أكاديمية كاوست
KAUST ACADEMY

Gene expression analysis & Enrichment

Day 03

The KAUST Academy & The Bioinformatics Platform

4-7 Feb 2026

Evaluation

أكاديمية كاوست
KAUST ACADEMY

1. Theoretical Exam

- **No coding required**
  - Questions will evaluate your conceptual understanding in bioinformatics, the process of genomics data analysis, etc.
- **Weight:** 75% of the total grade
- **Exam Duration:** 2 hours
- **Question style:**
  - Conceptual, Comprehension & Familiarity with Bioinformatics Analysis
  - Designed to test **your understanding**, not memorization
- **Number of questions > 50**

2a. Project Presentation & Report

- **One report per group**
- **Number of presenters:** flexible (up to the group)
- **Presentation format:** no strict format required we look into:
  - Quality of the analysis
  - Clear explanation of main steps and relevance of the results
  - Problem solving skills
- **Time allocation: 5 minutes** presentation per group

2b. Delivery

Each group is responsible for uploading:
•One report
•One presentation
📁 Upload both files (**Report & PPT**) to the **Report & Presentation** folder
⏰ **Deadline: 9:00 AM, February 7**

Agenda – Day 03

أكاديمية كاوست
KAUST ACADEMY

Morning Session

9-9:15	S1: Recap of Day 2
9:15-10	S2: Differential Expression Concepts
10-11	*L1:* Perform differential expression analysis, inspect DE results
11-11:30	S3: Visualization & Reporting
11:30-12	L2: Visualization & Integrated QC

Afternoon Session

2-2:45	S4: Pathway & Enrichment Analysis
2:45-3:30	*L3: Hands-on Enrichment Analysis*
3:30-4	S5: Extended Applications of RNA-seq
4-5	S6: Project Work

Recap

- Quality Control
- Reference Genome Alignment
- Pseudo Alignment: Gene Quantification

Agenda – day 3

01

Differential Expression

Normalization, statistical modeling,
hypothesis testing, DESeq2 workflow

02

Visualization

PCA, volcano plots, MA plots,
heatmaps, quality assessment

03

Enrichment Analysis

Gene Ontology, pathway analysis

04

Extended Applications of RNA-seq

Variant detection, single cell, spatial transcriptomics

Agenda – Day 03

أكاديمية كاوست
KAUST ACADEMY

Morning Session

9-9:15	S1: Recap of Day 2
9:15-10	S2: Differential Expression Concepts
10-11	*L1:* Perform differential expression analysis, inspect DE results
11-11:30	S3: Visualization & Reporting
11:30-12	L2: Visualization & Integrated QC

Afternoon Session

2-2:45	S4: Pathway & Enrichment Analysis
2:45-3:30	*L3: Hands-on with Transcriptomics Data*
3:30-4	S5: Extended Applications of RNA-seq
4-5	S6: Team Formation & Project Overview

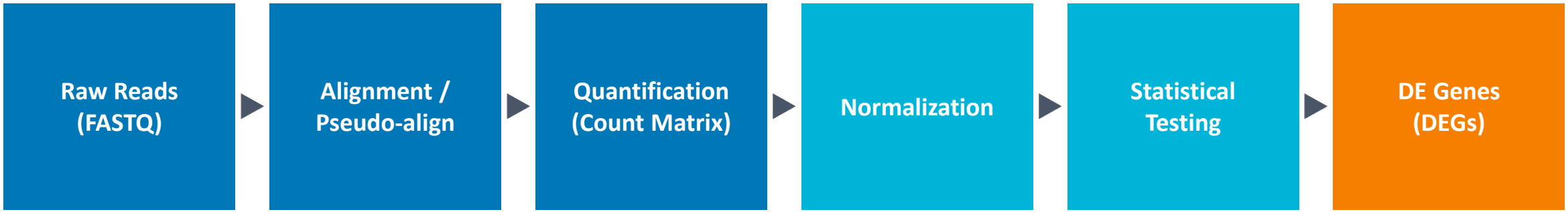
أكاديمية كاوست
KAUST ACADEMY

Differential Expression Concepts

Day 03 – Session 02

Differential Expression

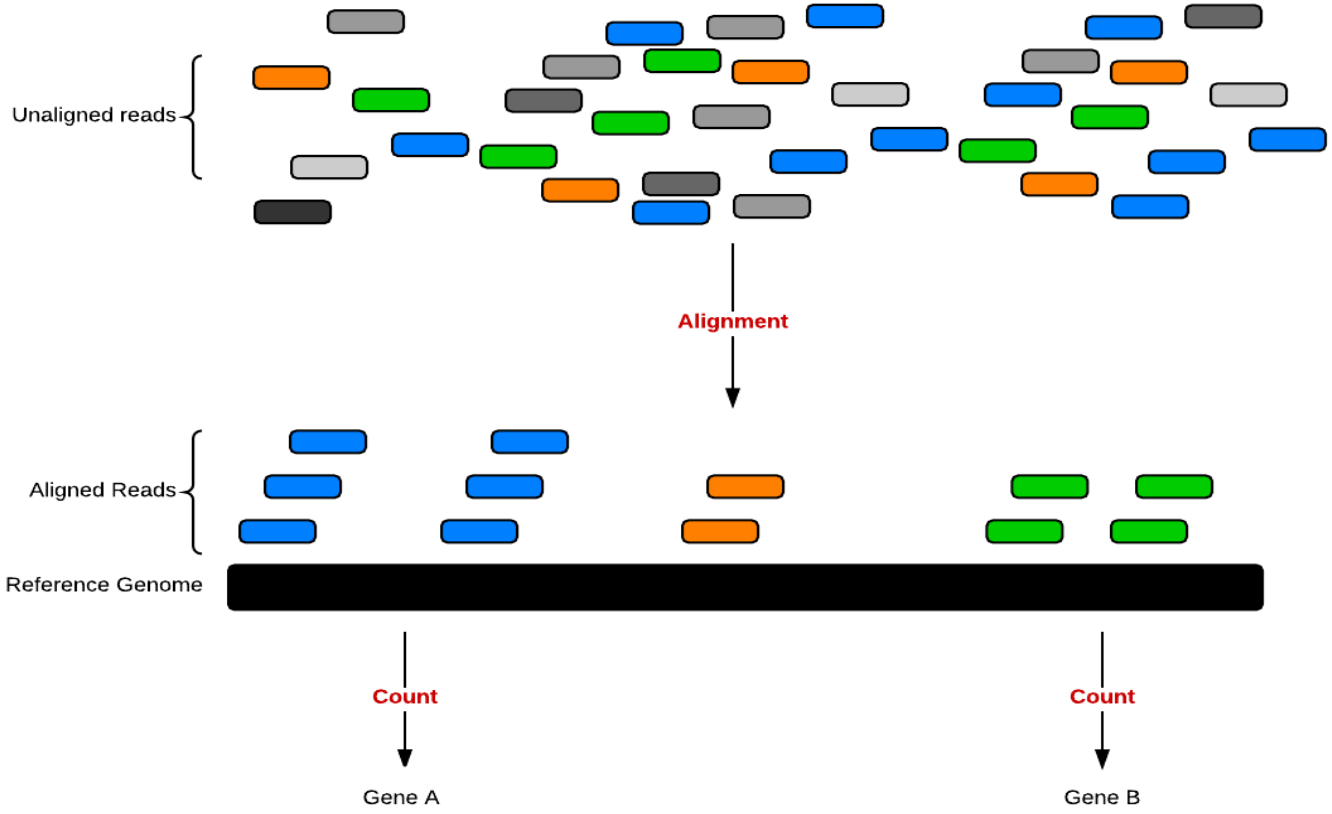
The goal of DE testing is to determine which genes are expressed at statistically different levels between experimental conditions.



This session covers steps from the count matrix through DE identification

*Key tools: **DESeq2**, edgeR, limma-voom | References: Love et al. 2014, Robinson et al. 2010, Law et al. 2014*

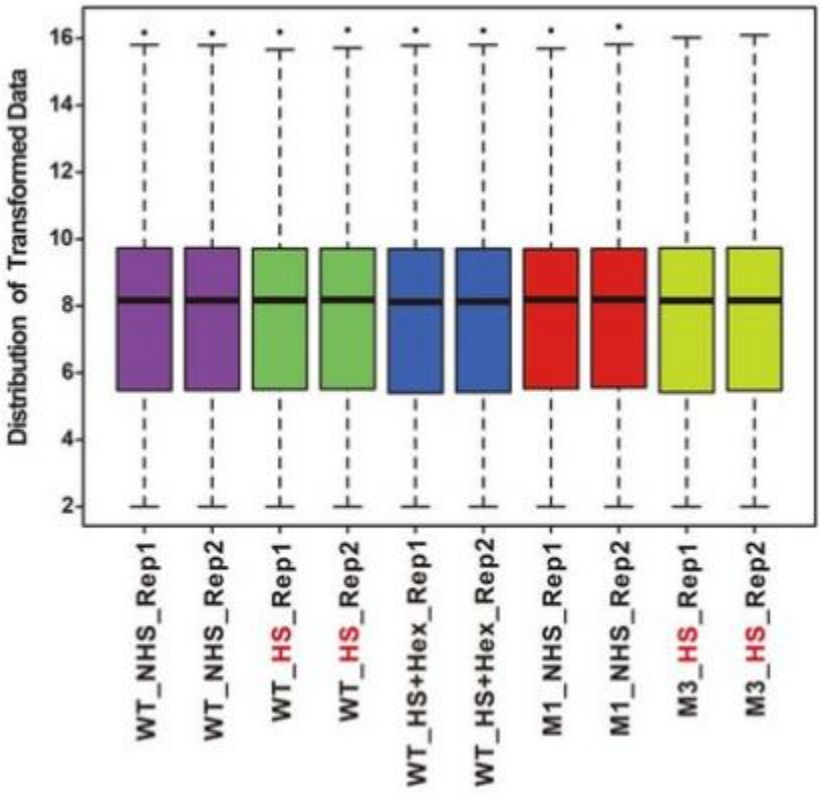
Differential Expression

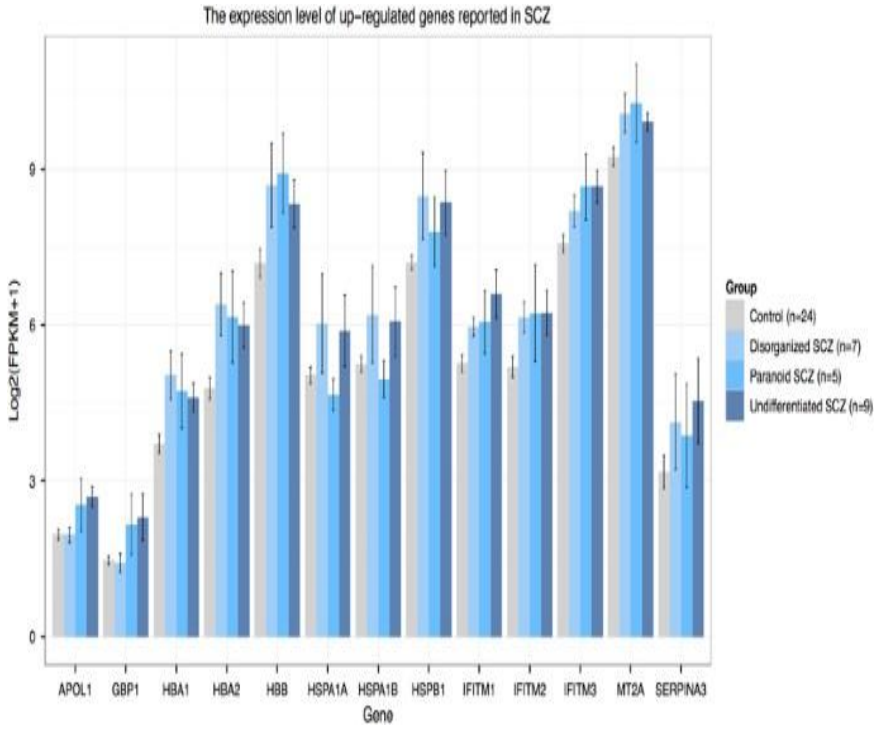


Differential expression (DE) identifies genes whose expression changes **beyond random noise** between conditions.

The goal of differential expression testing is to determine which genes are expressed at different levels between conditions.

Differential Expression





Key ideas:

- Counts vary naturally
- Replicates are essential
- Statistics separate signal from noise
- Even identical samples won't have identical counts.
- We must model **biological + technical variability**.

Count Matrix

RNA-seq quantification produces a matrix of integer read counts: rows = genes, columns = samples.

Gene	Ctrl_1	Ctrl_2	Ctrl_3	Treat_1	Treat_2	Treat_3
BRCA1	523	612	498	1247	1189	1356
TP53	3042	2987	3156	3201	3089	3245
MYC	876	923	845	412	378	445
GAPDH	15234	14876	15567	15012	14923	15345

Raw counts are influenced by technical factors and must be normalized before comparison.

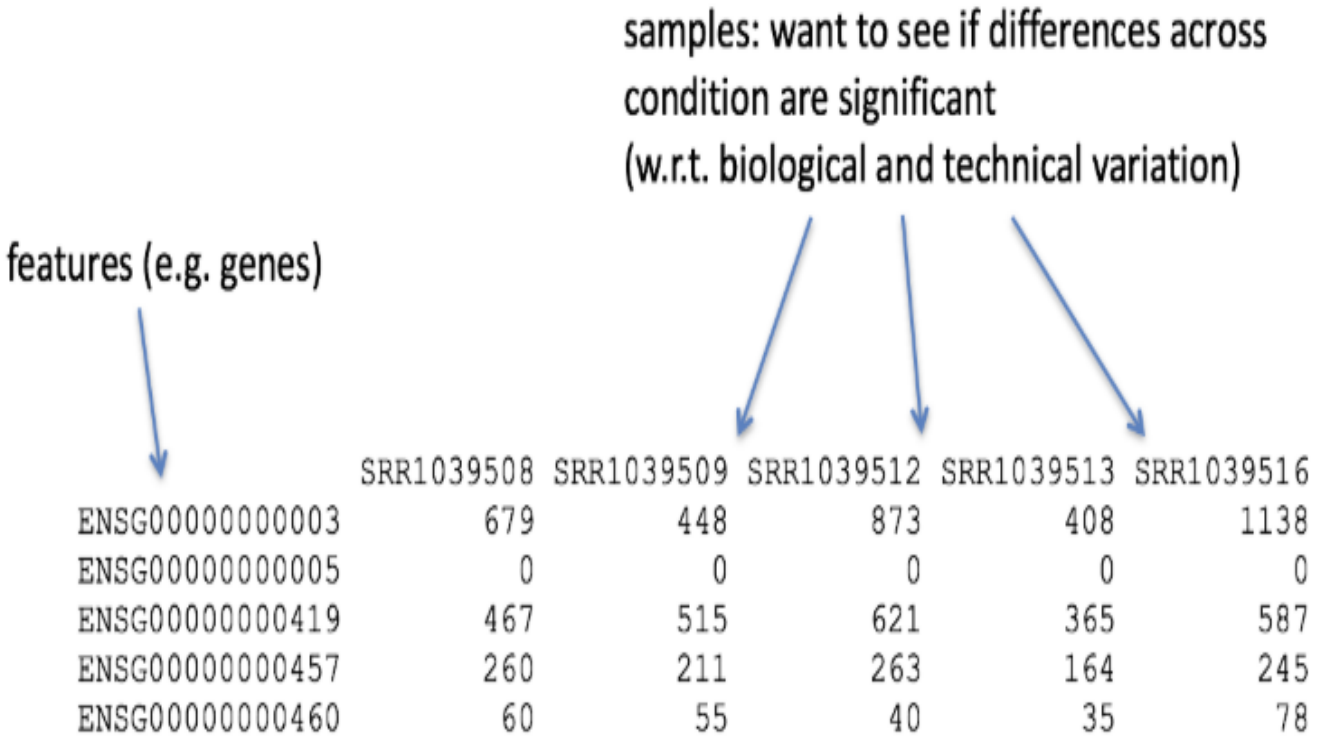
Higher counts ≠ higher expression without normalization!

Count Matrix (Chr11 count matrix)

Filter

	KO_1	KO_2	KO_3	WT_1	WT_2	WT_3
ENSG00000002330	894.001	945.225	828.000	709.835	784.982	676.000
ENSG00000005801	1112.096	1195.335	1372.418	2031.598	2640.474	1953.730
ENSG00000006071	6.976	8.123	6.126	5.052	5.156	6.898
ENSG00000006118	2359.351	2656.001	2388.003	3827.001	3935.999	3365.999
ENSG00000006534	2154.000	2106.336	2144.988	5160.683	5213.987	4869.684
ENSG00000006611	56.999	63.999	50.000	98.000	84.999	56.001
ENSG00000007372	692.885	645.840	724.059	942.793	914.968	848.283
ENSG00000011347	901.000	948.018	894.999	412.999	401.999	388.000
ENSG00000011405	5696.433	6079.254	5782.191	6658.913	7492.005	6401.650
ENSG00000013725	1.000	1.000	0.000	1.000	1.000	2.000
ENSG00000014138	1949.034	1974.479	1942.497	2357.336	2592.020	2295.673
ENSG00000014216	3296.001	3477.000	3288.059	3344.152	3406.679	3258.001
ENSG00000019102	0.000	1.000	0.000	0.000	0.000	0.000
ENSG00000019144	2172.117	2415.428	2487.230	2119.817	2210.612	1997.139
ENSG00000019485	1035.000	1096.037	966.000	1696.001	1734.000	1482.000
ENSG00000019505	4.220	10.315	11.003	4.148	2.043	2.789
ENSG00000020922	2921.648	3050.366	3095.713	4064.267	4674.470	3831.009
ENSG00000021300	1.000	0.000	0.000	0.000	2.000	0.000
ENSG00000021762	1888.003	2091.558	1903.999	2534.285	2633.878	2115.262
ENSG00000023171	3899.036	4222.015	3920.074	3614.194	3667.034	3122.008

Count Matrix cnt.



	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

What does the count data actually represent?

The count data used for differential expression analysis represents the number of sequence reads that originated from a particular gene.

The higher the number of counts, the more reads associated with that gene, and the assumption that there was a higher level of expression of that gene in the sample.

Try to interpret me :)

	KO_1	KO_2	KO_3	WT_1	WT_2	WT_3
ENSG00000177830	149.999	141.000	144.000	1644.998	1761.001	1692.999
ENSG00000284057	43.565	33.351	0.000	245.576	263.860	277.515
ENSG00000151364	14.000	21.000	19.340	111.990	110.693	114.346
ENSG00000175592	190.000	267.000	157.000	537.001	544.999	422.000
ENSG00000165905	108.000	106.000	114.001	299.999	261.001	241.999
ENSG00000172927	44.000	45.000	31.001	74.000	111.001	106.000
ENSG00000006534	2154.000	2106.336	2144.988	5160.683	5213.987	4869.684
ENSG00000243964	329.668	477.683	1603.927	1934.557	1905.867	1830.601
ENSG00000259112	22.501	40.467	48.717	74.813	99.510	91.088
ENSG00000110514	1439.999	1575.443	1489.000	3438.018	3840.000	3066.998
ENSG00000204529	41.000	45.007	29.000	74.146	108.027	85.723
ENSG00000148926	254.000	306.001	226.001	504.001	677.000	482.001
ENSG00000214756	32.000	26.000	34.000	75.129	59.000	54.000

Data structure – Experiment Design

أكاديمية كاوست
KAUST ACADEMY

countData

	ctrl_1	ctrl_2	exp_1	exp_1
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

colData

	treatment	sex
ctrl_1	control	male
ctrl_2	control	female
exp_1	treatment	male
exp_2	treatment	female

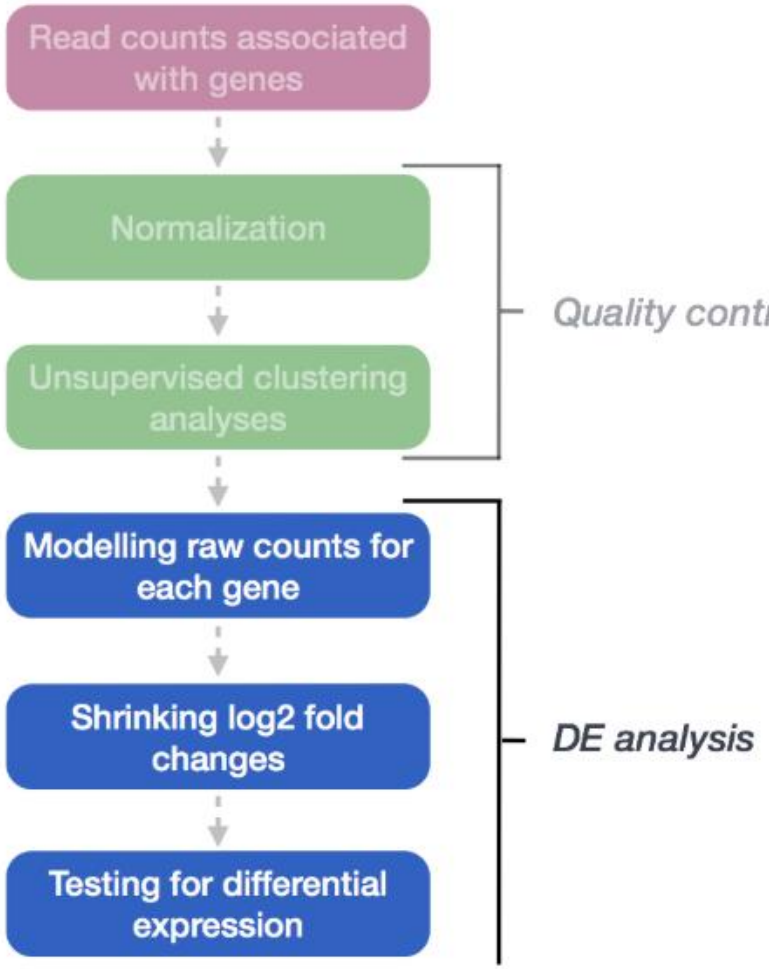
Sample names:
ctrl_1, ctrl_2, exp_1, exp_2

countData is the count matrix
(number of reads mapping to each gene for each sample)

colData describes metadata about the *columns* of countData

colnames(countData) == rownames(colData)

DEG pipeline



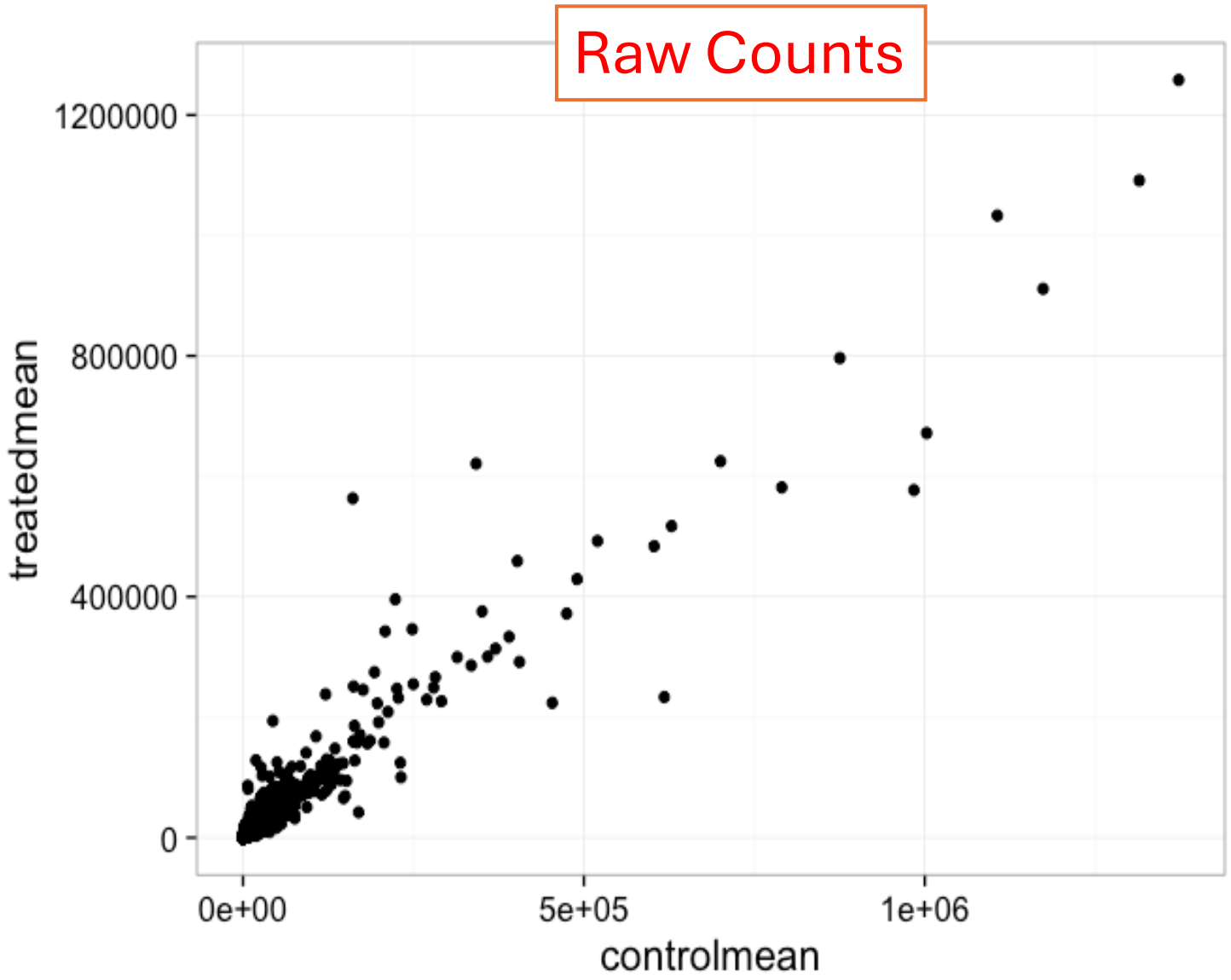
Why this steps is important?

Why do we need normalization?

Raw counts are influenced by technical artifacts that must be removed before comparing expression between samples.

Sequencing Depth	Different total reads per sample. A sample sequenced 2x deeper will have ~2x counts for every gene.
Gene Length	Longer genes capture more fragments. Important for within-sample comparisons (RPKM/TPM), NOT needed for DE of same gene across conditions.
RNA Composition	A few highly expressed genes consume disproportionate reads, making other genes appear under-expressed. Library-size normalization alone cannot fix this.

Why Are Raw Counts Not Comparable

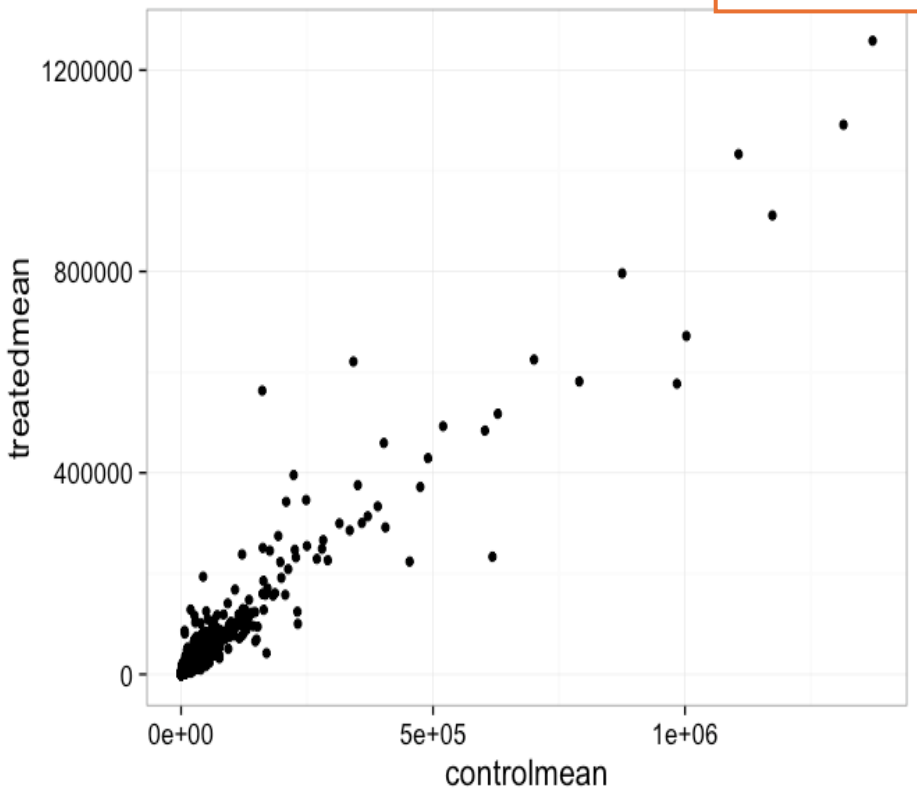


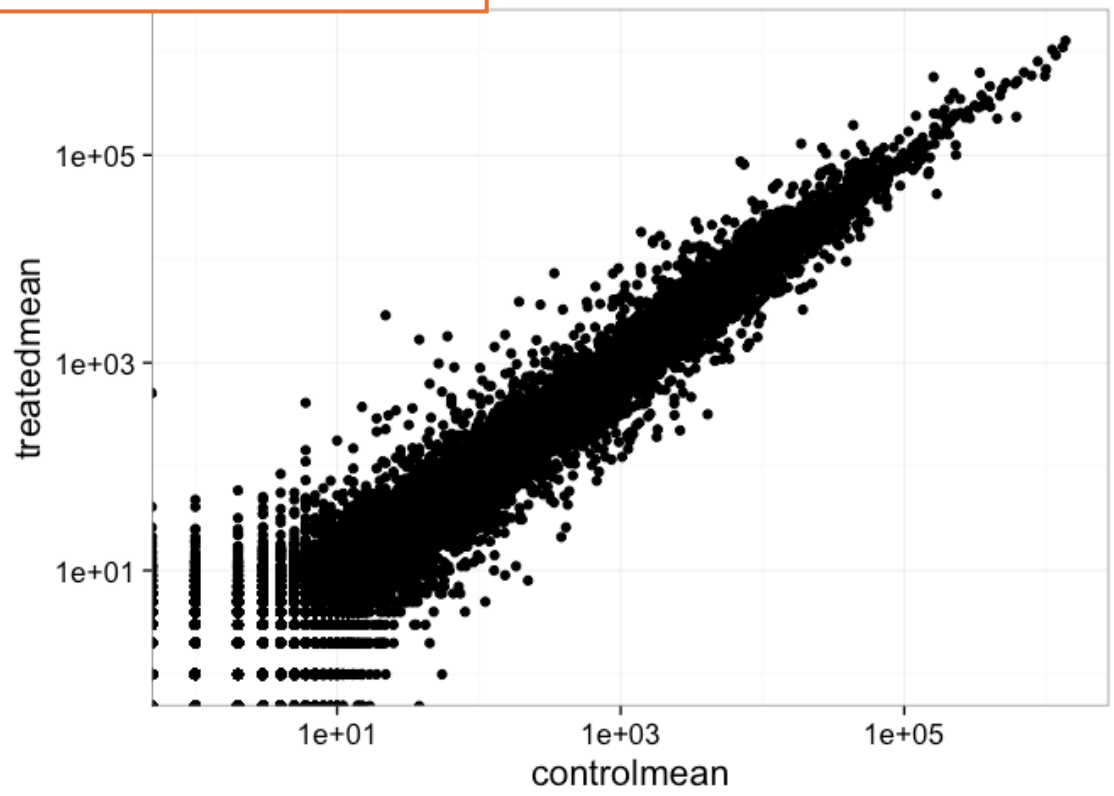
1.Create a scatter plot showing the mean of the treated samples against the mean of the control samples.

2.Wait a sec. There are 60,000-some rows in this data, but we only seeing a few dozen dots at most outside of the big clump around the origin.

Why Are Raw Counts Not Comparable

Transformed Counts



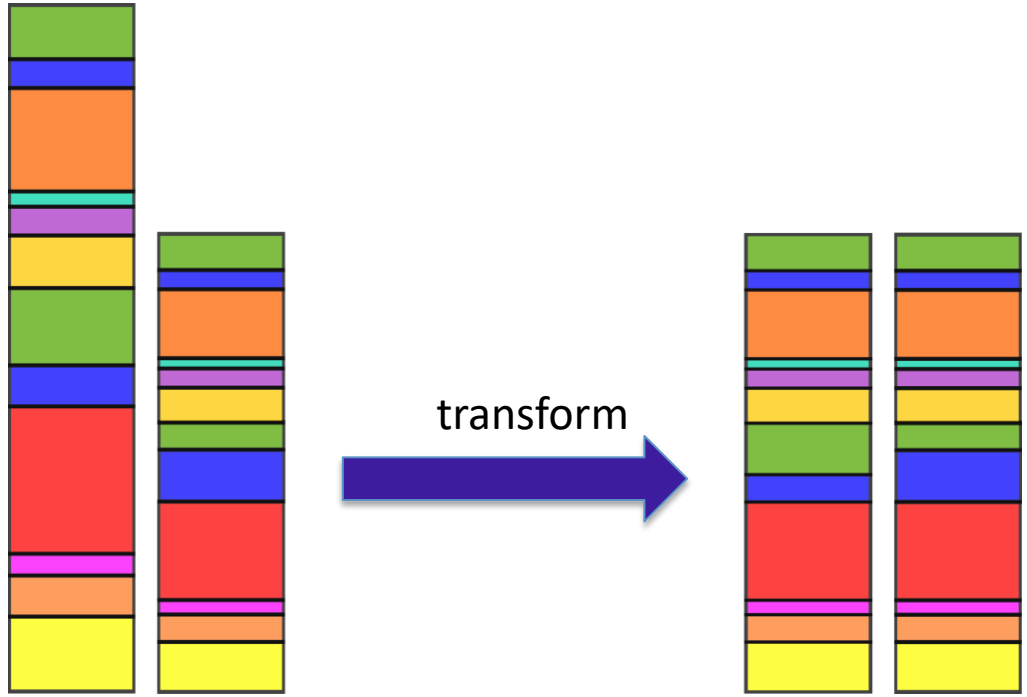


1.Try plotting both axes on a log scale (*hint*: `... + scale_..._log10()`)

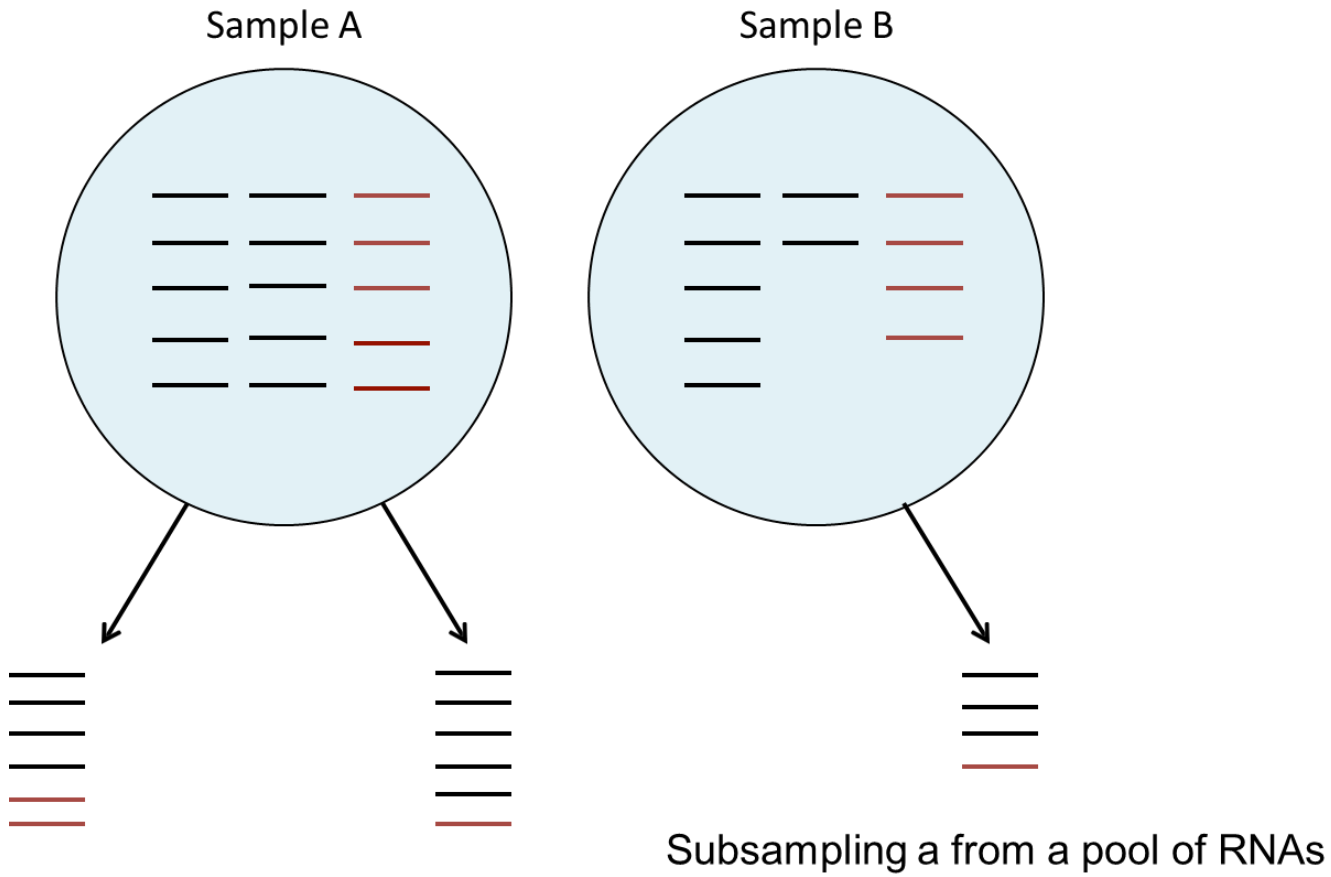
Transformation

Total Count

- Normalise each sample by total number of reads sequenced.
- Can also use another statistic similar to total count; eg. median, upper quartilex



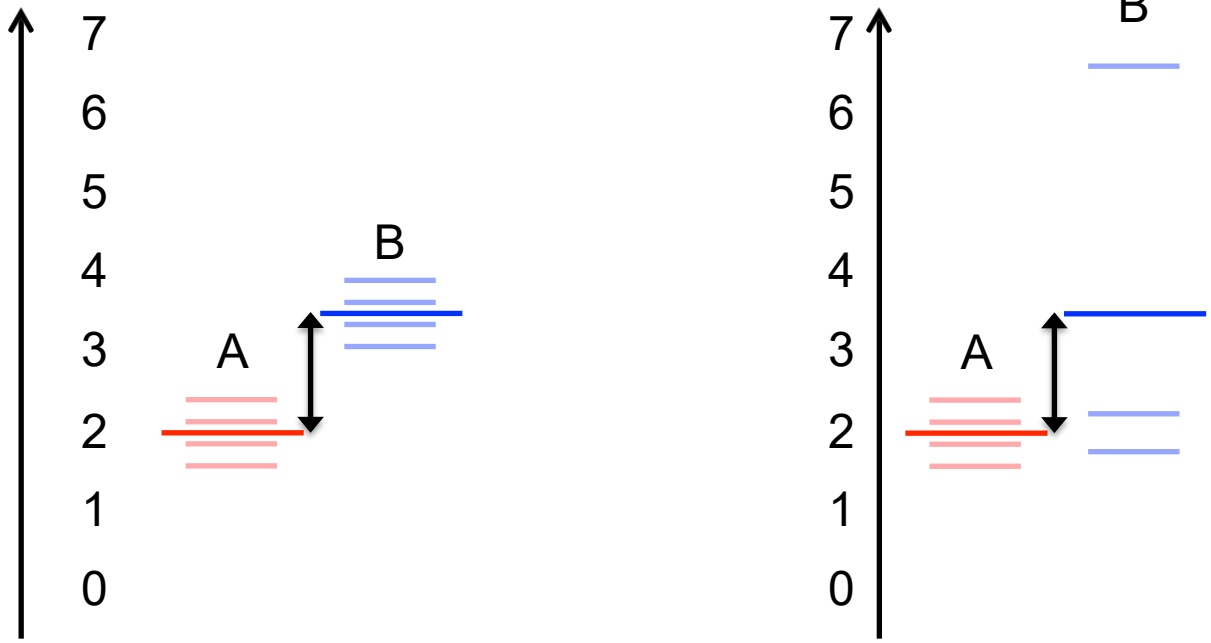
Source of variation: Sampling Bias



Necessary to make accurate comparisons of gene expression between samples.

Source of variation: Variations in replicates

- Simple difference in means



Differential expression (DE) identifies genes whose expression changes **beyond random noise** between conditions.

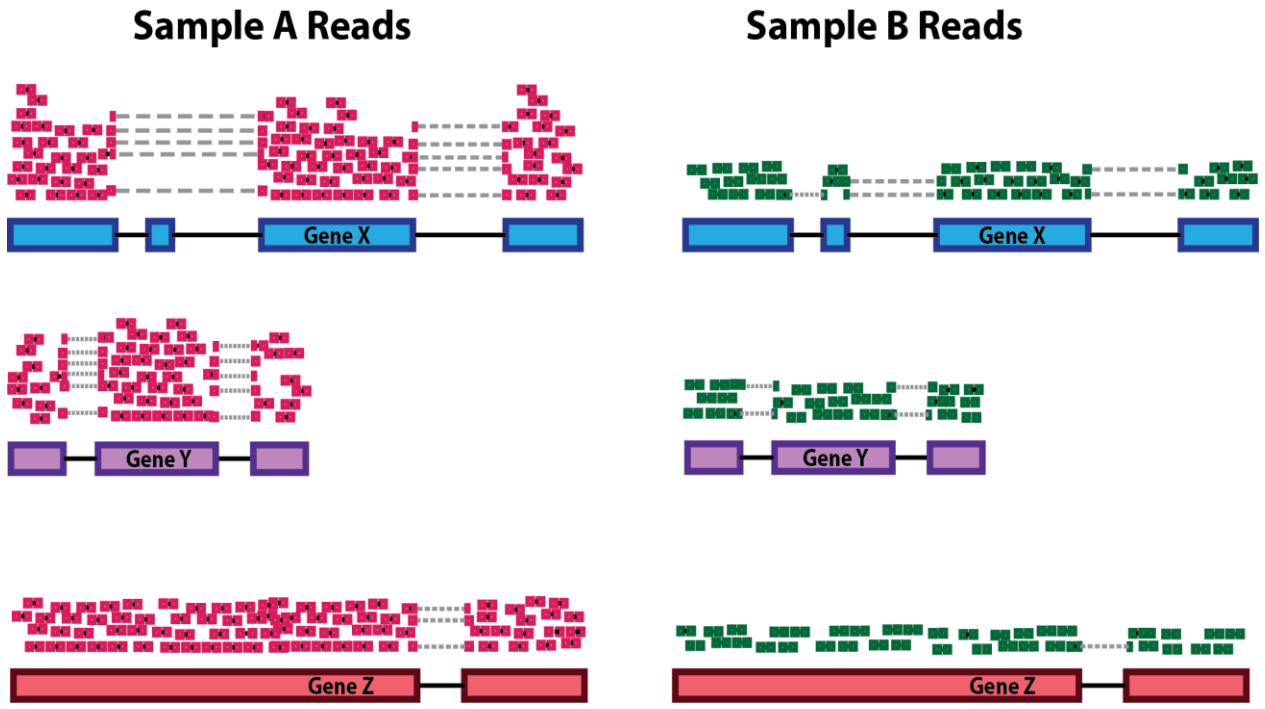
Key ideas:

- Counts vary naturally
- Replicates are essential
- Statistics separate signal from noise

- Replication introduces variance

Source of variation: Sequencing Depth

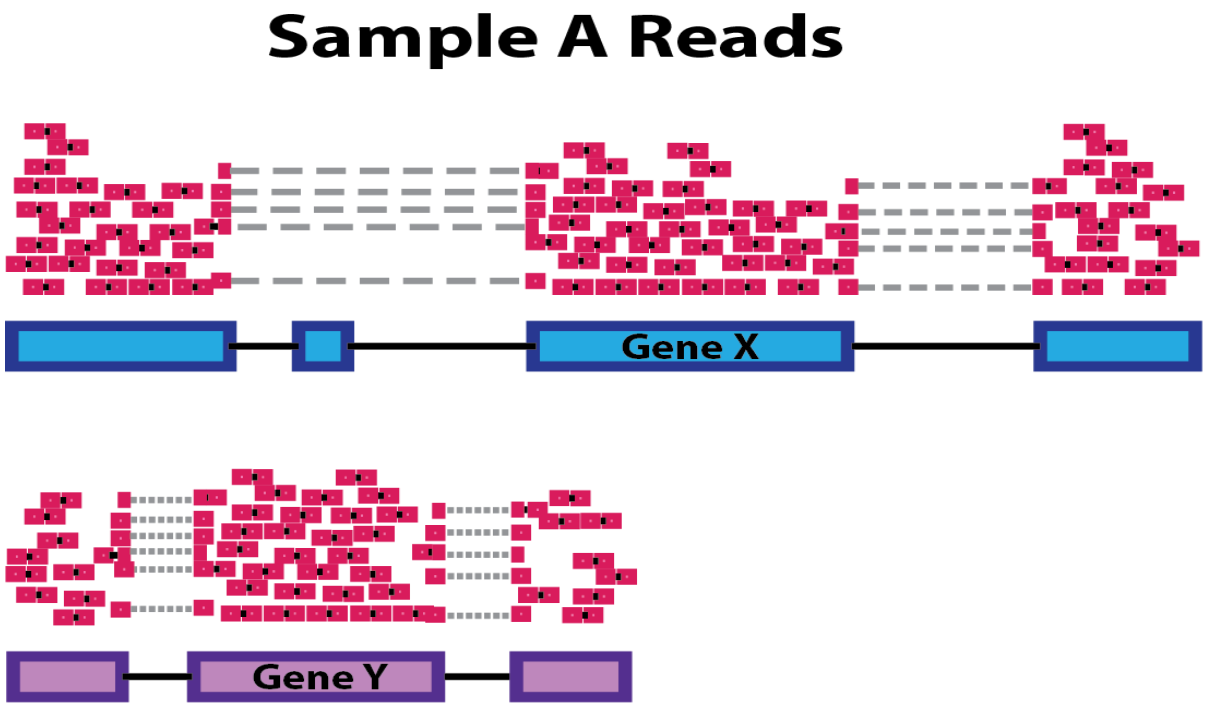
The main factors often considered during normalization are:



1. **Sequencing depth:** Accounting for sequencing depth is necessary for comparison of gene expression between samples. In the example below, each gene appears to have doubled in expression in *Sample A* relative to *Sample B*, however this is a consequence of *Sample A* having double the sequencing depth.

Source of variation: Gene Length

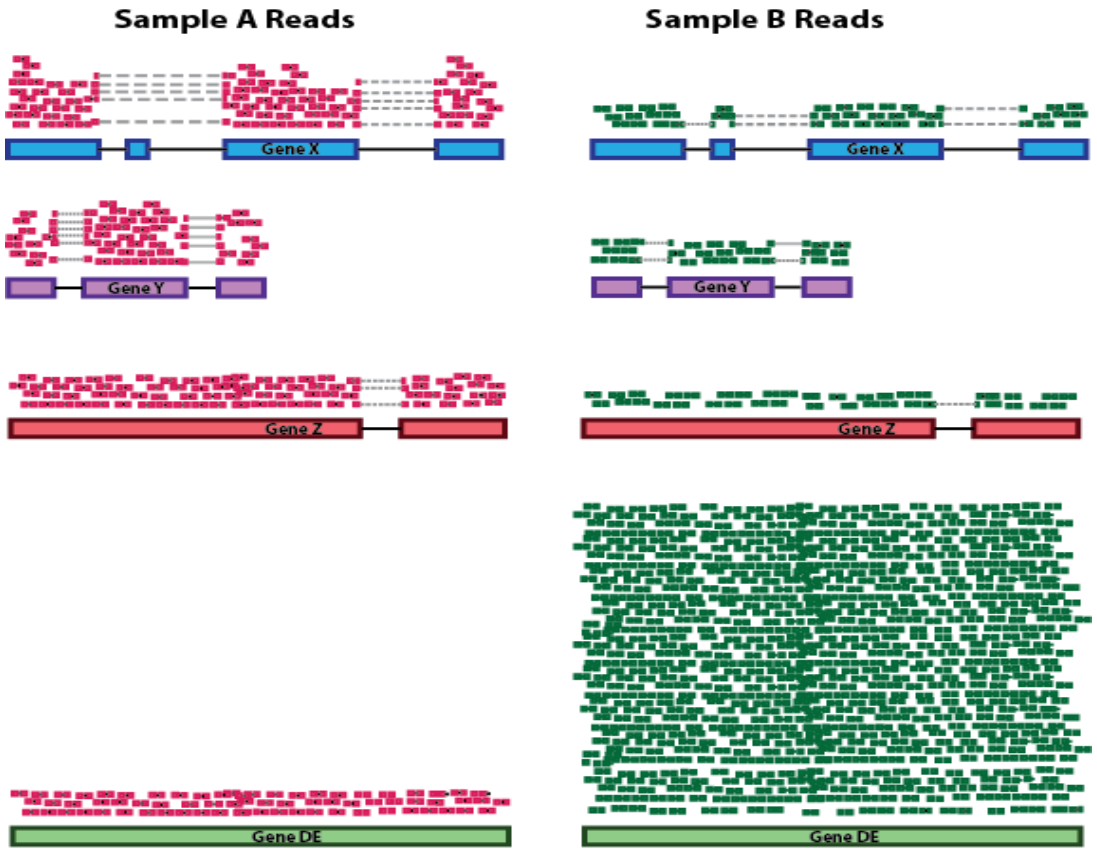
The main factors often considered during normalization are:



2. Gene length:
Accounting for gene length is necessary for comparing expression between different genes within the same sample. In the example, *Gene X* and *Gene Y* have similar levels of expression, but the number of reads mapped to *Gene X* would be many more than the number mapped to *Gene Y* because *Gene X* is longer.

Source of variation: RNA Composition

The main factors often considered during normalization are:



3. RNA composition: A few highly differentially expressed genes between samples, differences in the number of genes expressed between samples, or presence of contamination can skew some types of normalization methods. Accounting for RNA composition is recommended for accurate comparison of expression between samples, and is particularly important when performing differential expression analyses.

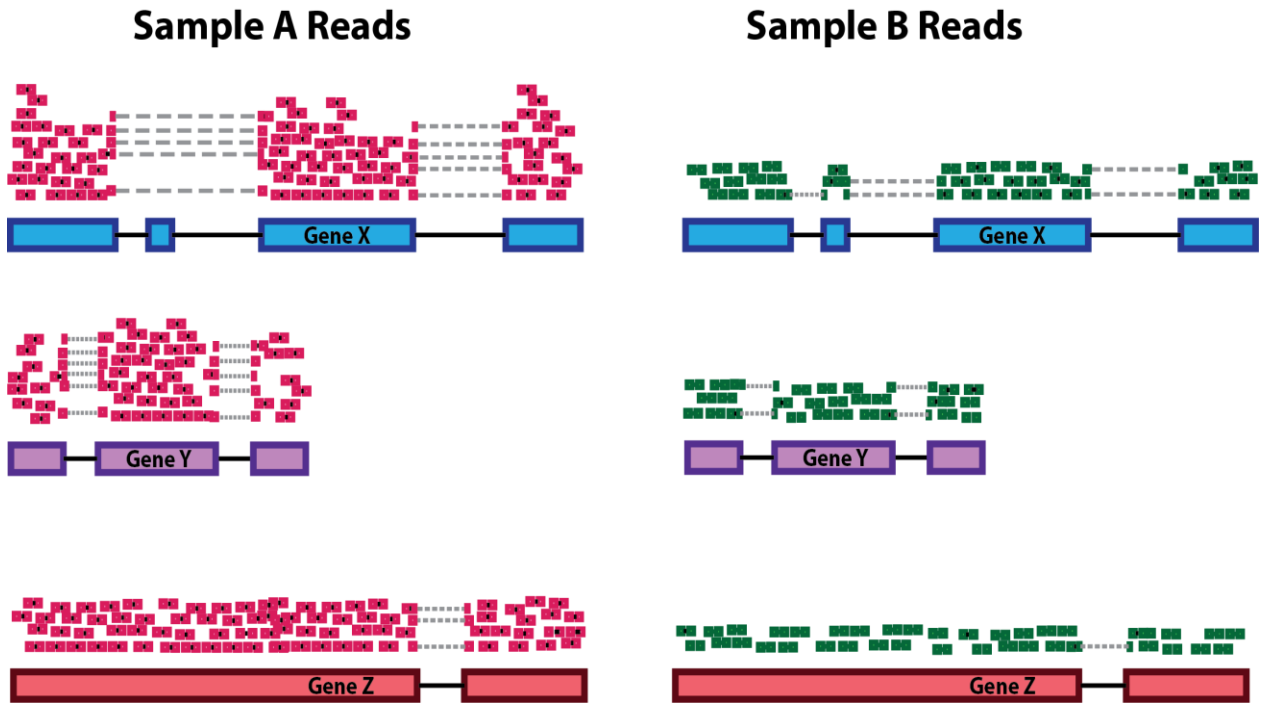
Normalization Methods Comparison

Method	Accounts For	Suitable For	Limitations
CPM	Sequencing depth	Quick exploratory analysis	Ignores RNA composition
RPKM / FPKM	Depth + gene length	Within-sample comparisons	Not comparable across samples
TPM	Depth + gene length	Cross-sample gene comparison	No RNA composition correction
DESeq2 (Median of Ratios)	Depth + RNA composition	Differential expression	Assumes most genes not DE
TMM (edgeR)	Depth + RNA composition	Differential expression	May over-trim in extremes

For DE analysis, use DESeq2 or TMM — NOT RPKM/FPKM/TPM
For Visualization, use TPM

Normalization: Counts per million (CPM)

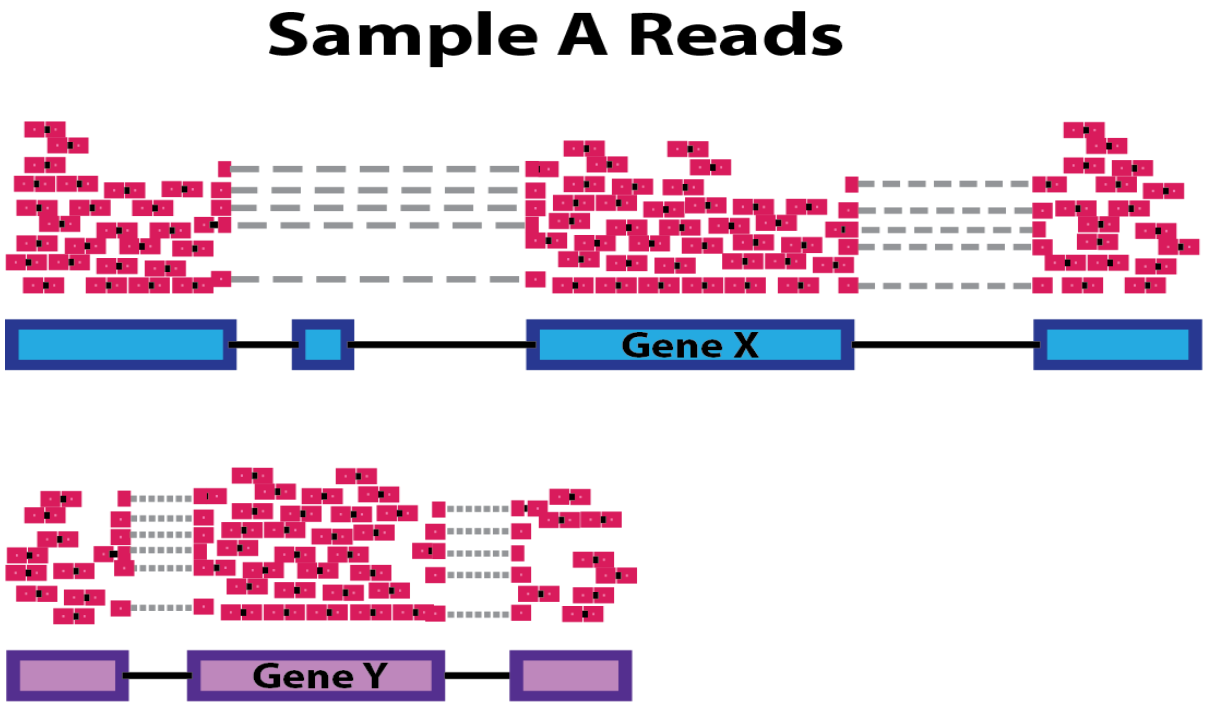
The main factors often considered during normalization are:



$$\text{CPM} = \left(\frac{\text{Raw Counts}}{\text{Total Mapped Reads}} \right)$$

- Counts per million (CPM) mapped reads are the number of raw reads mapped to a transcript, scaled by the number of sequencing reads in your sample, multiplied by a million.
- It normalizes RNA-seq data for sequencing depth but not gene length.
- Therefore, although it is a within sample normalization approach, CPM normalization is unsuitable for within sample comparisons of gene expression.
- Between sample comparisons can be made when CPMis used alongside ‘within a dataset’ normalization methods.

Normalization: Transcripts per Million (TPM)



The calculation involves two main steps, often described as normalizing for gene length first, then sequencing depth:

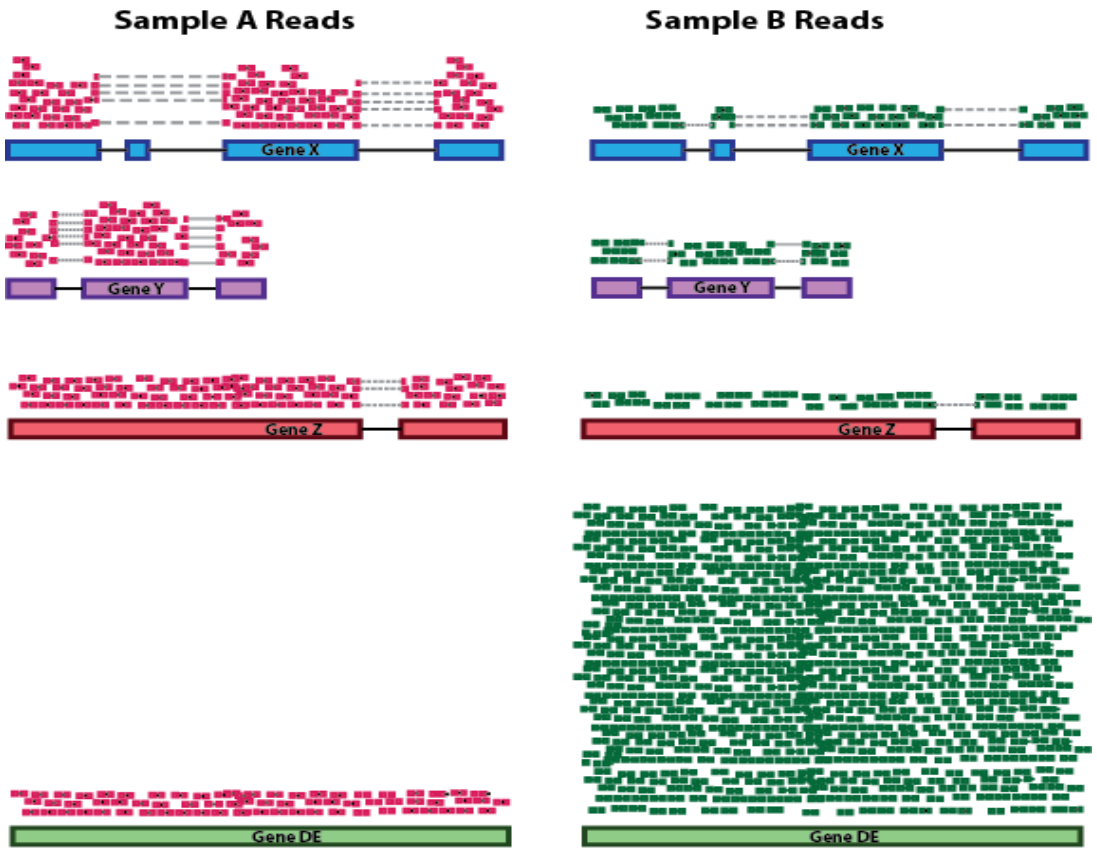
1. **Calculate Reads Per Kilobase (RPK)** for each gene (i):

$$\mathrm{RPK}_i = \frac{\text{Mapped Reads}_i}{\text{Length of gene } i \text{ in kb}}$$

2. **Calculate TPM** for gene (i):

$$\mathrm{TPM}_i = \left(\frac{\mathrm{RPK}_i}{\sum_j \mathrm{RPK}_j} \right) \times 10^6$$

Normalization: RPKM



RPKM

- Reads per kilobase per million =

$$\frac{\text{reads for gene } A}{\text{length of gene } A \text{ X Total number of reads}}$$

Oshlack, A. & Wakefield, M.J. (2009) *Biology Direct*

Normalization: DESeq2: Median of Ratios Normalization

DESeq2 uses the "median of ratios" method (Anders & Huber, 2010), accounting for sequencing depth and RNA composition.

1	**Pseudo-reference**	For each gene, compute geometric mean across all samples = pseudo-reference.
2	**Calculate ratios**	For each gene per sample: ratio = count / geometric_mean. Most genes aren't DE, so ratios reflect technical scaling.
3	**Median = size factor**	For each sample, take median of all gene ratios. This is the size factor (s_j), robust to DE outliers.
4	**Normalize**	Divide each raw count by its sample's size factor: normalized_ij = count_ij / s_j

Key Assumption: The majority of genes are NOT differentially expressed between conditions.

أكاديمية كاوست
KAUST ACADEMY

Normalization answers:

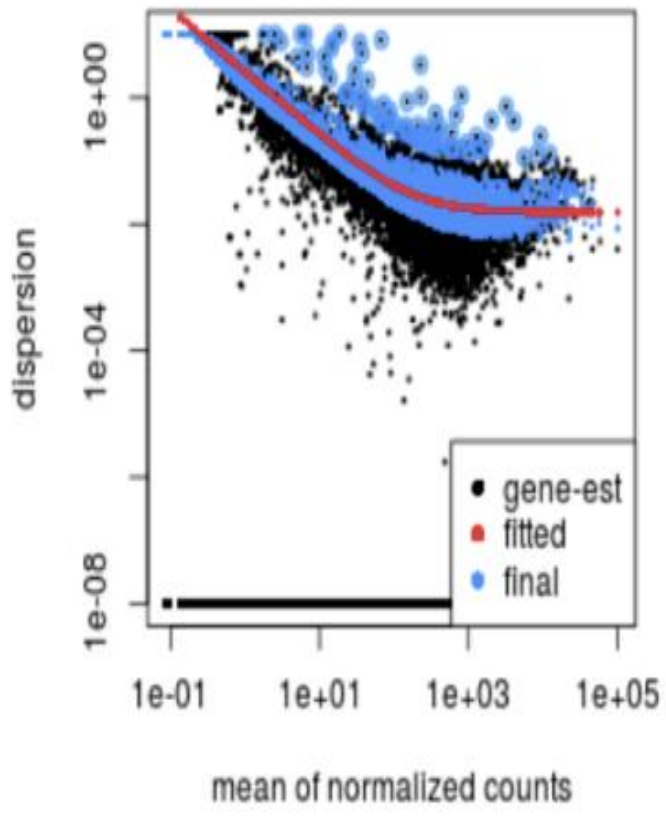
“Are samples comparable?”

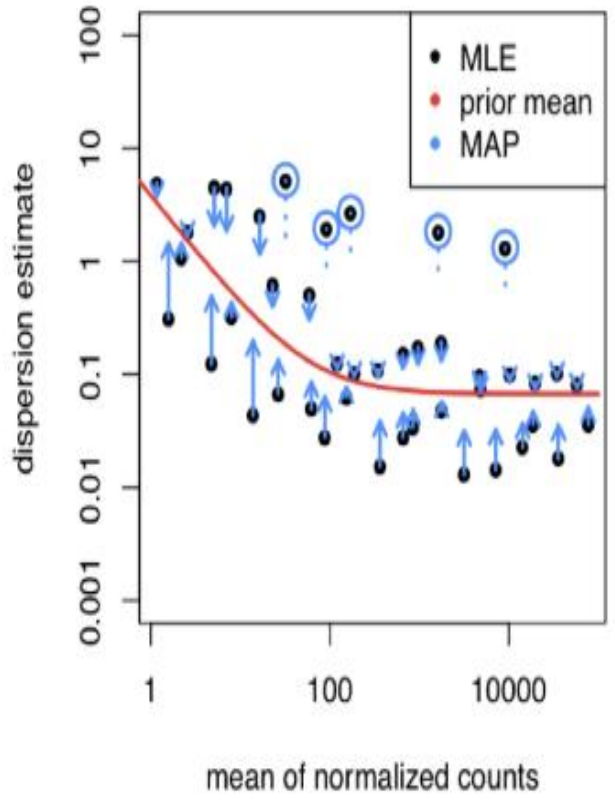
Differential expression answers:

“Which genes actually change between conditions?”

“Now the data are comparable… how do we detect biology?”

We Need Special Statistical Models

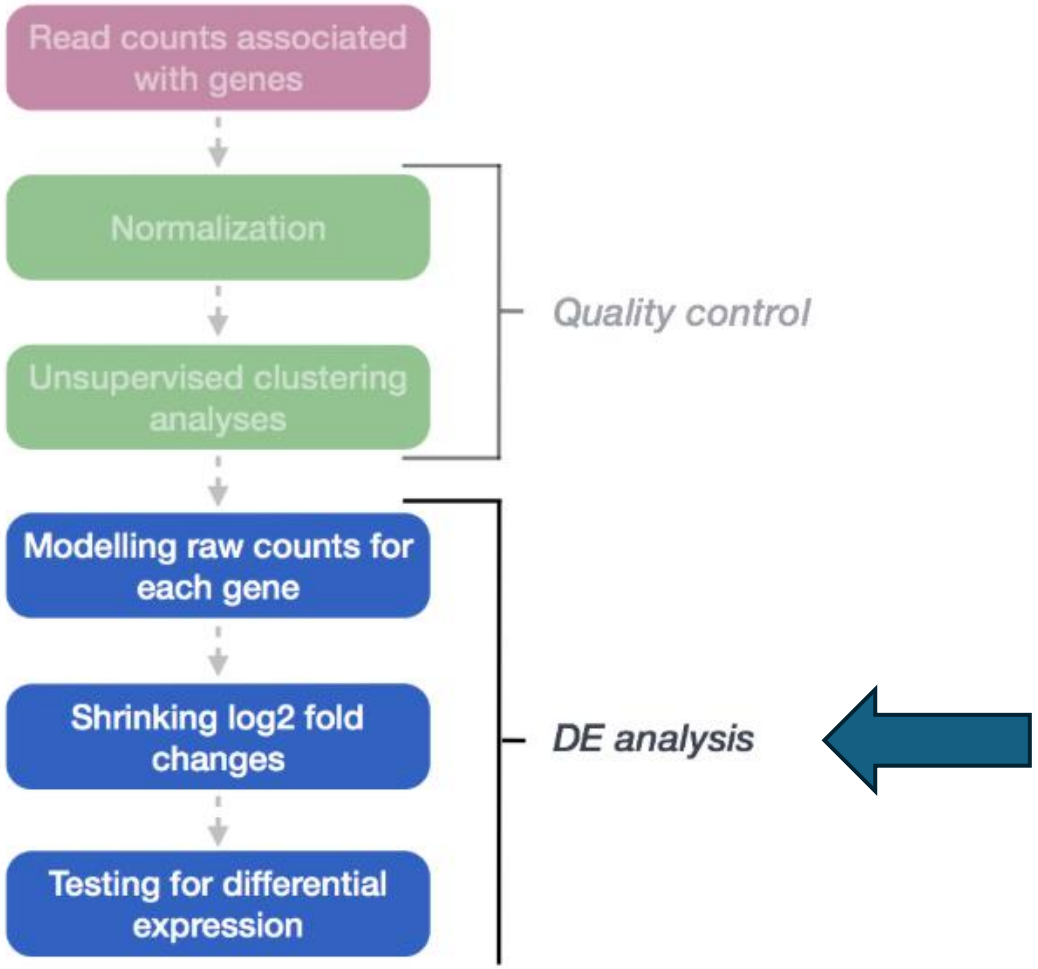




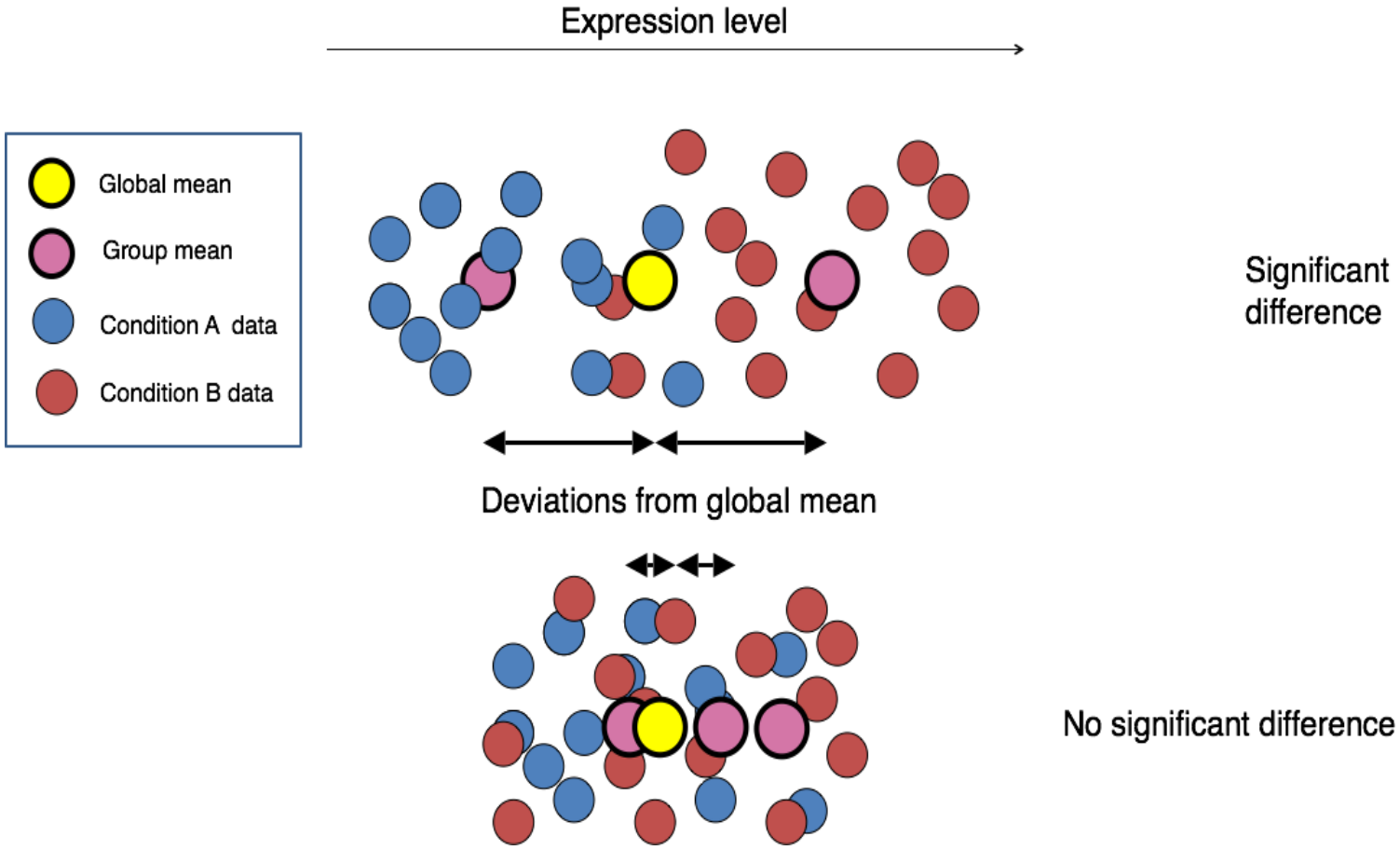
RNA-seq data

- Discrete counts
- Over-dispersed
- Not normally distributed

DEG pipeline



Differential expression analysis with DESeq2



Model RNA-seq count data using a Negative Binomial distribution to account for biological variability, then statistically test each gene to determine whether its expression differs significantly between experimental conditions.

We model how RNA-seq counts behave statistically.

Confounding factors: complex design

	sex	age	litter	treatment	treat_sex
sample1	M	11	1	Ctrl	CtrlM
sample2	M	13	2	Ctrl	CtrlM
sample3	M	11	1	Treat	TreatM
sample4	M	13	1	Treat	TreatM
sample5	F	11	1	Ctrl	CtrlF
sample6	F	13	1	Ctrl	CtrlF
sample7	F	11	1	Treat	TreatF
sample8	F	13	2	Treat	TreatF

DESeq2 also allows for the analysis of complex designs.

Example: we can explore the effect of sex on the treatment

```
design <- ~ sex + age + treatment + sex:treatment
```

Key Considerations for DE Analysis

Biological replicates essential	At least 3 per condition. Technical replicates do not substitute.
Use raw integer counts	Never use TPM/FPKM as input. DE tools normalize internally.
Filter lowly expressed genes	Remove genes with <10 total counts. Reduces testing burden.
Check for batch effects	Include batches in design formula. Use PCA to detect confounders.
Do NOT use fold-change alone	Large FC without stats is meaningless. Always report padj.
Do NOT use t-tests on counts	Violates normality. Use NB-based or voom-transformed methods.

Interpreting DE Results Table

Understanding each column of the DESeq2 results table is critical for proper interpretation.

Column	Description
`baseMean`	Average normalized count across all samples. Overall expression level.
`log2FoldChange`	Effect size. Positive = upregulated. log2FC=1 means 2-fold change.
`lfcSE`	Standard error of log2FC. Larger for lowly expressed genes.
`stat`	Wald statistic = log2FC / lfcSE. Distance from zero in SE units.
`pvalue`	Raw p-value. NOT corrected for multiple testing.
`padj`	BH-adjusted p-value (FDR). USE THIS for calling DE. Threshold: <0.05.

Typical DE criteria: padj < 0.05 AND |log2FC| > 1 (2-fold change)

DESeq2: Size of change: Log2FC

$$\log2FC = \log_2\left(\frac{\text{Expression in Condition B}}{\text{Expression in Condition A}}\right)$$

Gene	baseMean	log2FC	SE	p-value	FDR
GeneA	120	1.5	0.3	1e-6	1e-4
GeneB	15	−0.8	0.6	0.12	0.4

DESeq2 is an R package for analyzing count-based NGS data like RNA-seq.

Input

- Raw count matrix

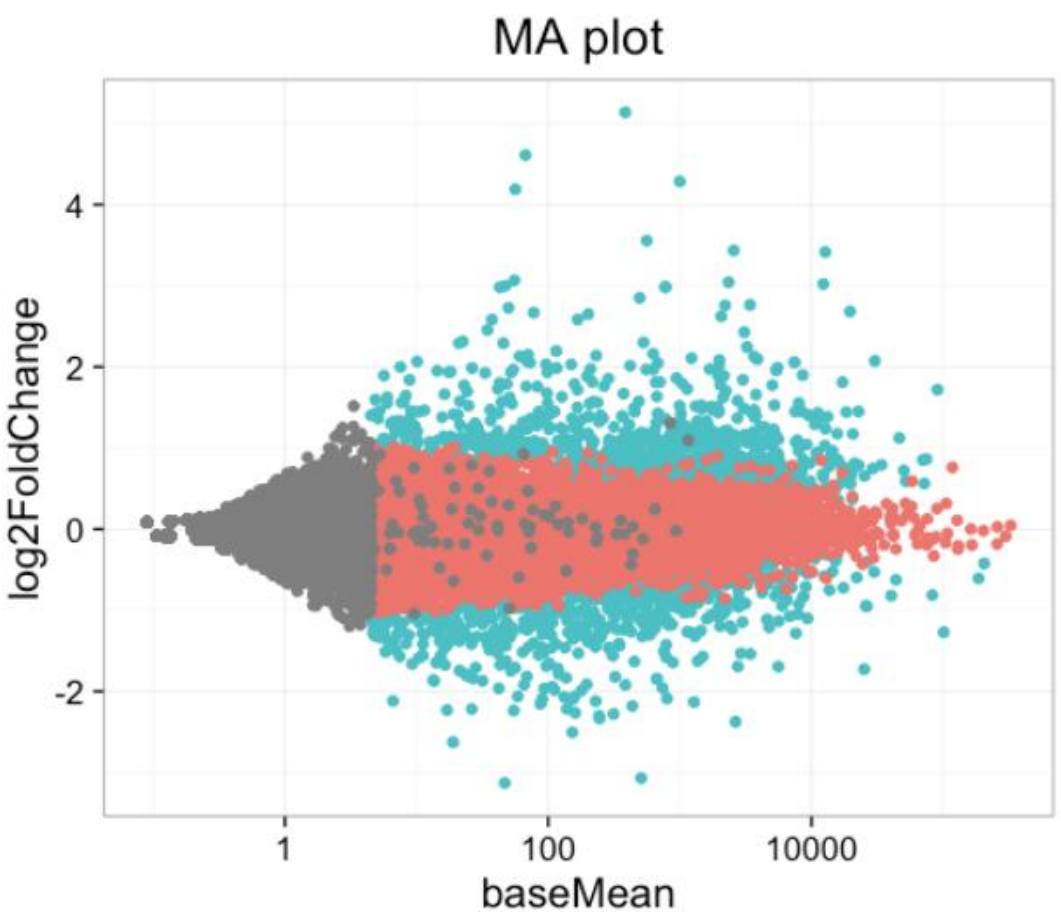
Output

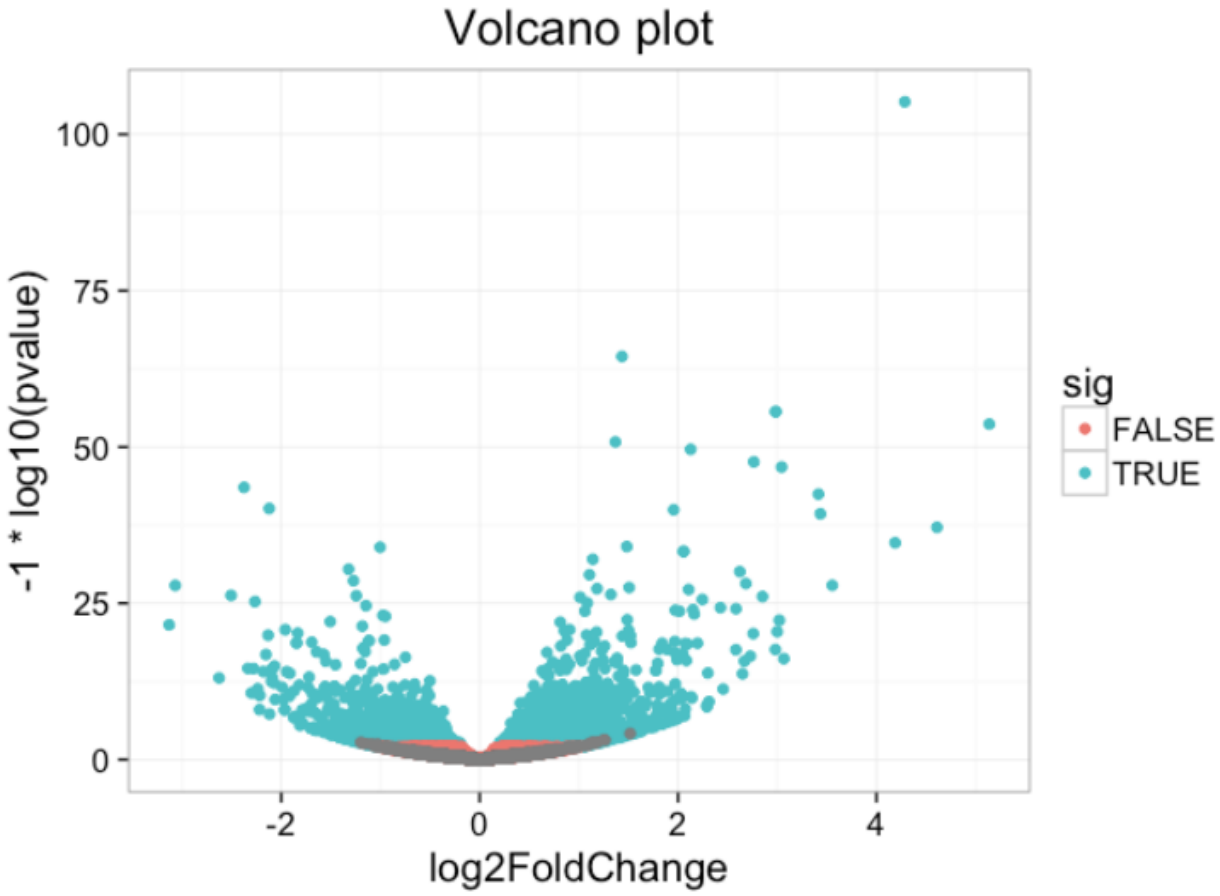
- Log2 Fold Change(effect size)
- p-value (statisticalevidence)

A gene is considered DE if:

- **|log2FC| ≥ 1**
- **padj ≤ 0.05**

Visualize the result:





Agenda – Day 03

أكاديمية كاوست
KAUST ACADEMY

Morning Session

9-9:15	S1: Recap of Day 2
9:15-10	S2: Differential Expression Concepts
10-11	*L1:* Perform differential expression analysis, inspect DE results
11-11:30	S3: Visualization & Reporting
11:30-12	L2: Visualization & Integrated QC

Afternoon Session

2-2:45	S4: Pathway & Enrichment Analysis
2:45-3:30	*L3: Hands-on with Transcriptomics Data*
3:30-4	S5: Extended Applications of RNA-seq
4-5	S6: Team Formation & Project Overview

Agenda – Day 03

أكاديمية كاوست
KAUST ACADEMY

Morning Session

9-9:15	S1: Recap of Day 2
9:15-10	S2: Differential Expression Concepts
10-11	*L1:* Perform differential expression analysis, inspect DE results
11-11:30	S3: Visualization & Reporting
11:30-12	L2: Visualization & Integrated QC

Afternoon Session

2-2:45	S4: Pathway & Enrichment Analysis
2:45-3:30	*L3: Hands-on with Transcriptomics Data*
3:30-4	S5: Extended Applications of RNA-seq
4-5	S6: Team Formation & Project Overview

أكاديمية كاوست
KAUST ACADEMY

Towards Biological Meaning: Visualization & Reporting

Day 03 – Session 03

Good analysis without good visualization is invisible science.

Sample-Level QC: PCA & Clustering

Assess data quality before DE. Identify outliers, batch effects, sample swaps.

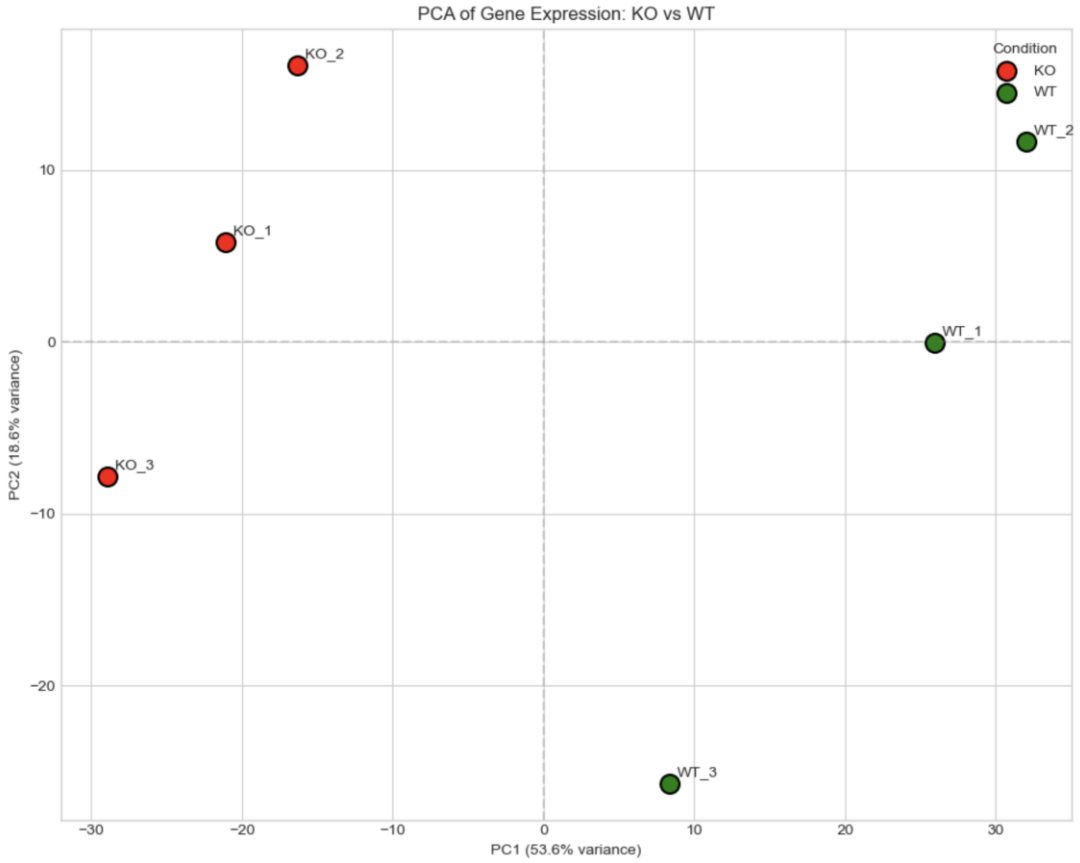
PCA Plot

Reduces gene dimensions to 2D.

Look for:

- Separation by condition
- Tight replicate clustering
- Outliers far from group
- Batch effects

Use plotPCA() on vst/rlog data.
PC1 captures most variance.



Always transform data first

Sample-Level QC: PCA & Clustering

Assess data quality before DE. Identify outliers, batch effects, sample swaps.

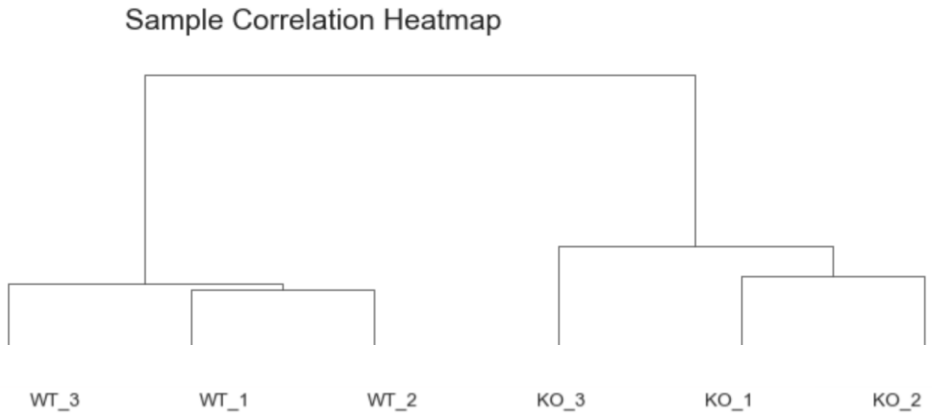
Hierarchical Clustering

Sample-to-sample distance heatmap.

Look for:

- Samples cluster by condition
- Dark diagonal = similar replicates
- Unexpected patterns = problems

Use Euclidean distance on variance-stabilized data.
pheatmap or ComplexHeatmap.



Always transform data first

Volcano Plot

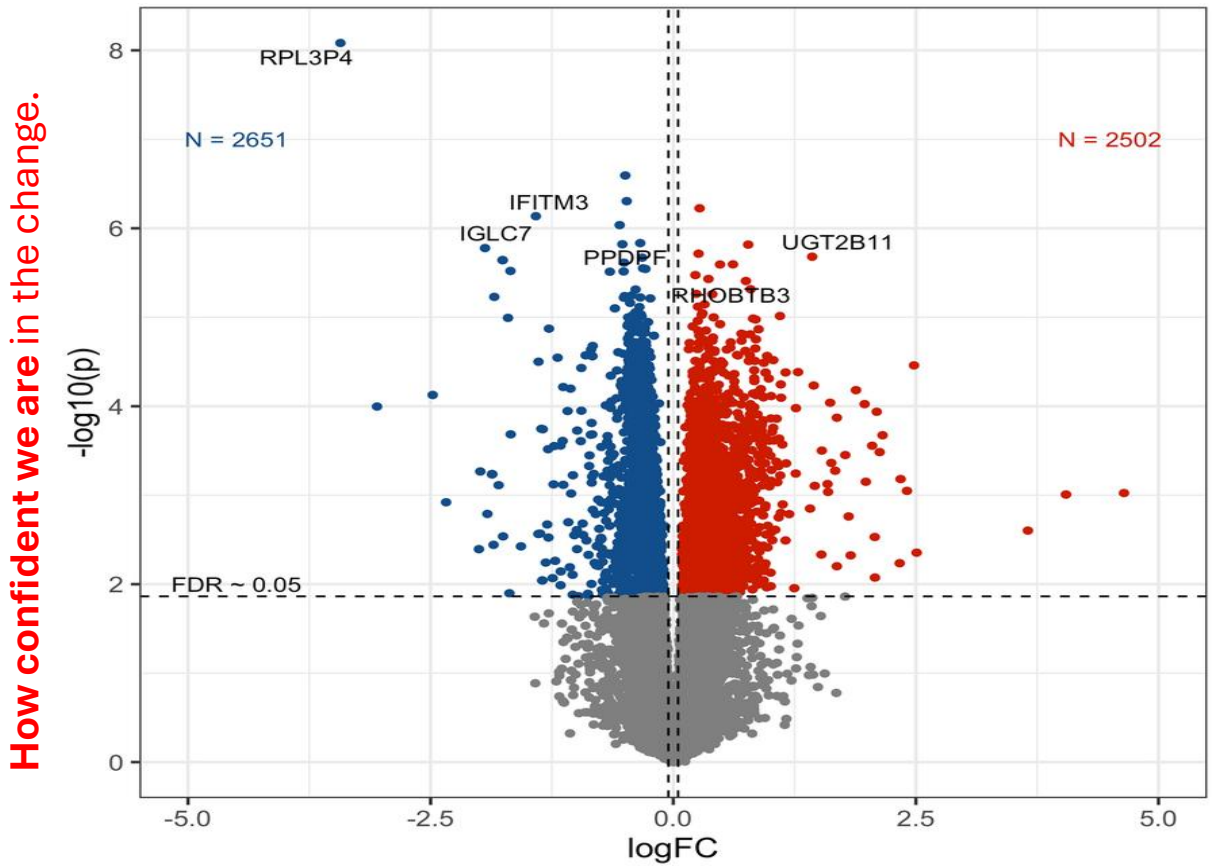
Most common DE visualization: statistical significance vs. magnitude of change for every gene.

Axes & Interpretation

X-axis: log2 Fold Change
- Positive = up in treatment
- Negative = down in treatment

Y-axis: –log10(adjusted p-value)
- Higher = more significant
- Cutoff at –log10(0.05) ≈ 1.3

- **Which genes changed between two conditions? (Biologically)**
- **How confident are we about that change? (Statistically)**
- **------- Threshold**



Log fold change: **how much a gene changed**.

MA Plot

Shows mean expression vs. fold change. Reveals expression-dependent bias and effect of LFC shrinkage.

X: baseMean (log10 scale)
  Left = low expression, right = high

Y: Log2 Fold Change
  Cloud centered at y=0 = good normalization
  Low-count genes (left) show high variance

Heatmaps for Gene Expression

Display expression patterns across samples for selected genes, revealing clusters and co-expression patterns.

Best Practices

Input: transformed values, NOT raw counts

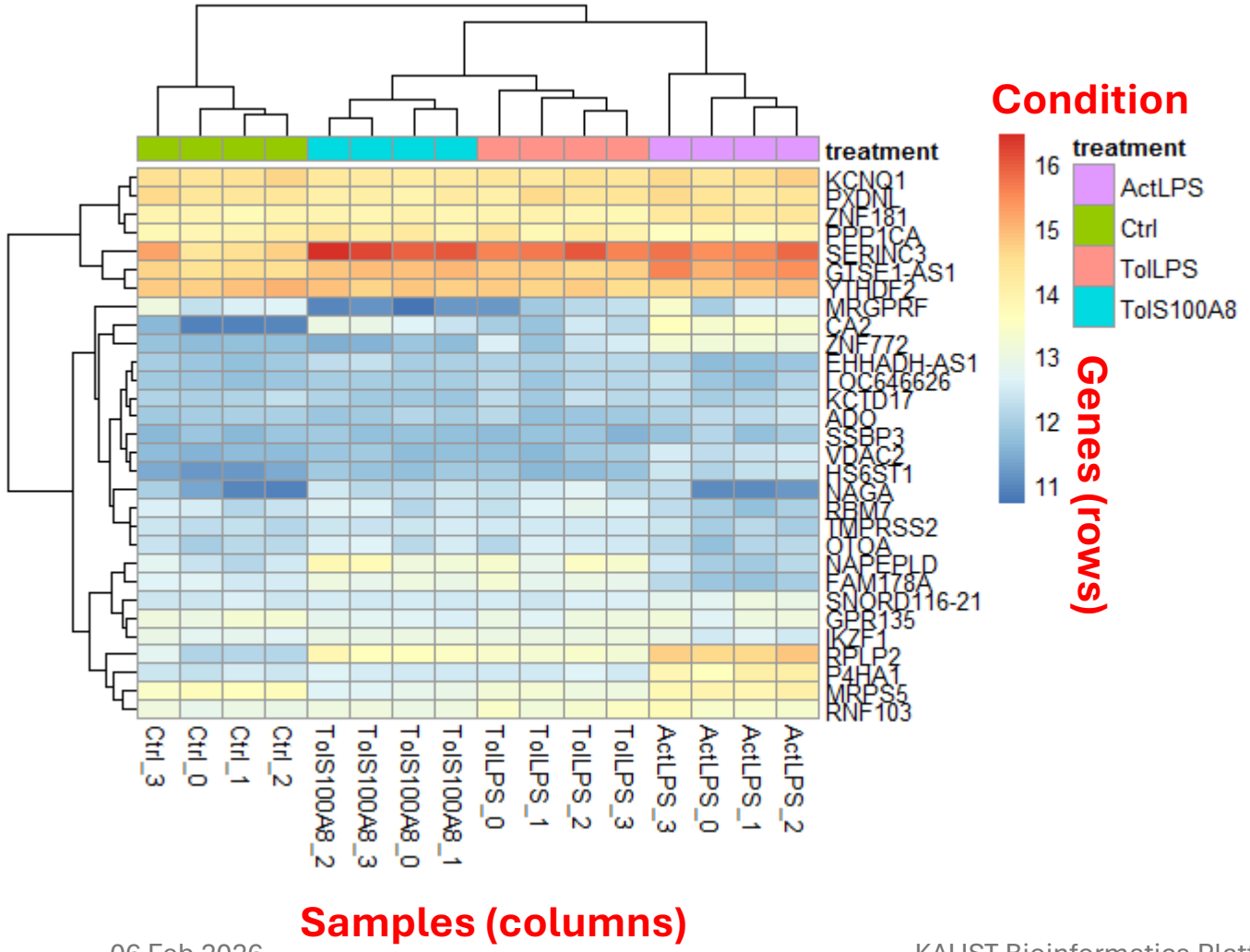
Scaling: Z-score by row (gene) to focus on relative patterns

Gene selection: Top 20–50 DE genes by padj, or genes of interest.

Annotations: Column annotations for condition, batch, etc.

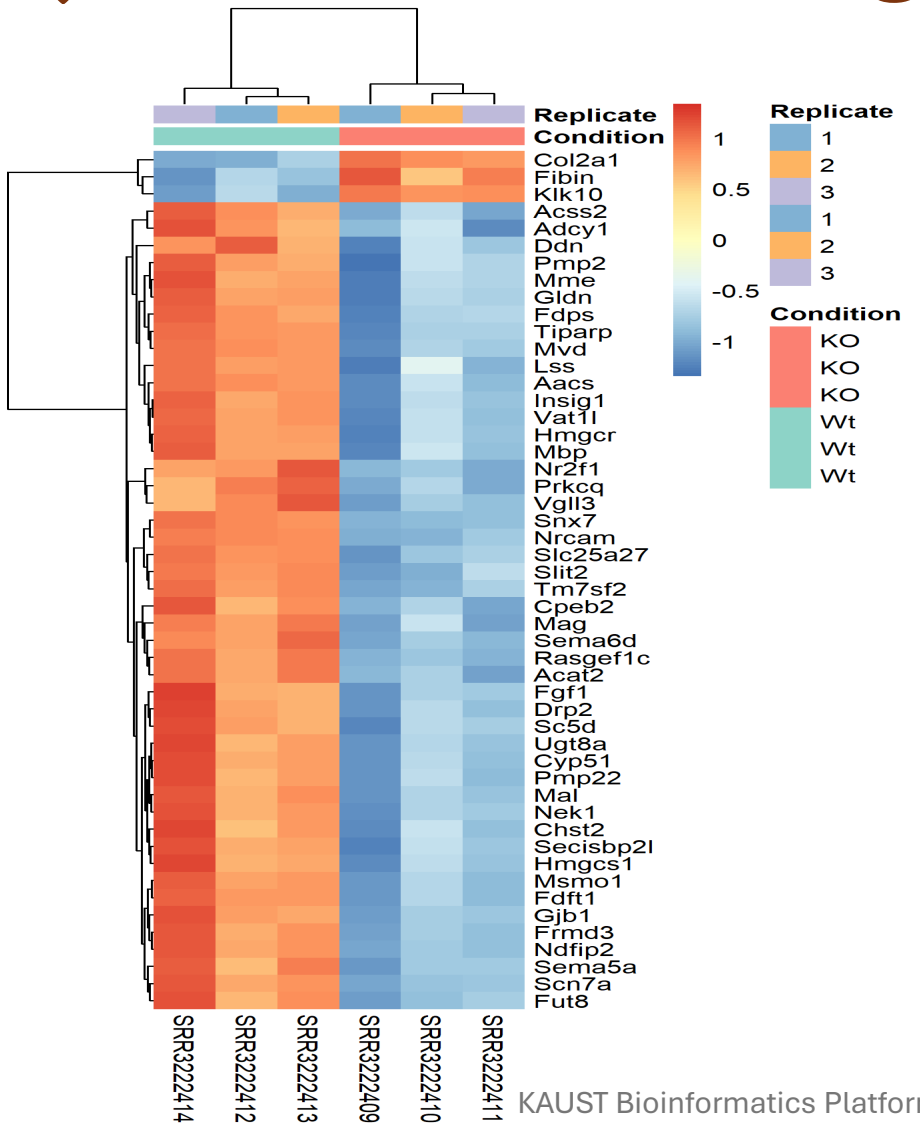
Clustering: Hierarchical clustering of rows + columns reveals modules

Heatmaps and Clustering



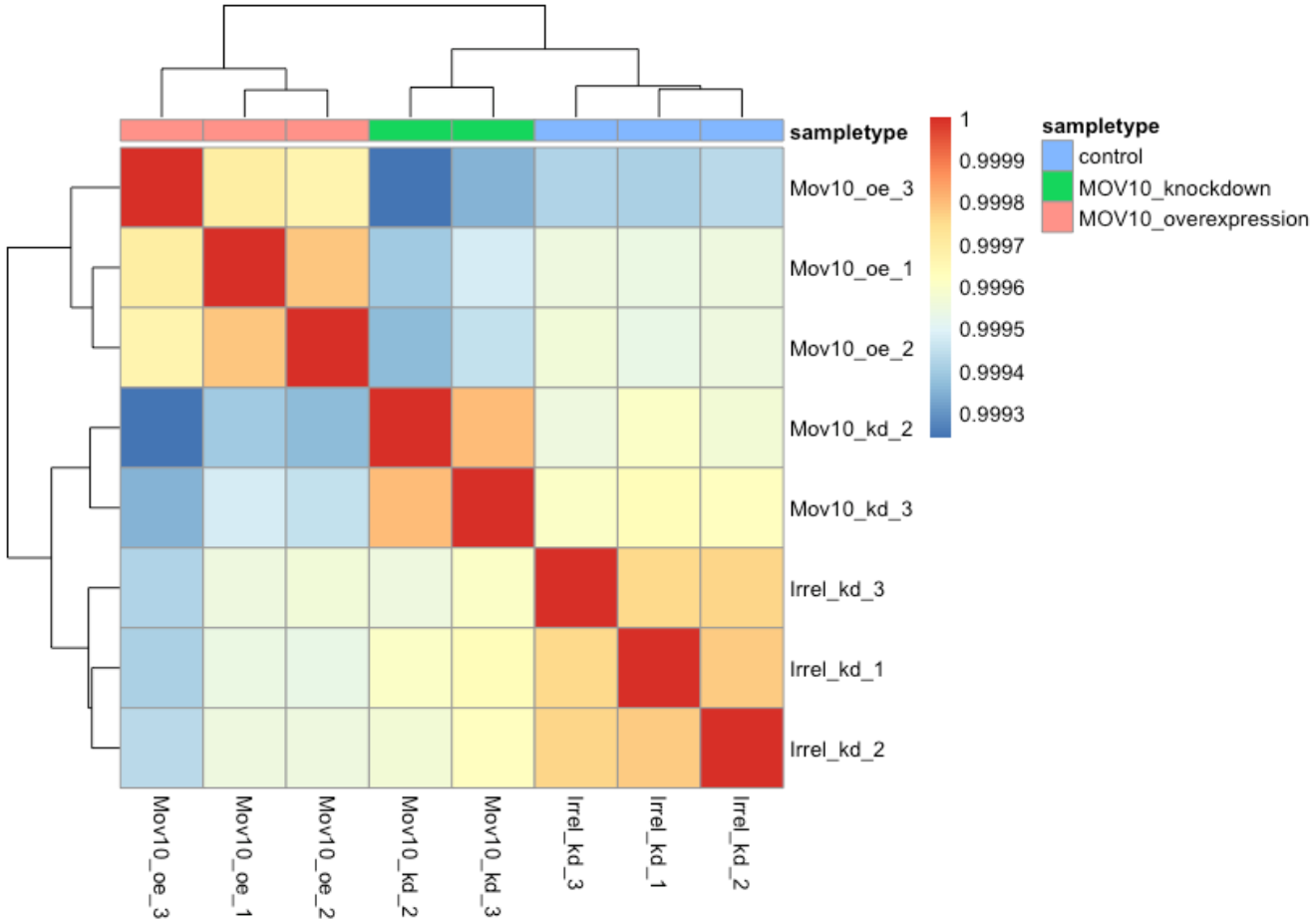
1. Heatmap of top genes with clear clustering by condition
2. Which genes behave similarly across samples
3. Do top DE genes show a clear pattern?

Heatmaps and Clustering



- Do samples cluster by condition?
- heatmap where the top annotation bars clearly label Condition and Replicate.
- how to read the annotation bars and how clustering supports the experimental design.

Heatmaps and Clustering



- Sample-to-sample correlation heatmap
- Correlation matrix heatmap showing samples clustering nicely by group.
- Usually uses for “QC check: do replicates look similar?”

Agenda – Day 03

أكاديمية كاوست
KAUST ACADEMY

Morning Session

9-9:15	S1: Recap of Day 2
9:15-10	S2: Differential Expression Concepts
10-11	*L1:* Perform differential expression analysis, inspect DE results
11-11:30	S3: Visualization & Reporting
11:30-12	L2: Visualization & Integrated QC

Afternoon Session

2-2:45	S4: Pathway & Enrichment Analysis
2:45-3:30	*L3: Hands-on with Transcriptomics Data*
3:30-4	S5: Extended Applications of RNA-seq
4-5	S6: Team Formation & Project Overview

أكاديمية كاوست
KAUST ACADEMY

Pathway & Enrichment Analysis

Day 03 – Session 04

From Gene Lists to Biological Meaning

DE gives a gene list. Enrichment analysis connects it to known biology: pathways, processes, functions.

The Problem

500 DE genes. Looking at individuals is overwhelming. A pathway effect may be distributed across many genes with small changes.

The Solution

Enrichment asks: "Are genes from a biological pathway over-represented in my results more than by chance?" Provides higher-level interpretation.

Three Generations of Methods

1st: ORA
(Over-Representation)

2nd: FCS / GSEA
(Functional Class Scoring)

3rd: Topology-Based
(Network context)

Knowledge Bases for Enrichment

أكاديمية كاوست
KAUST ACADEMY

Enrichment relies on curated databases defining gene sets.

Gene Ontology (GO)	BP, MF, CC ontologies. Hierarchical DAG. ~45,000 terms.
KEGG Pathways	Curated metabolic/signaling/disease pathways. ~350 human pathways.
Reactome	Open-source, peer-reviewed. Hierarchical. ~2,600 human pathways.
MSigDB (Broad)	8 collections: Hallmark (50 sets), curated, motif, GO, oncogenic, immunologic, cell type.
WikiPathways	Community-curated open pathway database with visual diagrams.

Use multiple databases for comprehensive analysis.

Gene Ontology: Structure & Three Domains

أكاديمية كاوست
KAUST ACADEMY

GO is a DAG: child terms are more specific. A gene annotated to a child inherits all parent annotations.

Biological Process (BP)

What the gene does biologically.
Ex: apoptosis, immune response, cell cycle, DNA repair

Molecular Function (MF)

Biochemical activity.
Ex: kinase activity, DNA binding, receptor activity

Cellular Component (CC)

Where in the cell.
Ex: nucleus, mitochondrion, plasma membrane

Redundancy is common in GO results. Use REVIGO or rrvgo to summarize and reduce redundant terms.

Over-Representation Analysis (ORA)

Tests whether pathway genes appear more often in DE list than expected by chance (2×2 contingency table).

	In Gene Set	Not in Set	Total
DE Genes	**x**	n–x	n
Non-DE	M–x	rest	N–n
Total	M	N–M	N

Fisher's Exact / Hypergeometric

N = background genes
n = DE genes
M = genes in set S
x = DE genes in set S

Tests: is x > expected by chance?

Key pitfalls: (1) Background must be expressed genes, not full genome; (2) Arbitrary DE threshold affects results; (3) Assumes gene independence (often violated)

ORA tools: clusterProfiler, g:Profiler, DAVID, Enrichr, PANTHER

Gene Set Enrichment Analysis (GSEA)

Uses the ENTIRE ranked gene list (no threshold). Tests whether pathway genes cluster at the top or bottom.

1. **Rank all genes**

   Order by metric (Wald stat, shrunken log2FC, signed –log10(p)×sign(log2FC), or t-statistic).

2. **Walk down ranked list**

   Running-sum: +increment for gene set members (weighted by rank metric), –decrement otherwise.

3. **Enrichment Score**

   Max deviation from zero. Positive ES = members at TOP (upregulated). Negative = at BOTTOM (downregulated).

4. **Significance testing**

   Compare ES to null from gene label permutation. NES = ES / mean(null). FDR for multiple testing.

Subramanian et al. (2005) PNAS | Tools: fgsea, clusterProfiler::GSEA(), GSEA (Broad Institute)

ORA vs GSEA: When to Use Which

أكاديمية كاوست
KAUST ACADEMY

Feature	ORA	GSEA
Input	DE gene list (binary: DE or not)	All genes ranked by metric
Threshold	Requires arbitrary cutoff	No threshold needed
Quantitative info?	No — membership only	Yes — uses magnitude of change
Subtle effects?	Misses many-gene-small-change	Detects coordinated small changes
Direction?	Not inherent (split up/down)	Yes — positive/negative NES
Speed	Very fast	Slower (permutation)
Best for	Quick analysis; clustering results; non-ranked lists	DE results with rankings; subtle pathway effects

Recommendation: If you have DE results with gene rankings, prefer GSEA. Use ORA for quick checks or non-ranked lists.

Critical: Choosing the Right Background

For ORA, background gene set dramatically affects results. Wrong choice inflates false positives.

WRONG: Full genome as background

Using ~20,000 genes when only 12,000 expressed. Inflates denominator → false enrichment of tissue-specific pathways.

RIGHT: Expressed genes only

Use all detected/expressed genes (passing your pre-filter). Asks: "Among expressed genes, are DE genes enriched in this pathway?"

This is the #1 most common mistake in enrichment analysis. Always specify the 'universe' parameter.

Visualizing Enrichment Results

أكاديمية كاوست
KAUST ACADEMY

Dot Plot

Gene sets vs gene ratio. Size=count, color=padj.

```
enrichplot::dotplot()
```

Bar Plot

Horizontal bars for enriched terms. Clean for talks.

```
barplot(ego)
```

GSEA Plot

Running sum + barcode + rank metric for one set.

```
plotEnrichment()
```

Enrichment Map

Network of overlapping gene sets. Reveals themes.

```
enrichplot::emapplot()
```

Concept Network

Genes shared across terms. Identifies hub genes.

```
enrichplot::cnetplot()
```

Pathway Render

KEGG diagram colored by DE. Shows gene interactions.

```
pathview::pathview()
```

Use enrichplot and clusterProfiler for publication-quality figures. Combine multiple views for full picture.

Common Pitfalls in Enrichment Analysis

أكاديمية كاوست
KAUST ACADEMY

Avoid these mistakes to prevent misleading results.

✗ Full genome as ORA background	✓ Use only expressed/detected genes
✗ Reporting raw p-values	✓ Always apply and report FDR correction
✗ Cherry-picking one enriched term	✓ Report all significant; reduce redundancy
✗ Running ORA when GSEA is better	✓ Use GSEA when you have DE rankings
✗ Ignoring gene ID mapping issues	✓ Verify conversions; check unmapped genes
✗ Using only one database	✓ Cross-validate: GO + KEGG + Reactome
✗ Ignoring gene set size	✓ Filter: keep 15–500 genes per set

Practical Tips for Interpretation

أكاديمية كاوست
KAUST ACADEMY

Look for themes, not individual terms

Are multiple immune terms enriched? That's more convincing than one. Cluster related results.

Check gene overlap across terms

Use cnetplot/UpSet plots. If one gene drives many terms, may be artifact.

Validate with orthogonal data

Cross-reference with protein data, literature, independent datasets. Enrichment is hypothesis-generating.

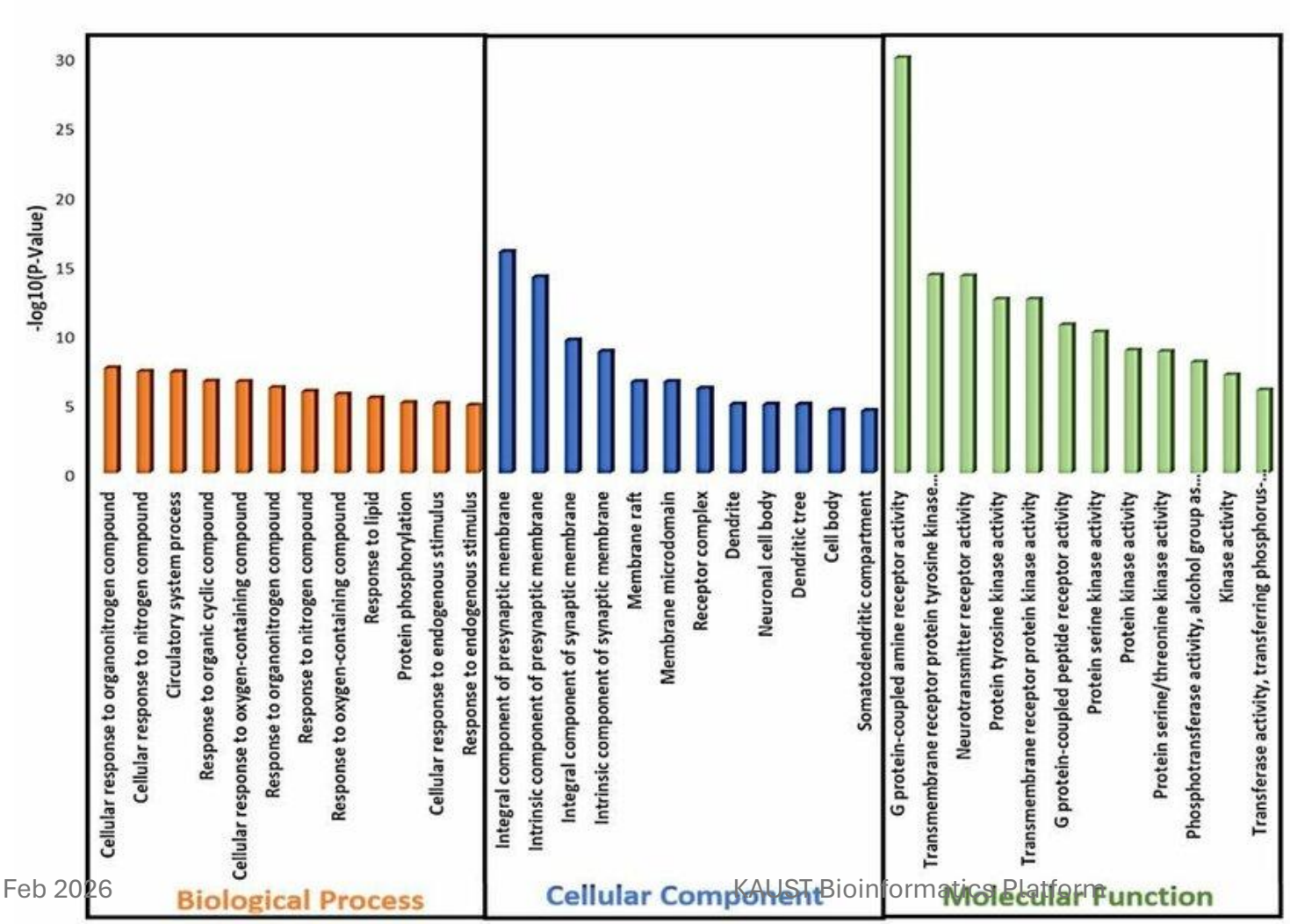
Consider biological context

Does the enrichment make biological sense given your experimental system? Unexpected results need careful scrutiny.

Report methods transparently

Database(s), method (ORA/GSEA), ranking metric, background, correction, software versions.

Gene Ontology (GO)



What does this gene do?

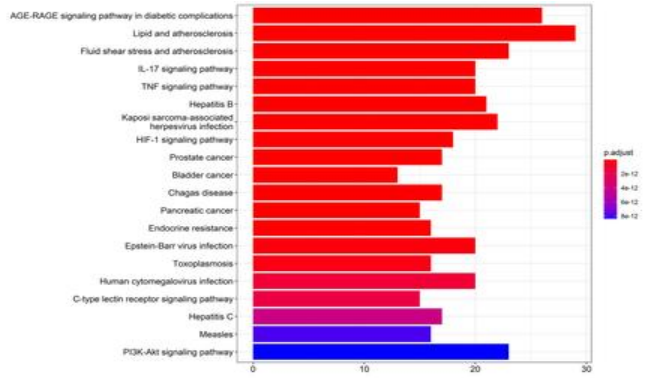
- **Biological Process** → what is happening
  (e.g. cell division, neuron signaling)

- **Molecular Function** → what the gene does
  (e.g. enzyme, receptor)

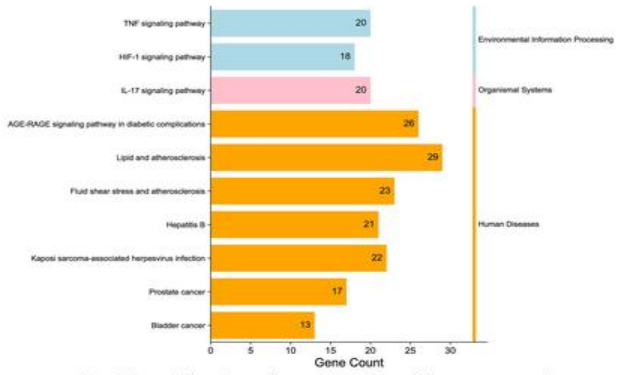
- **Cellular Component** → where it works
  (e.g. nucleus, synapse)

Think of GO as **labels describing gene functions**.

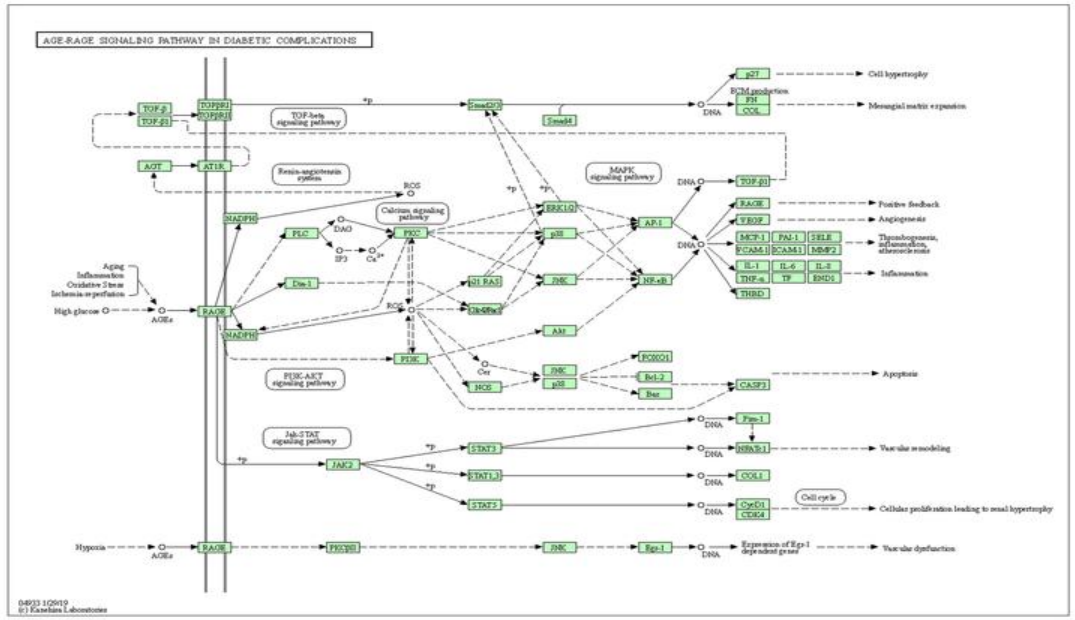
KEGG



A. KEGG signal pathway diagram



B. Classification diagram of enrichment result



C. AGE-RAGE signaling pathway in diabetic complications pathway

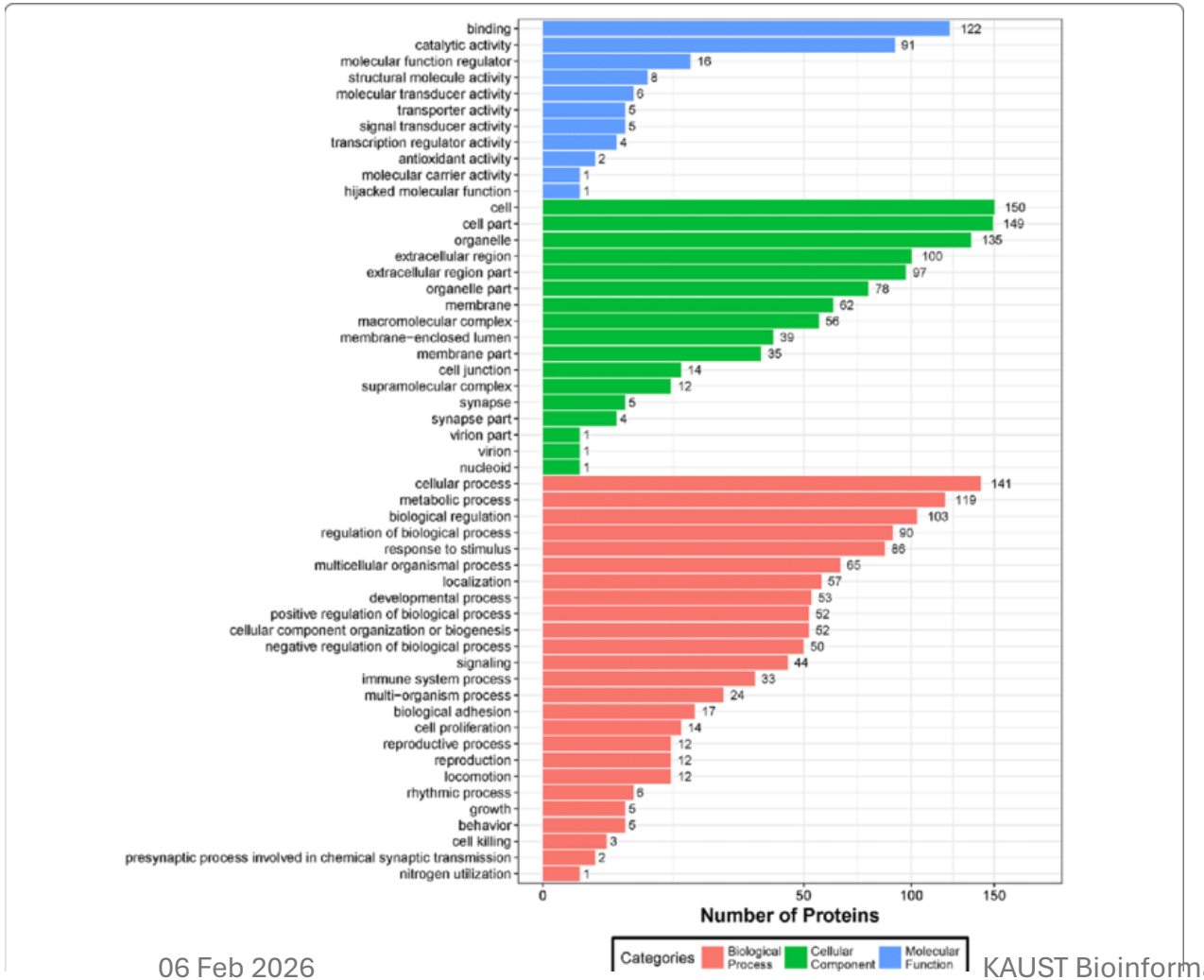
How do genes work together?

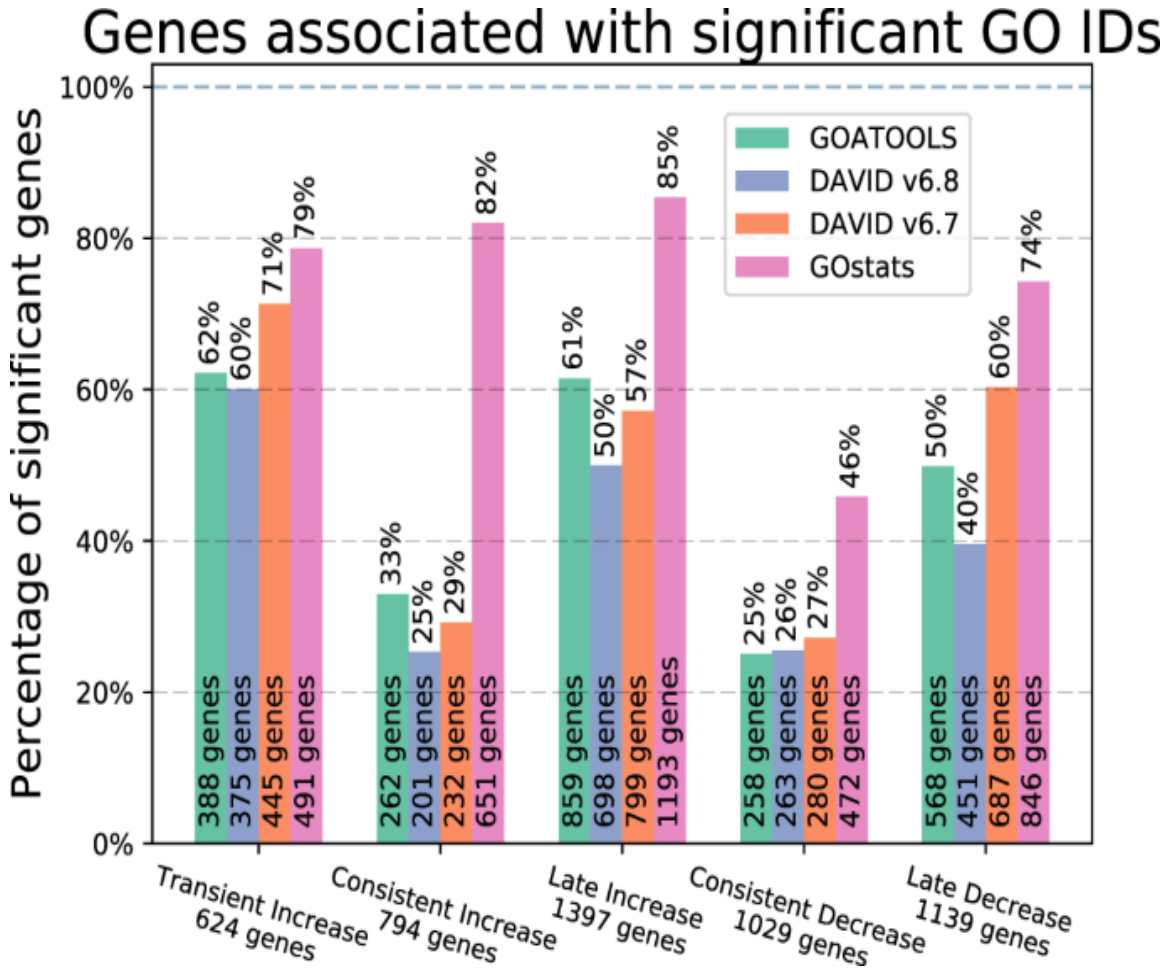
- Shows **pathways** (step-by-step biological processes)

- Genes are connected by **arrows and interactions**
- Examples: signaling pathways, metabolism pathways

Think of KEGG as a **map of biological pathways**.

Example of Dot Plot for Enrichment





Agenda – Day 03

أكاديمية كاوست
KAUST ACADEMY

Morning Session

9-9:15	S1: Recap of Day 2
9:15-10	S2: Differential Expression Concepts
10-11	*L1:* Perform differential expression analysis, inspect DE results
11-11:30	S3: Visualization & Reporting
11:30-12	L2: Visualization & Integrated QC

Afternoon Session

2-2:45	S4: Pathway & Enrichment Analysis
2:45-3:30	*L3: Hands-on with Transcriptomics Data*
3:30-4	S5: Extended Applications of RNA-seq
4-5	S6: Team Formation & Project Overview

أكاديمية كاوست
KAUST ACADEMY

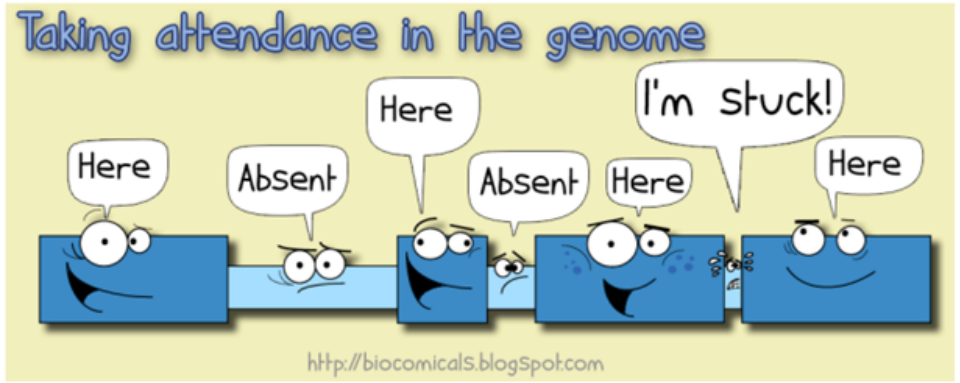
Extended Applications of RNA-seq &Emerging Transcriptomics Domains

Day 03 – Session 05

RNA-Seq applications

Discovery

- Transcripts
- Isoforms
- Splice junctions
- Fusion genes



Differential expression

- Gene level expression changes
- Relative isoform abundance
- Splicing patterns

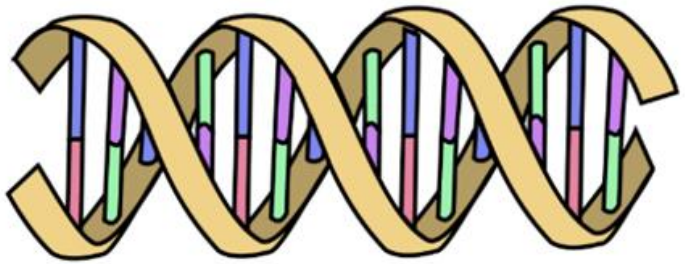
Variant calling

أكاديمية كاوست
KAUST ACADEMY

Introduction to Variant Calling

Size of Base Pairs

DNA



3 billions base pair long

....ATTTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....

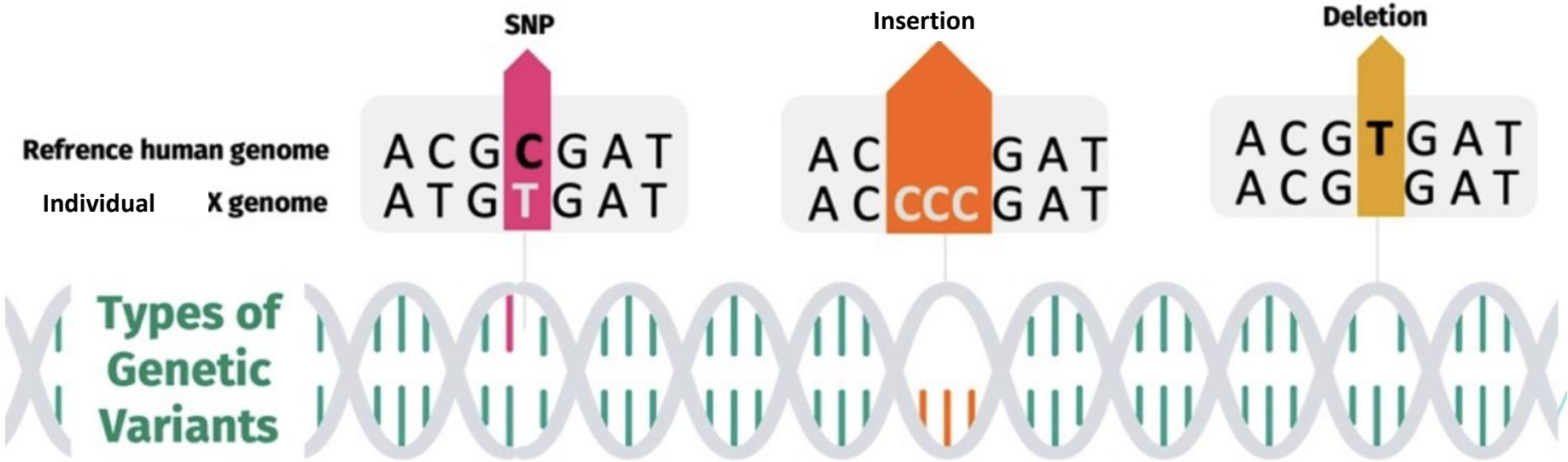
....TAAACGGTCAGTCATGGGTCCTACGACCTTGCCTA....

What is a variant?

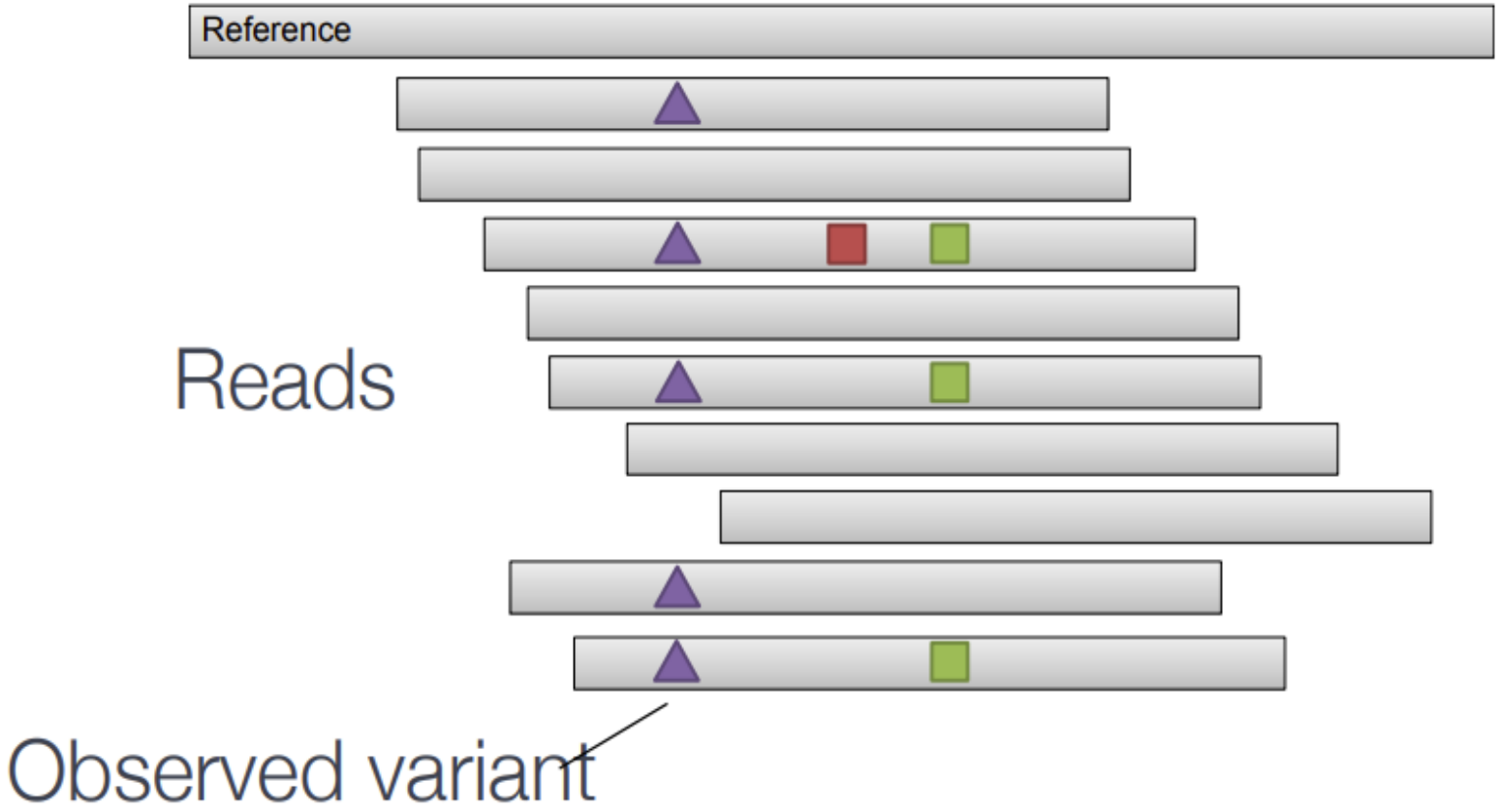
A variant refers to a difference or alteration in the DNA sequence compared to a reference genome.

Types of Variants

Variants can occur in a single nucleotide (single nucleotide polymorphism, or SNP), involve insertions or deletions of nucleotides (indels), or include larger structural changes such as duplications or rearrangements.



Variant Calling & Filtering:

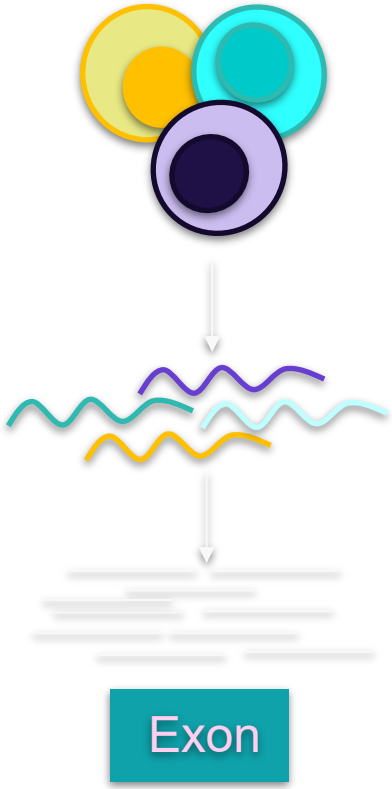


After alignment, the BAM file is analyzed to identify variants, such as SNPs (Single Nucleotide Polymorphisms) and INDELs (insertions/deletions), by comparing the aligned reads to the reference genome. Tools like GATK, Samtools, or FreeBayes are commonly used for this step.

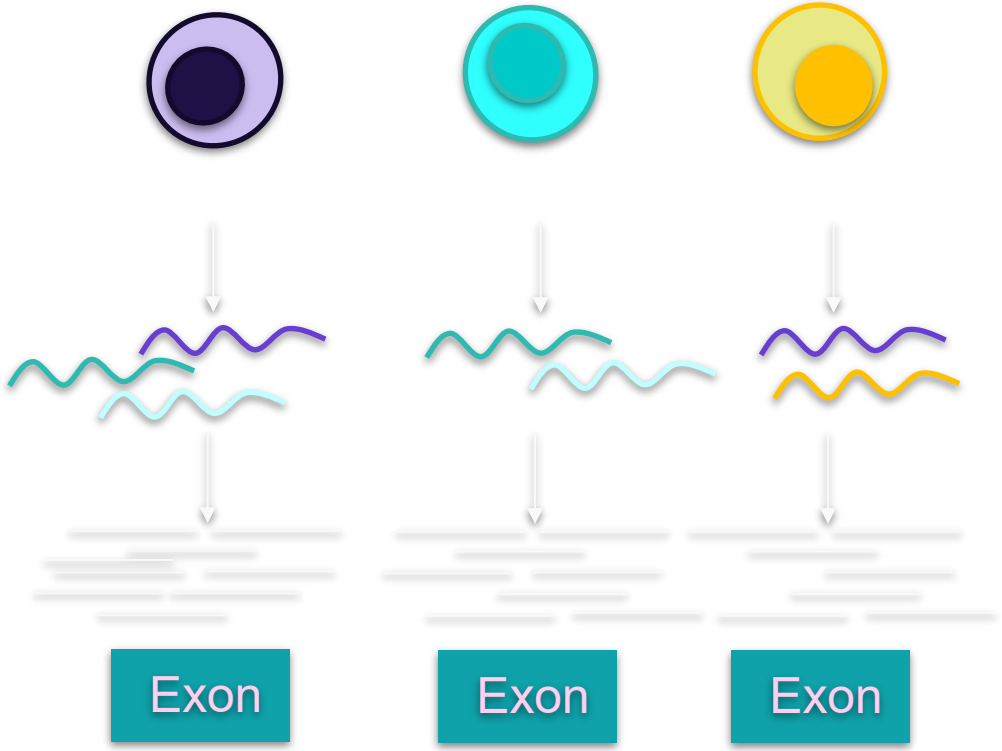
أكاديمية كاوست
KAUST ACADEMY

Introduction to scRNA

Bulk vs. Single-Cell RNA-Seq

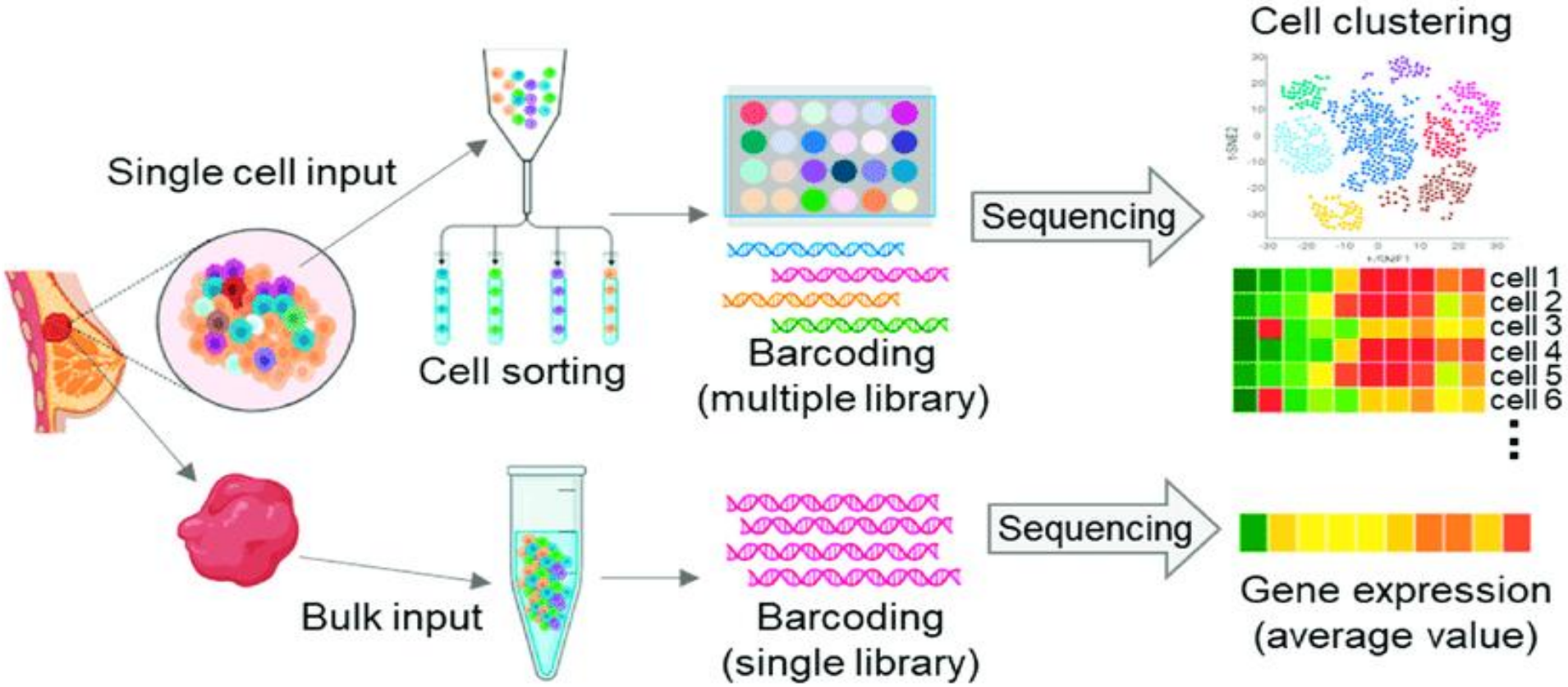


Bulk RNA seq



Single-Cell RNA seq

Bulk vs. Single-Cell RNA-Seq



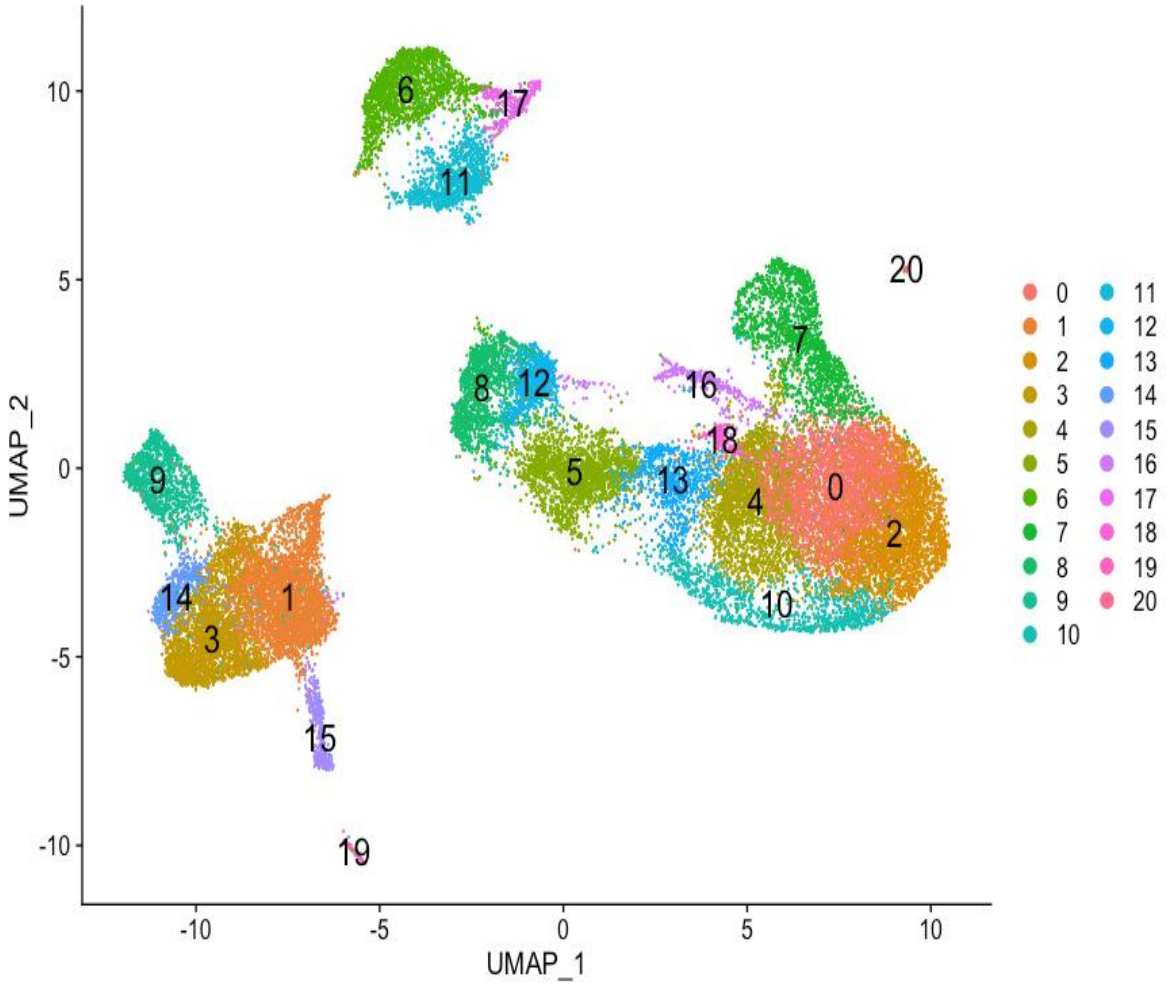
Differential Expression Analysis and Cell Annotation tools:

Single-cell Expression Analysis (scRNA-seq)

- Single-cell RNA sequencing (scRNA-seq) is a technology that enables the measurement of gene expression profiles at the individual cell level.

- scRNA-seq provides a comprehensive view of cellular heterogeneity and allows for the identification and characterization of distinct cell types and states.

- This technique has wide-ranging applications in biology, including the study of developmental processes, disease mechanisms, immune responses, and the discovery of novel cell populations and biomarkers.



Spatial transcriptomics

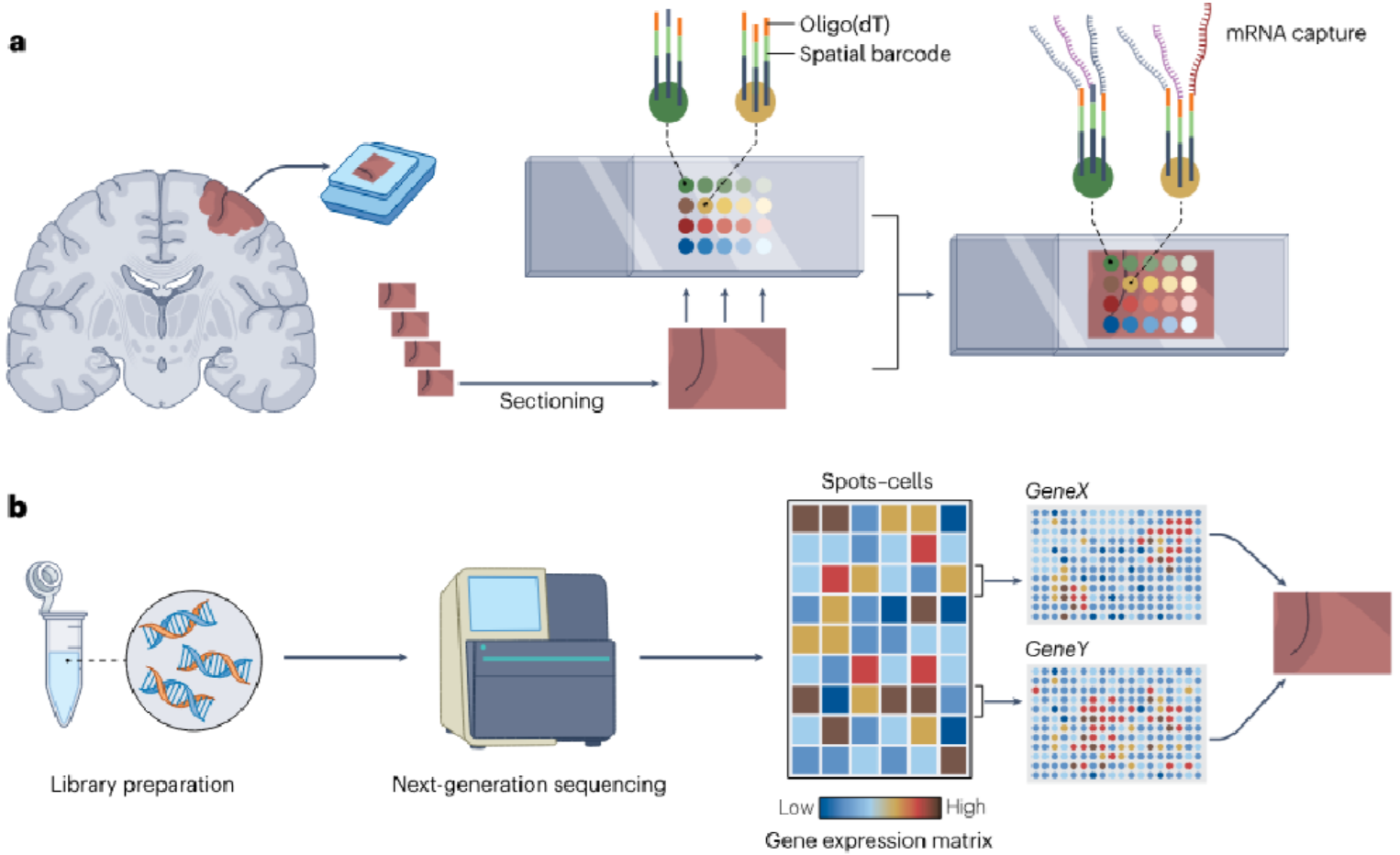


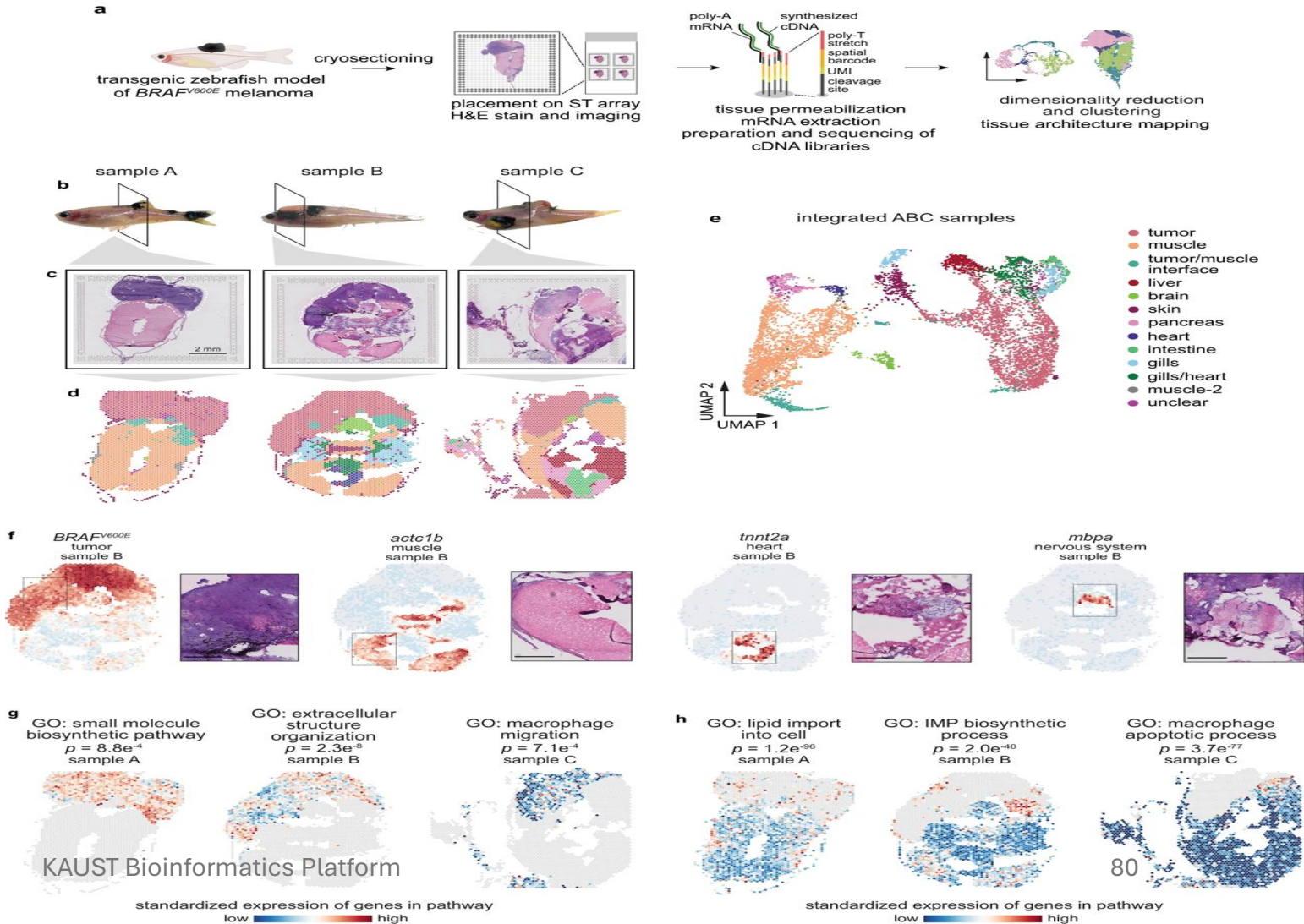
Fig. 3 | Spatial transcriptomics: the principle and a workflow. a, The tissue is subjected to cryosectioning on an mRNA capture slide, fixation and permeabilization to release RNA. The poly-A tail of the mRNA binds to an oligo(dT) ending fragment (single-stranded sequence of deoxythymines) on the capture DNA probes, which also contain embedded positional barcodes. **b**, After library preparation and sequencing, the computational analysis includes retrieval of the positional barcodes and tissue coordinates to reconstruct the relationship between transcripts and their locations.

Integrating spatial transcriptomics with single-cell transcriptomics

- Integrating spatial transcriptomics with single-cell transcriptomics is a powerful approach that combines the benefits of both technologies to unravel the spatial organization of gene expression within complex tissues.

- By overlaying spatial information onto single-cell gene expression data, this integrative approach provides a comprehensive understanding of cellular heterogeneity, cell-cell interactions, and tissue architecture.

- This technique enables the identification of spatially distinct cell populations, characterization of cell type distributions, and exploration of spatially coordinated gene expression patterns.



Future directions: Integrative omics approach

1- Sequencing DNA or RNA is not enough to understand a disease or define a cell type!

2- Integrate analysis using bioinformatics, AI, machine learning platforms.