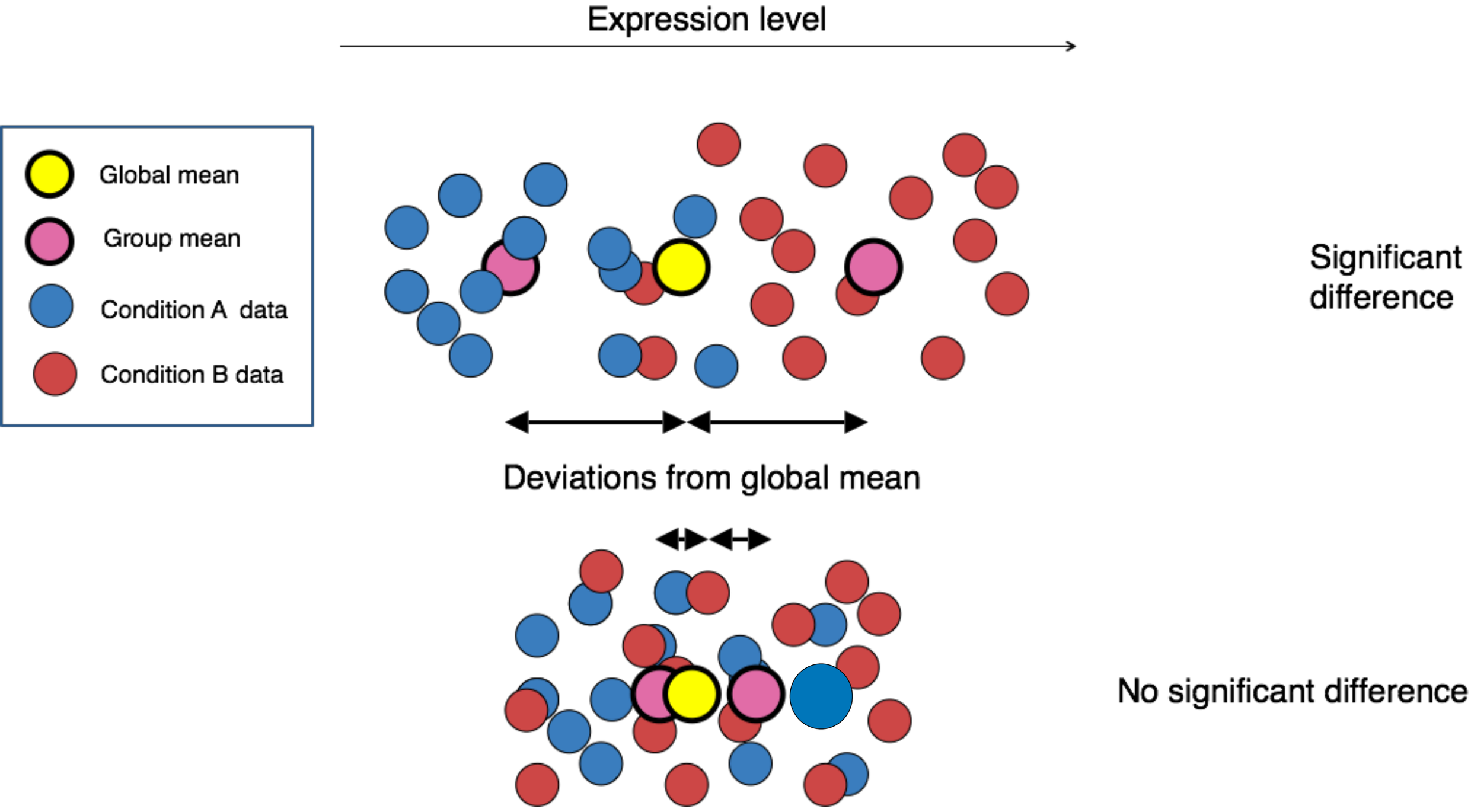
What does differential expression mean?

What is differential expression analysis?

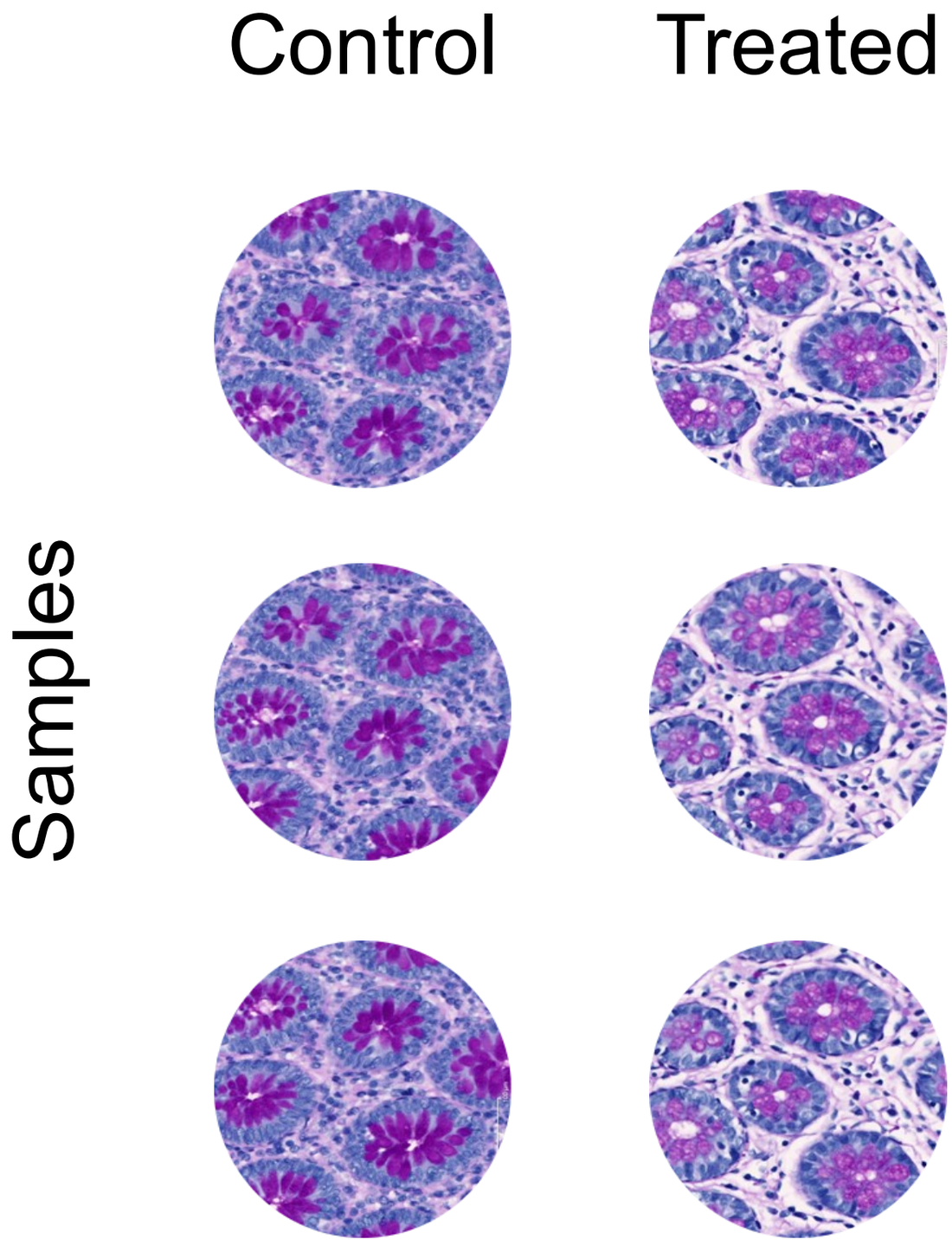
Through the process of differential gene expression, the activation of different genes within a cell that define its purpose, each cell expresses only those genes which it needs.

Differential expression analysis means taking the normalised read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups. For example, we use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.

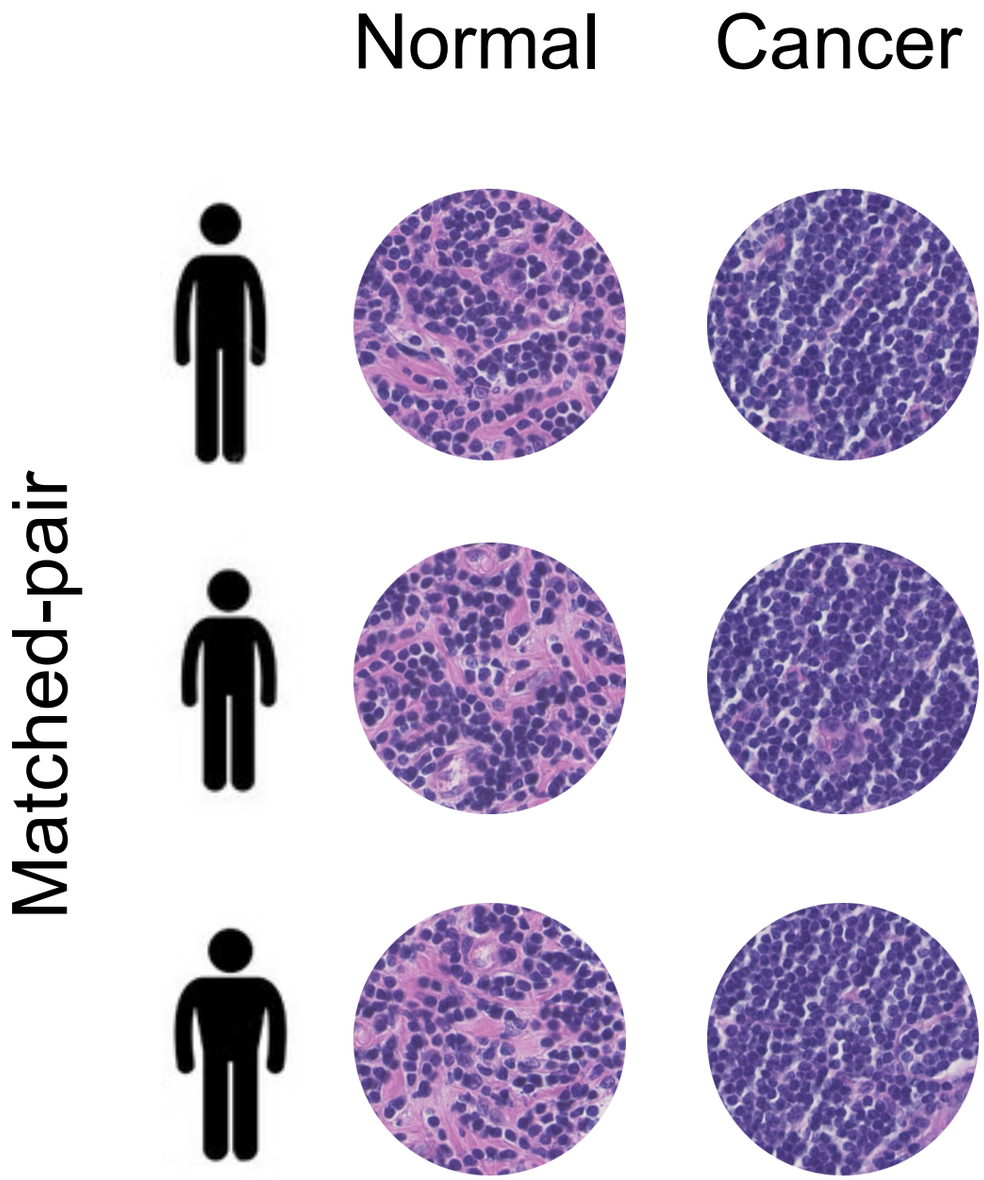
Whether gene g is differentially expressed between condition A and B?



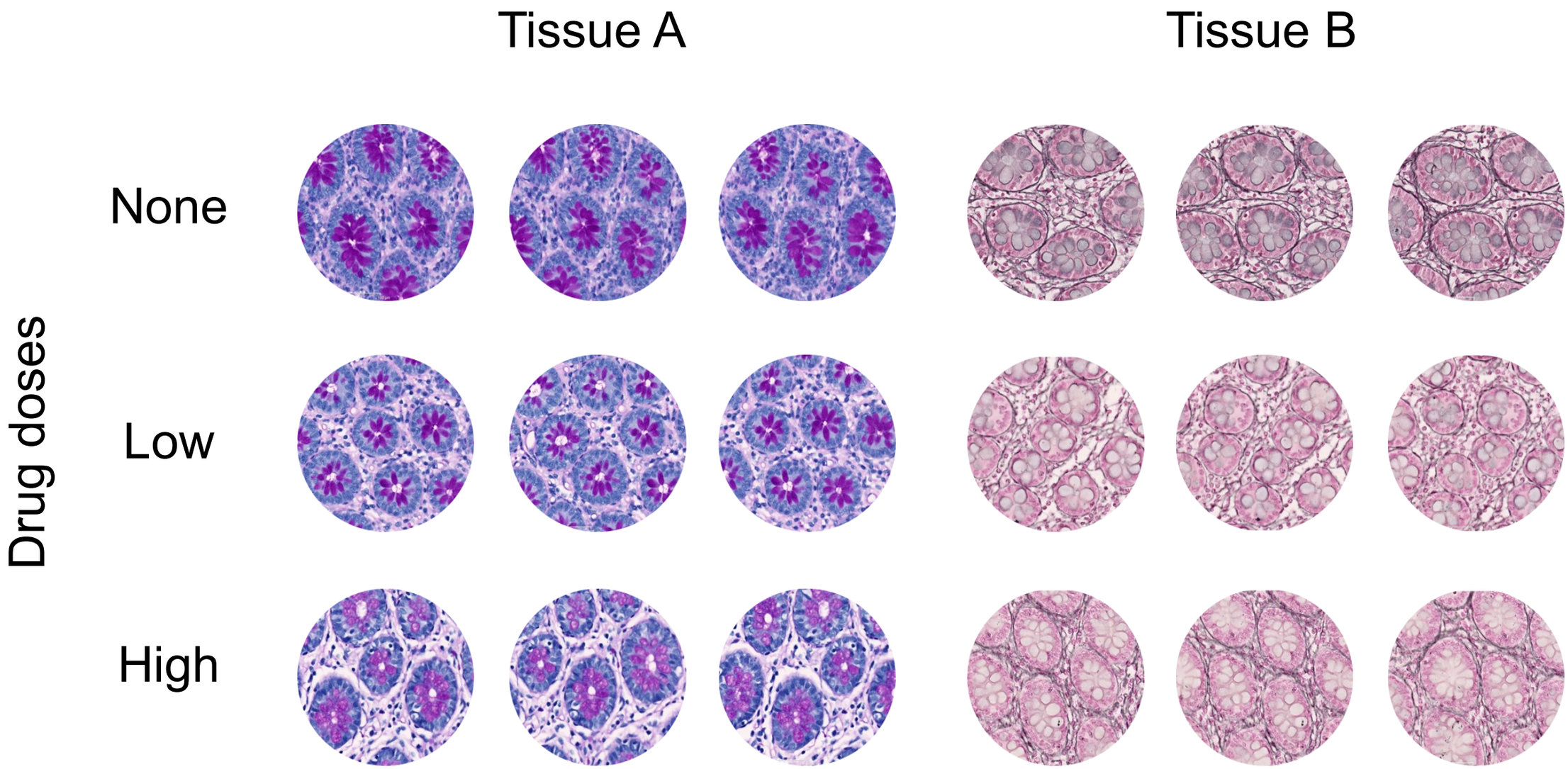
Study designs that support DE analysis

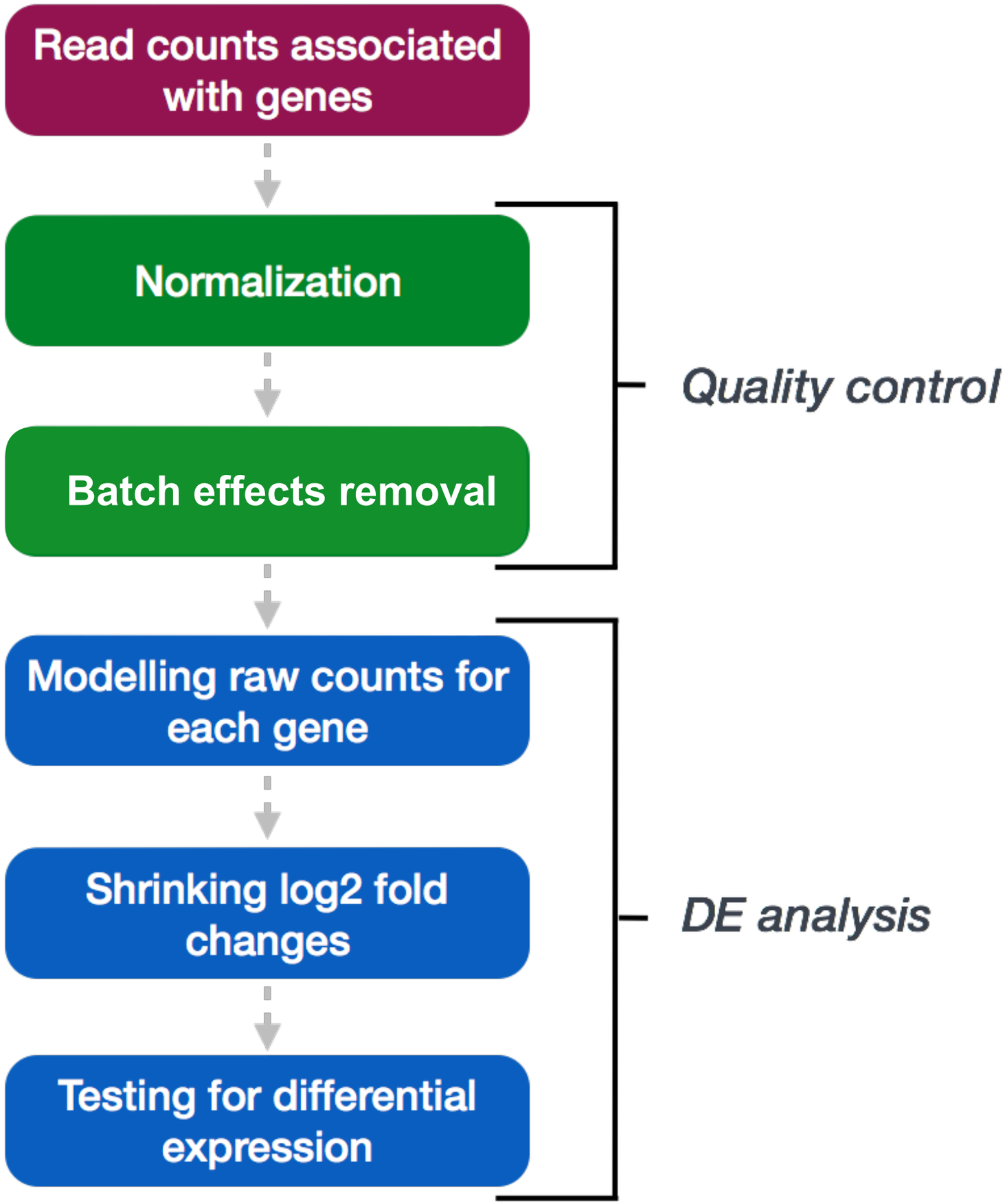


Study designs that support DE analysis



Study designs that support DE analysis




Read counts associated with genes
Normalization
Batch effects removal
Quality control
Modelling raw counts for each gene
Shrinking log2 fold changes
Testing for differential expression
DE analysis

What are the issues to consider in quantification?

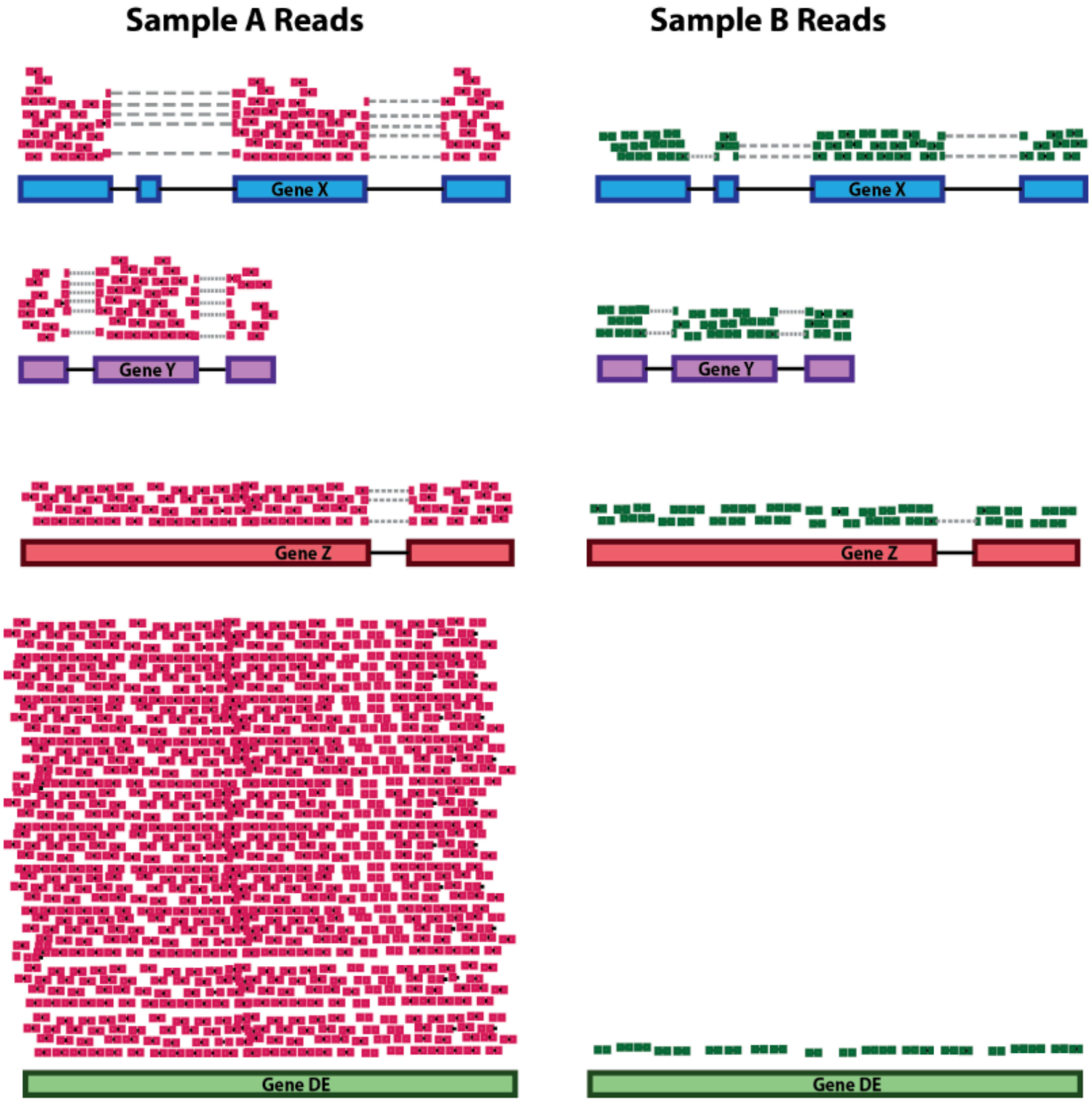
Well, many:

- **Duplicates**: PCR artifact, remove using *Picard*
- **Multiple mappable reads**: remove or *RSEM*
- **Comparison across samples**: normalization method such as *TMM*
- **Batch effects**: unwanted technical variation introduced by sequence efficiency, library preparation, different experiment conditions, unknown sources. Methods to use include *sva, RUVseq*

RNA-seq measures relative abundance of transcripts

Gene	Sample 1 absolute abundance	Sample 1 relative abundance	Sample 2 absolute abundance	Sample 2 relative abundance
1	20	10%	20	5%
2	20	10%	20	5%
3	20	10%	20	5%
4	20	10%	20	5%
5	20	10%	20	5%
6	100	50%	300	75%

RNA-seq measures relative abundance of transcripts



Normalization method 1: TPM

- TPM - Transcripts Per Million

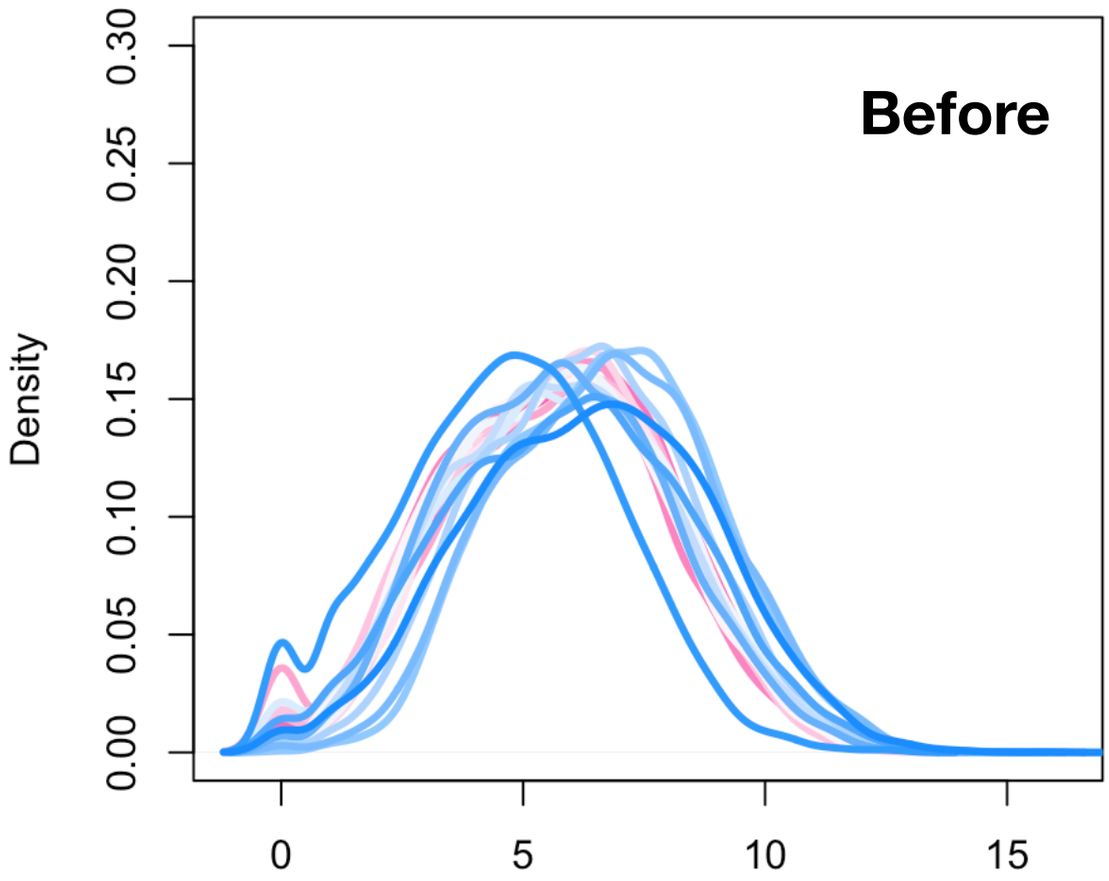
(estimate of) $\mathrm{TPM\ for\ isoform\ i} = 10^6 \times Z \times \frac{c_i}{\ell_i' N}$

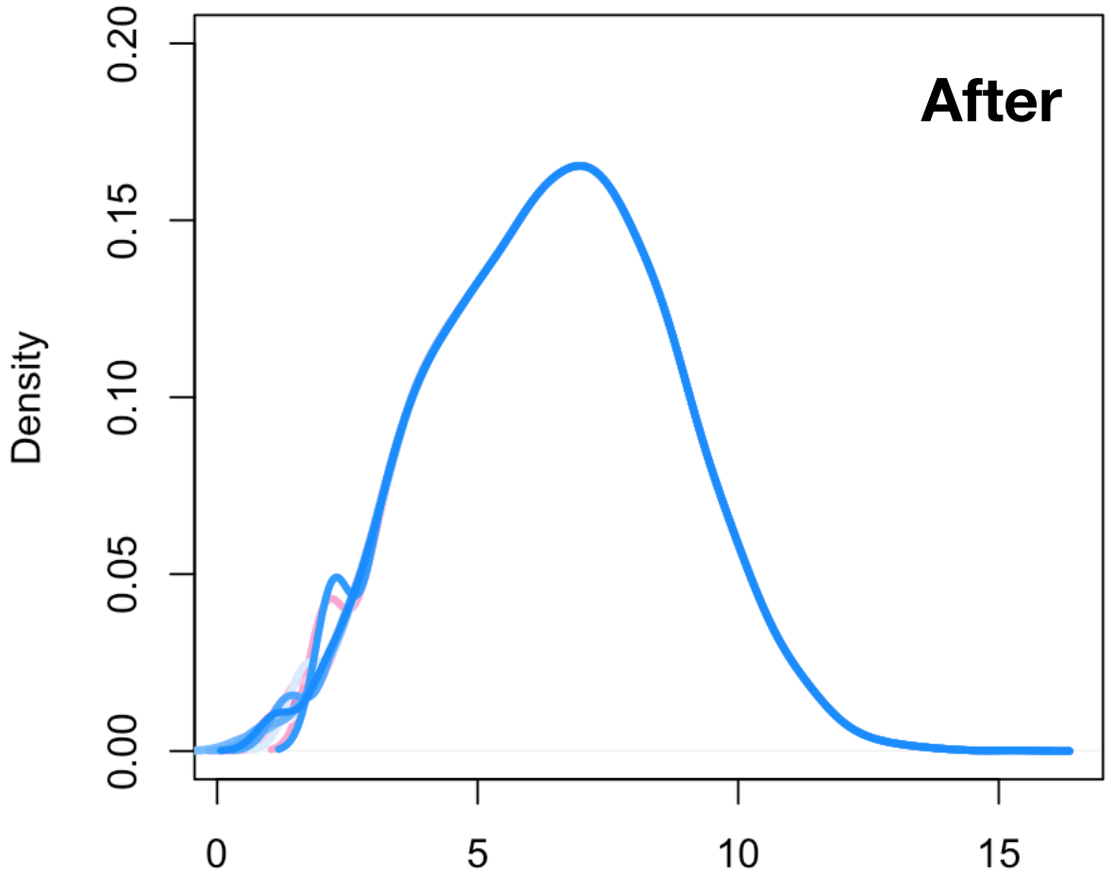
Factors considered: Gene length, Library size

Normalization method 2: Quantile Normalization

- **Assumption: The distribution of gene expression measures does not change across the samples.**

- **This assumption is unlikely to be true when testing treatments with severe effects on the transcription apparatus or studying cancer samples with severe genomic aberrations**





Factors considered: distribution, outliers

Normalization method 3: Median of Ratios method

Do: counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene

Factors considered: library sizes, RNA composition

gene	sampleA	sampleB	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = **1161.5**
ABCD1	22	13	sqrt(22 * 13) = **17.7**
...	...	...	...

Step 1: creates a pseudo-reference sample (row-wise geometric mean).

Normalization method 3: Median of Ratios method

Factors considered: library sizes, RNA composition

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**

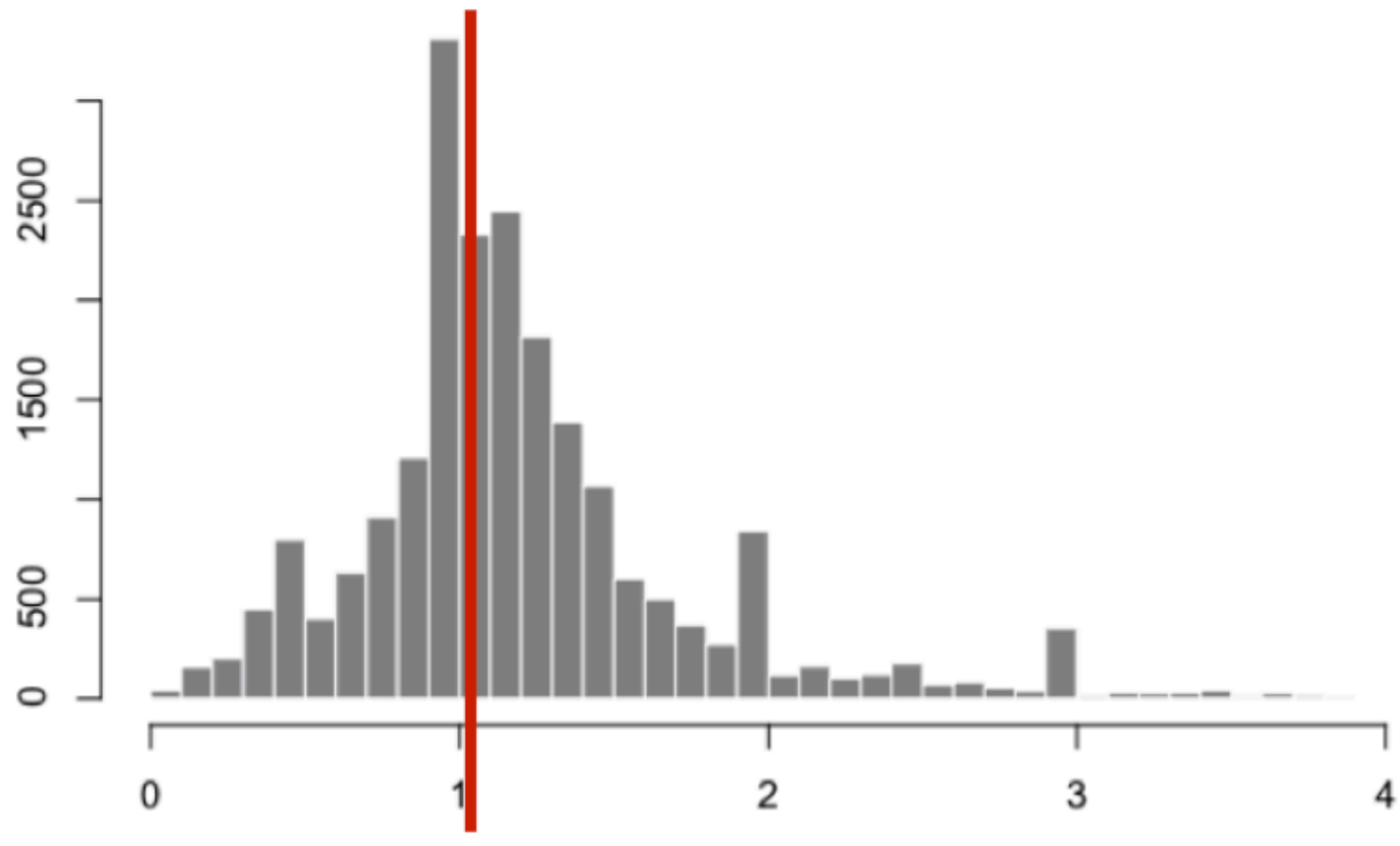
Step 2: calculates ratio of each sample to the reference.

Normalization method 3: Median of Ratios method

Factors considered: library sizes, RNA composition

```
normalization_factor_sampleA <- median(c(1.28, 1.3, 1.39, 1.35, 0.59))
```

```
normalization_factor_sampleB <- median(c(0.78, 0.77, 0.72, 0.74, 1.35))
```



Step 3: calculates the normalization factor for each sample (size factor).

Normalization method 3: Median of Ratios method

Factors considered: library sizes, RNA composition

SampleA median ratio = 1.3

SampleB median ratio = 0.77

Raw Counts

gene	sampleA	sampleB
EF2A	1489	906
ABCD1	22	13
...	...	...

Normalized Counts

gene	sampleA	sampleB
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**
...	...	...

Step 4: calculate the normalized count values using the normalization factor.

Normalization Method Summary

Method	Description	Accounted factors	Recommedations for use
TPM	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM	reads/fragments per kilobase of exon per million reads/ fragments mapped	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
median of ratios	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for DE analysis; **NOT for within sample comparisons**
trimmed mean of M values (TMM)	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for DE analysis

Normalization Method Summary

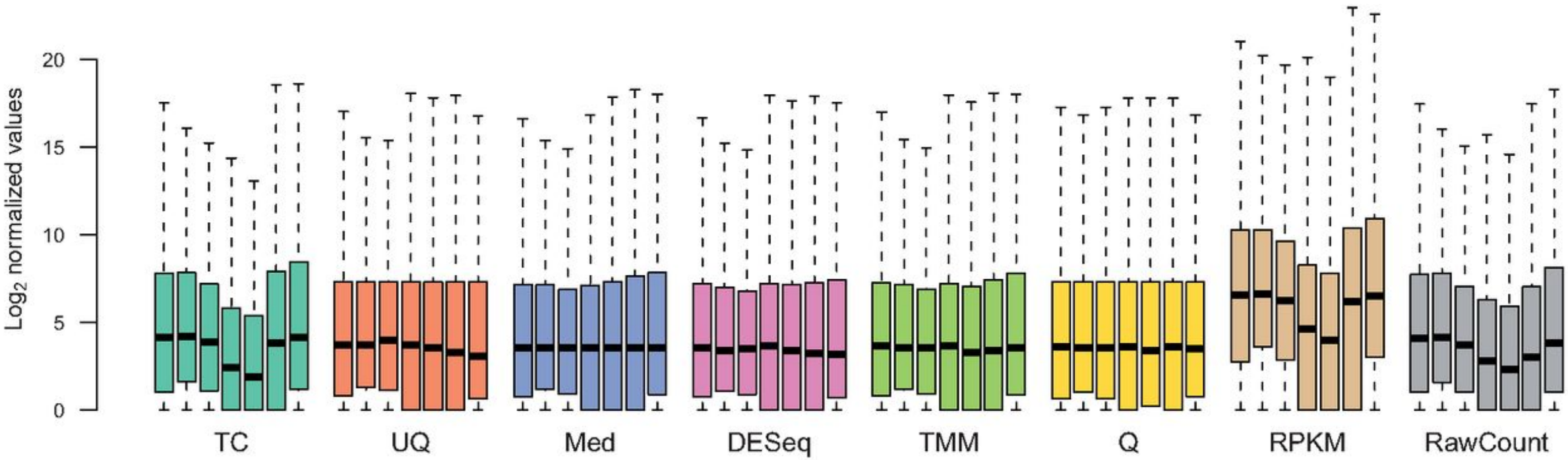
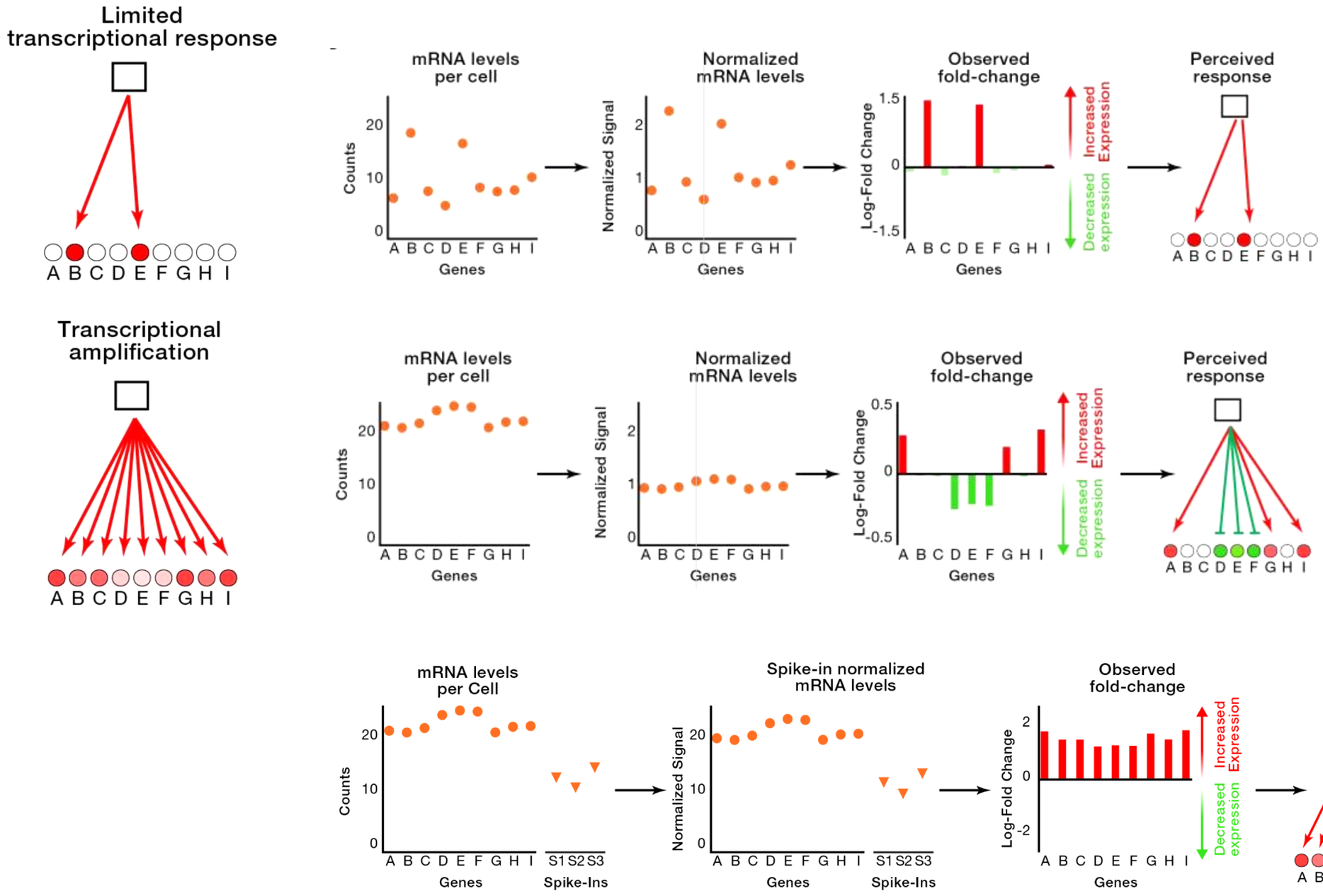


Figure from Dillies et al. (2013) that shows the effects of different approaches to normalize for read count differences due to library sizes (TC, total count; UQ, upper quartile; Med, median; DESeq, size factor; TMM, Trimmed Mean of M-values; Q, quantile) or gene lengths (RPKM).

Can we use these normalization methods to compare samples that are very different?

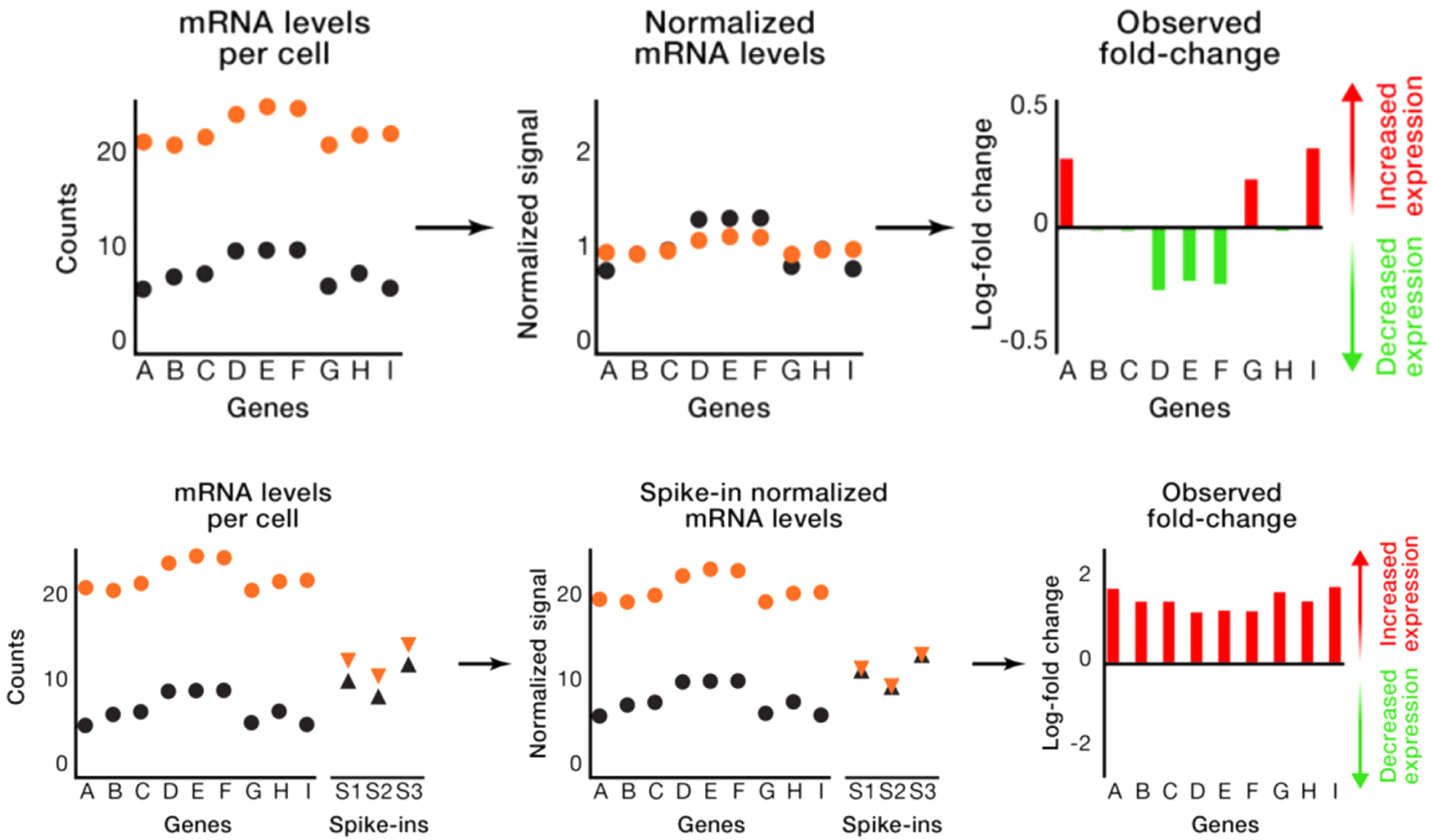
Widespread gene expression effects can be misinterpreted in the absence of adequate experimental and analytical steps. The normalization methods we talked about assume that expression of only a subset of genes is affected. Therefore, the presence of a global expression change in our samples (e.g. transcriptional amplification when an oncogene is activated) can lead to incorrect interpretations. To prevent this, we can use spike-ins, addition of synthetic RNA molecules of known abundance into our samples.

Normalization assumptions and issues



Loven *et al.*, *Cell* 2012

Normalization assumptions and issues



Loven *et al.*, *Cell* 2012

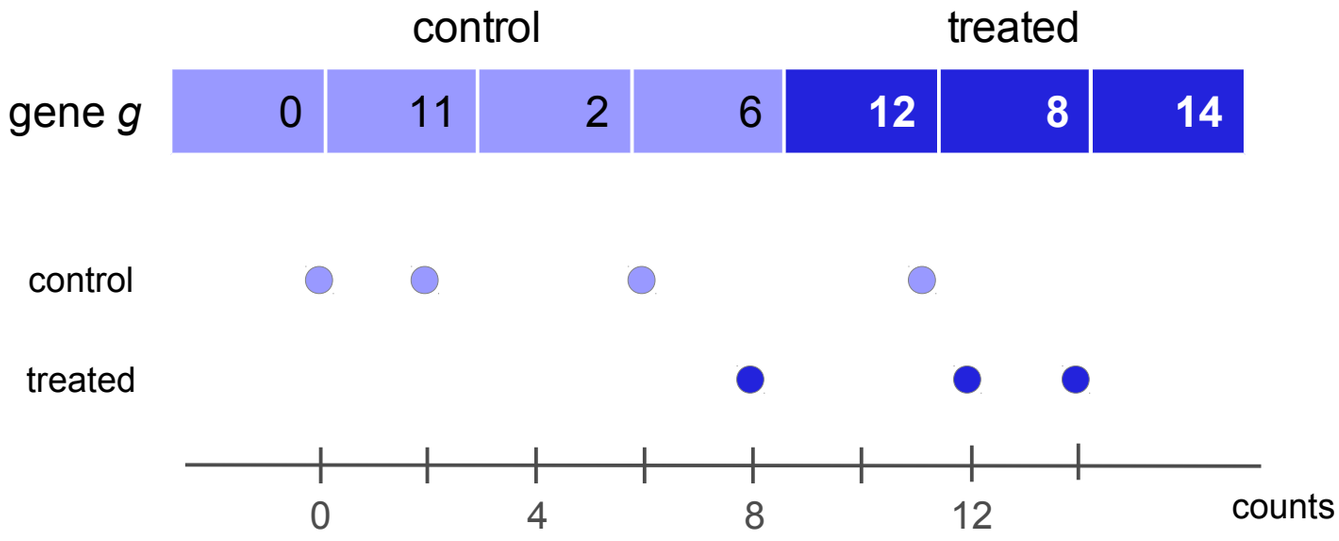
What to do next if we want to test for differential expression?

First of all, we need to use a statistical model to describe the data generating process of sequencing counts.

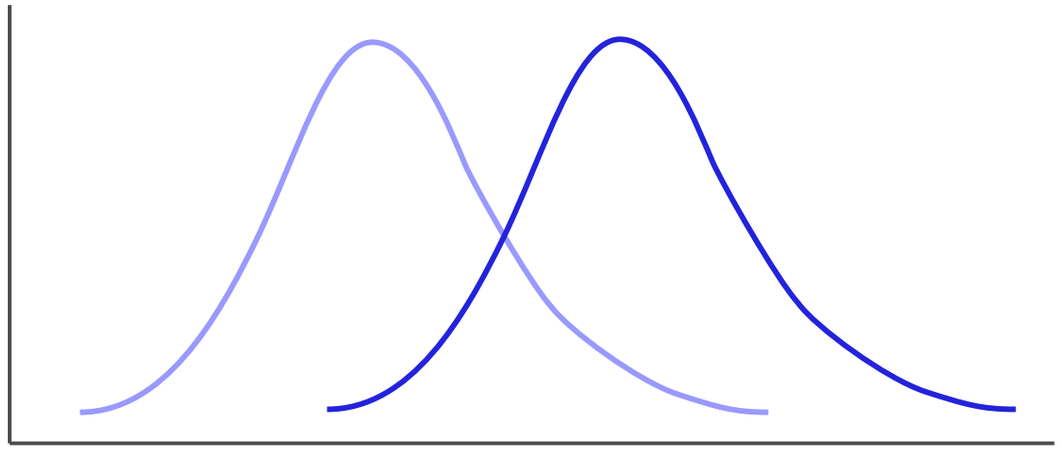
In probability theory and statistics, a probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. We treat the number of read sampled from the sequencing experiments as a random variable. We use distributions to describe the outcomes.

Testing for differential expression

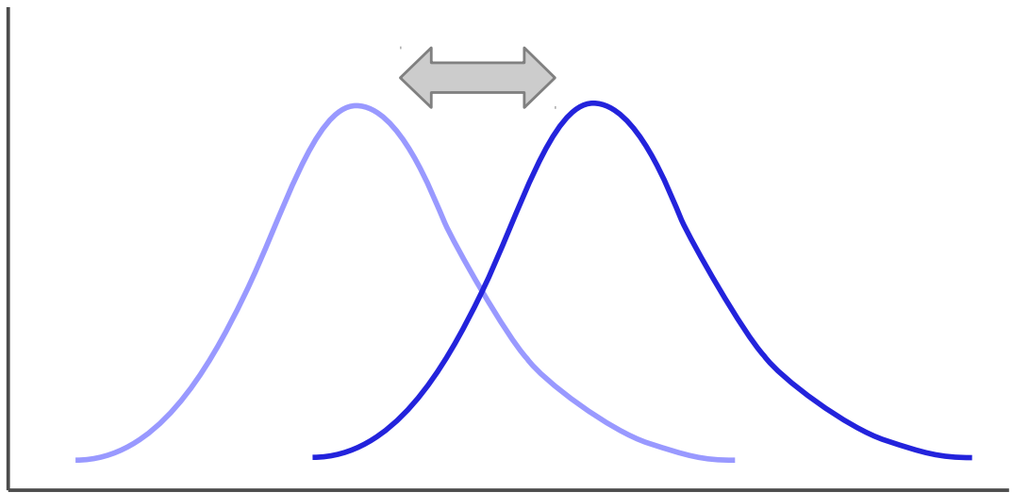
Data



Fit a distribution



Quantify the difference



Conclusion: differential or not?

Testing for differential expression

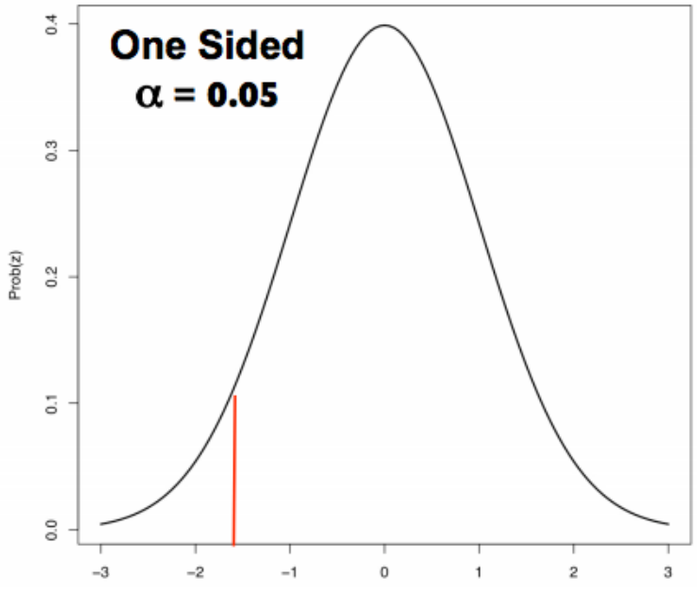
Null Hypothesis (H_0): The gene g is not differentially expressed between the conditions

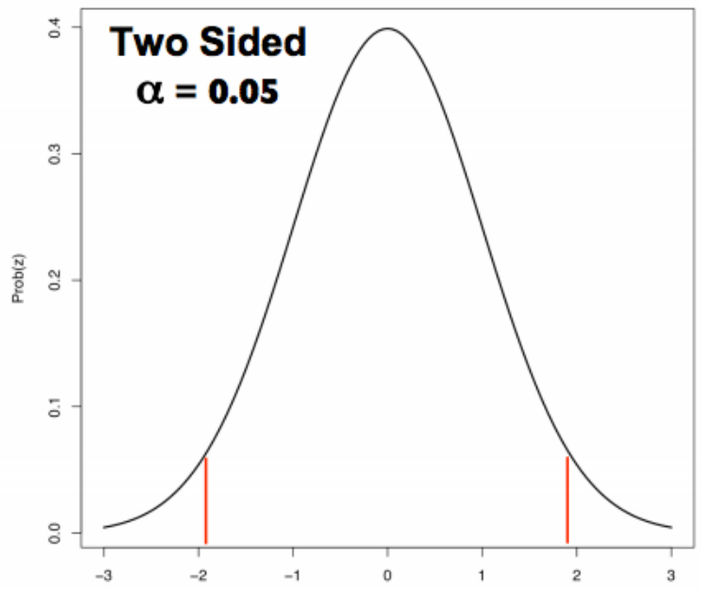
Alternative Hypothesis (H_A): The gene g is differentially expressed between the conditions

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

P-values: the probability of seeing a result as extreme or more extreme than the observed data, when the null hypothesis is true.

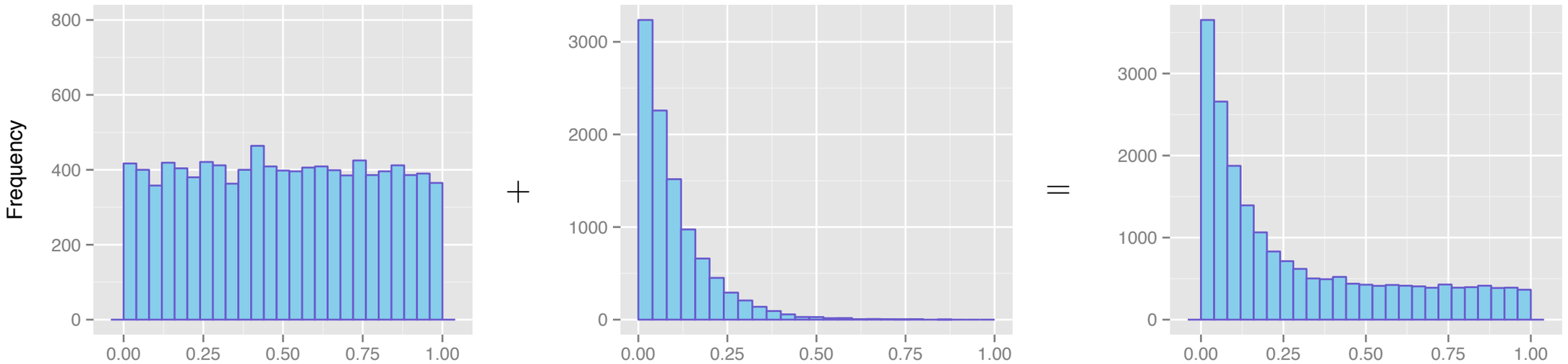




If p-value greater than 0.05, we fail to reject the null hypothesis.

There is nothing magical about p-value < 0.05, it is just a convention.

Diagnostic plots for multiple testing

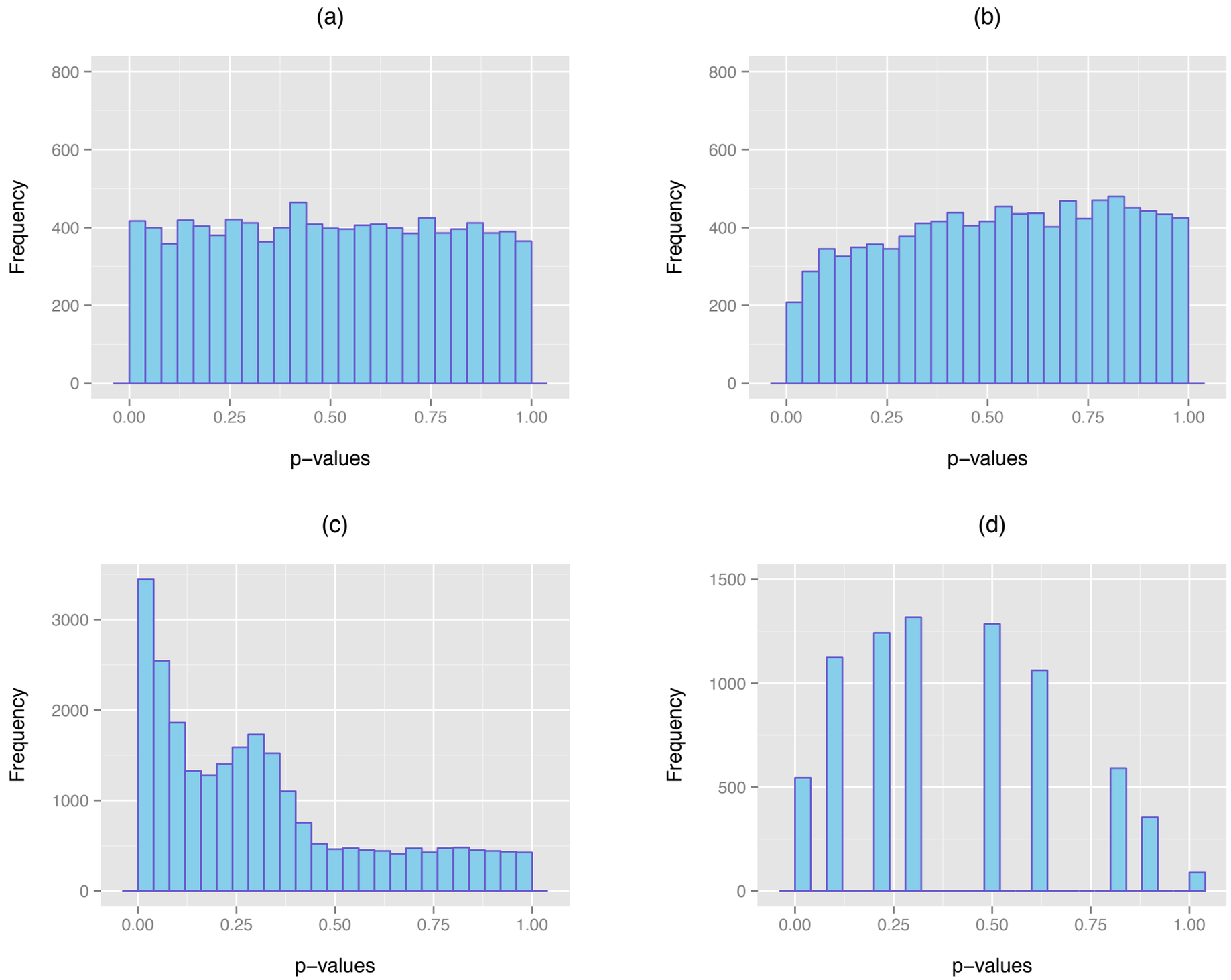


Distribution under H_0 | Distribution under H_1 | Resulting mixture distribution

Observed p-values can be considered as mix of samples from:

- **a uniform distribution (true nulls) and**
- **distributions concentrated at 0 (true alternatives)**

Unreliable differential expression tests



Understand DE analysis output

DESeq2 analysis example:

```
## Simulated data
set.seed(1)
dds = makeExampleDESeqDataSet(n = 1000, m = 6, betaSD = 1)

## DESeq2 analysis
dds = estimateSizeFactors(dds)
dds = estimateDispersions(dds)
dds = nbinomWaldTest(dds)

## Extract results from DESeq2 analysis
resDESeq2 = results(dds)
resDESeq2
```

```
log2 fold change (MAP): condition B vs A
Wald test p-value: condition B vs A
```

```
DataFrame with 1000 rows and 6 columns
          baseMean log2FoldChange     lfcSE      stat    pvalue      padj
         <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
gene1       12.247         1.5333    0.8054    1.9039   0.05693    0.2096
gene2       22.568         1.3896    0.6556    2.1197   0.03403    0.1554
gene3        3.961        -1.8592    0.9755   -1.9059   0.05666    0.2096
gene4      143.035        -0.5476    0.3421   -1.6010   0.10939    0.3082
gene5       16.301        -0.2533    0.6843   -0.3702   0.71125    0.8570
...            ...            ...       ...       ...       ...       ...
gene996     9.5873         1.33286    0.8254   1.61480    0.1064    0.3014
gene997     6.6044         0.08247    0.8567   0.09627    0.9233    0.9695
gene998     8.5560        -0.78911    0.7933  -0.99472    0.3199    0.5937
gene999     0.9542         0.66981    0.9694   0.69098    0.4896        NA
gene1000    3.7779         0.68939    1.0170   0.67789    0.4978    0.7372
```

References:

https://www.rna-seqblog.com/comparison-of-tmm-edger-rle-deseq2-and-mrn-normalization-methods/

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html