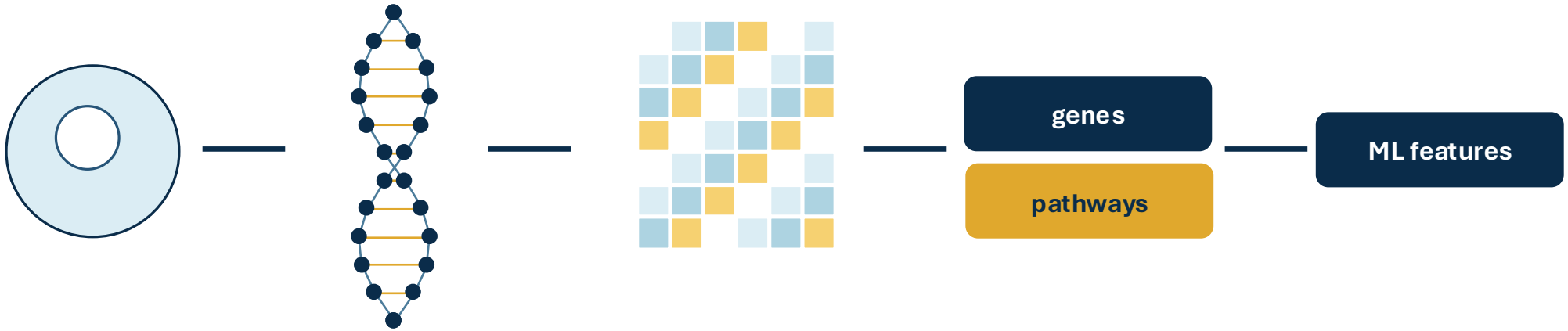
Genomics and Transcriptomics for Clinical Machine Learning

From TCGA-BRCA RNA-seq to interpretable biological features



Raquel Romão

Sofia Torres

26/05/2026
iSTARS Training school day 2

Today’s Schedule

09:00–09:30	Introduction to big clinical and omics datasets: how to extract knowledge from this data?
09:30–11:00	Omics data exploration: how genes drive clinical outcomes? Genomics
11:00–11:30	Break
11:30–13:00	Omics data exploration: how genes drive clinical outcomes? Transcriptomics
13:00–14:00	Lunch
13:30–15:40	From many genes to a few genes: Useful statistics and feature preparation for ML models
15:40–16:00	Break
16:00–17:00	Feature preparation practical + transition to Day 3

Who Is Guiding This Session?

Sofia Torres

iSTARS Computational Biologist

Currently I study the tumor micro environment using single cell RNA-seq

Email: sofia-torres@medicina.ulisboa.pt

Raquel Romão

iSTARS Computational Biologist

I use omics to understand how RNA-binding proteins shape immune system regulation

Email: raquel.romao@medicina.ulisboa.pt

Elias Barreira

Mostly interested in evolutionary biology and genotype–phenotype interactions. In this lab, I focused on predicting late-stage cancer phenotypes from early-stage data.

Email: eli@edu.ulisboa.pt

Examples of Questions We Work On

What in the tumor microenvironment drives metastasis growth?

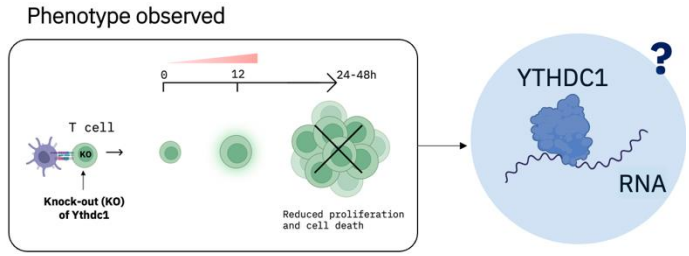
Profiling of the immune populations found in the liver through metastatic growth

Data used: **Single cell RNA Sequencing** (Transcriptomics)

We highlight some monocytes populations that have **cytotoxic potential**

How does YTHDC1 regulate proliferation or cell death?

RNA-binding protein



Data used: **RNA Sequencing** (Transcriptomics), **CLIP-seq** (Cross-linking immunoprecipitation followed by high-throughput sequencing)

We highlight **YTHDC1**, an RBP with a key role in **T cell proliferation**, involved in **apoptotic pathways**, and **cell cycle regulation** - essential for effective immune responses

What We Want You To Take Away

By the end you should feel more confident to...

Ask better data questions

Recognize what omics can and cannot answer

Know commonly used metrics

Discuss feature choices with data scientists

You will not be expected to...

Write code from scratch

Understand every statistical detail

Memorize RNA-seq pipeline tools

Build a full ML model today

09:00–09:30

Introduction to big clinical and omics datasets: how to extract knowledge from this data?

From clinical question to analyzable tables

Clinical Data Is The Starting Point

Before omics or ML, we need to define the clinical problem clearly.

Clinical data tells us:

- **Who is in the cohort?**
- **What disease state?**
- **What treatment?**
- **What outcome?**

Key idea

Data analysis is not just “run an algorithm”. It starts with turning a clinical question into variables that can be measured.

What Types Of Clinical Data Exist?

Healthcare data comes in many forms. Some are already tabular; others need extra processing.

Patient and diagnosis

Age, sex, diagnosis, tumour type, stage, subtype, comorbidities

Treatment data

Surgery, drugs, radiation, dose, timing, treatment response

Measurements over time

Lab tests, vital signs, symptoms, adverse events, follow-up visits

Clinical text

Pathology reports, radiology reports, clinical notes, discharge summaries

Images and signals

Radiology, pathology slides, ECG, wearables, digital health data

Outcomes

Survival, relapse, progression, complications, quality of life

Today we will mainly use structured clinical variables: ER status, stage, molecular subtype and survival information.

Healthcare Data Are Powerful, But Messy

Real clinical datasets rarely arrive as clean, complete spreadsheets.

Missing values

Not every patient has every measurement.

Different coding systems

The same concept may be recorded in different ways.

Repeated measurements

Labs and treatments change over time.

Confounding

Groups may differ in age, stage, treatment or sampling.

Censoring

For survival, some patients are still alive at last follow-up.

Data leakage risk

Future information must not enter a prediction model.

This is why study design and preprocessing matter as much as the algorithm.

From Clinical Question To Analyzable Table

Most methods need a clear table: rows are patients/samples, columns are variables.

1. Define the question

Example: Do ER-positive and ER-negative tumours show different biology?

2. Define the cohort

Which patients? Which tumour samples? Which timepoint?

3. Define variables

Features: age, stage, mutations, genes. Labels: ER status, outcome.

4. Analyze

Explore, test associations, build features, evaluate models.

Example analysis table

patient	age	stage	ER	gene/pathway
P1	59	II	+	high
P2	43	III	-	low
P3	70	I	+	medium

How Do We Extract Knowledge From Data?

Different approaches answer different questions. They are complementary, not competing.

Descriptive statistics

What do we see?
How many patients? What is the distribution of age, stage or ER status?

Statistical inference

Is a pattern likely to be real?
Are two groups different after accounting for uncertainty?

Machine learning

Can we predict for new patients?
Can a model learn useful patterns from clinical and omics features?

Bioinformatics / omics analysis

What biology might explain the pattern?
Which genes, pathways or mutations differ between groups?

Clinical interpretation

Does the result make sense clinically?
Could it change a hypothesis, biomarker, experiment or model?

Statistics And Machine Learning In Simple Terms

Both are useful, but they are usually used for different goals.

Statistics often asks

“Is there evidence for an association?”

Example today:
Which genes differ between ER-negative and ER-positive tumours?

Machine learning often asks

“Can we predict a label or outcome on new data?”

Example tomorrow:
Can clinical and molecular features help classify or predict patient outcomes?

For both approaches, the hardest part is often not the formula. It is defining the question, the cohort, the features, the label, and a fair evaluation.

Where Omics Enters The Story

Clinical data describes the patient and disease. Omics can add a molecular layer.

Clinical layer

What we observe in healthcare: stage, subtype, treatment, response, survival

Omics layer

What we measure molecularly: mutations, gene expression, pathways

Knowledge

What we want to learn: mechanisms, biomarkers, interpretable features

For example:

What data best describes biological outcomes such as the probability of cancer survival?

- Clinical variables - like smoking - alone are often not enough to describe phenotypes
  - Tumors might look similar clinically but have different biology
  - Omics can be a bridge between phenotype and mechanism

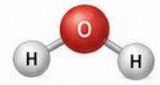
Biology works across many levels

- A pattern observed in a cohort may involve:

Genetic variation → gene expression → cellular behavior → tissue function → clinical phenotype

- To understand the phenomenon properly, we often need to connect **evidence across several biological levels.**

The arrangement of living things from simplest to most complex.

	LEVEL	DESCRIPTION	EXAMPLE
1	CHEMICAL (ATOMIC)	Basic building blocks of matter. Atoms combine to form molecules.	Hydrogen atom 
2	CHEMICAL (MOLECULAR)	Two or more atoms bonded together to form molecules.	Water (H_2O) 
3	ORGANELLE	Specialized structures within cells that perform specific functions.	Mitochondrion 
4	CELL	The basic structural and functional unit of life.	Animal cell 
5	TISSUE	A group of similar cells that work together to perform a specific function.	Muscle tissue 
6	ORGAN	Composed of two or more tissues working together to perform a specific job.	Heart 
7	ORGAN SYSTEM	A group of organs that work together to perform major functions.	Circulatory system 
8	ORGANISM	An individual living thing capable of carrying out all life processes.	Human 
9	POPULATION	A group of individuals of the same species living in the same area.	Population of deer 
10	COMMUNITY / ECOSYSTEM	All the populations of different species interacting in a particular area along with the nonliving environment.	Forest ecosystem 

Higher-level biological pressures can reshape lower-level biology

The relationship between levels is not purely linear:

- **Lower-level** changes, such as genetic variants, can influence higher-level outcomes.
- But **higher-level** processes can also feed back and affect lower-level states.

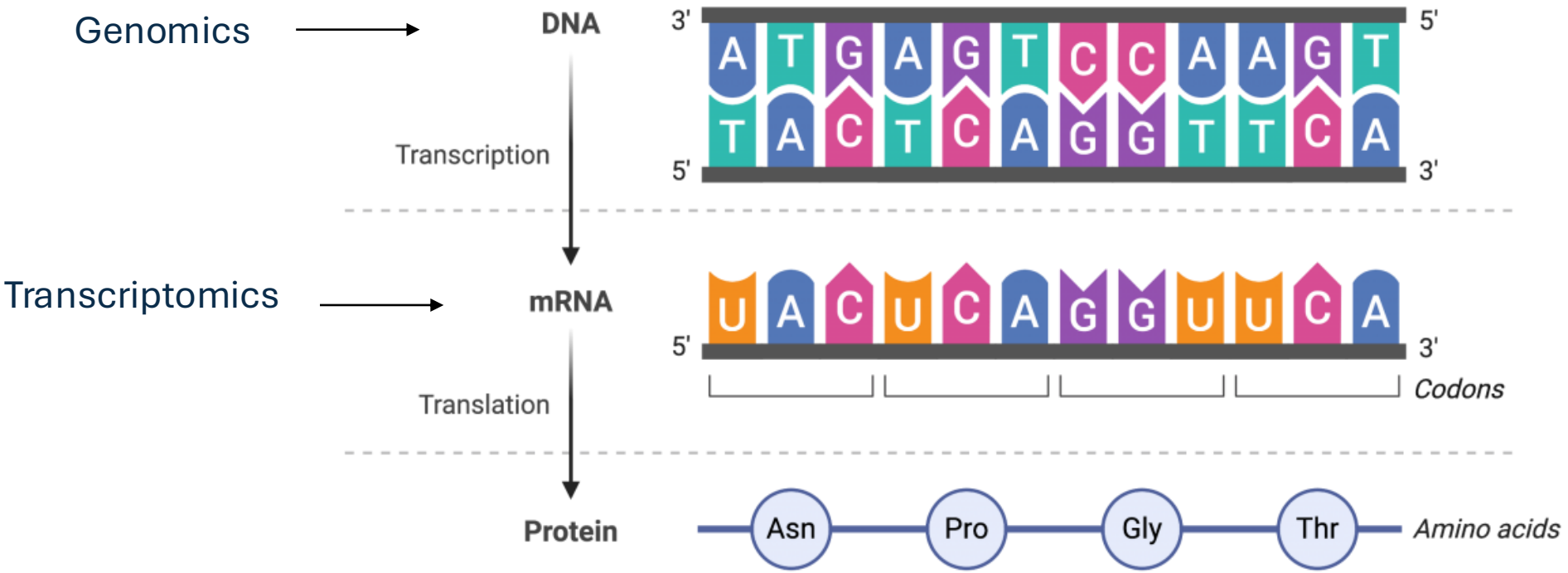
Not just a one-way chain

Higher-level biological pressures can reshape lower-level biology:

- **Inflammation** — activates stress and immune-response programs
- **Treatment** — selects resistant cells and changes survival pathways
- **Metabolism** — shifts gene expression through oxygen, nutrients, and energy stress
- **Tissue environment** — changes cell behavior through local signals and physical context
- **Immune pressure** — selects cells that evade detection or immune attack
- **DNA-binding proteins** — translate these signals into gene regulation by turning genes on/off

Introduction to omics datasets

Genomics vs transcriptomics



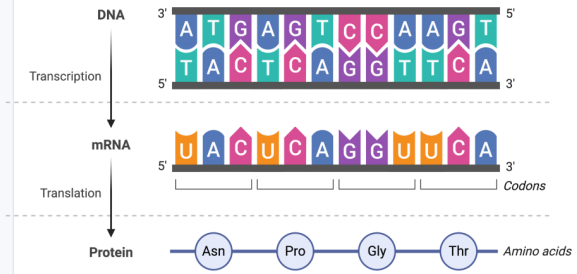
Genomics vs Transcriptomics

Genes: The blueprint

What it captures:
- Constant
- Same potential in all cells

Good for:
- Subtyping phenotypes (e.g. HER2-positive breast cancer)

- **Data**: Mutations



RNA: A snaphot in time

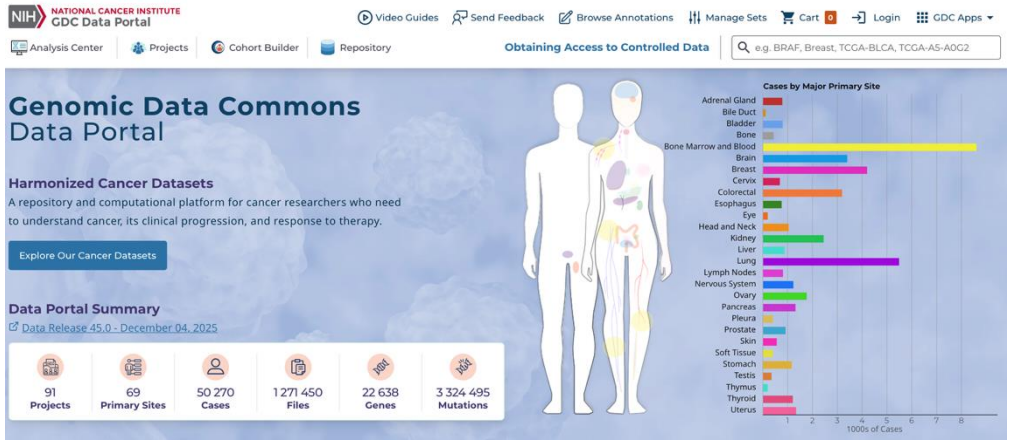
What it captures:
- Dynamic state of the organism
- Same cell has different expression patterns depending on state

Good for:
- Characterizing immune landscape, organism state

- **Data**: Gene expression (around 20k genes per sample) - high dimensional

The Cancer Genome Atlas Program (TCGA)

- 20,000 Primary cancer samples + Matched Normal Samples
- 33 types of cancer
- Includes genomic, epigenomic, transcriptomic, and proteomic data
- Open access for anyone wanting to do a cancer research project

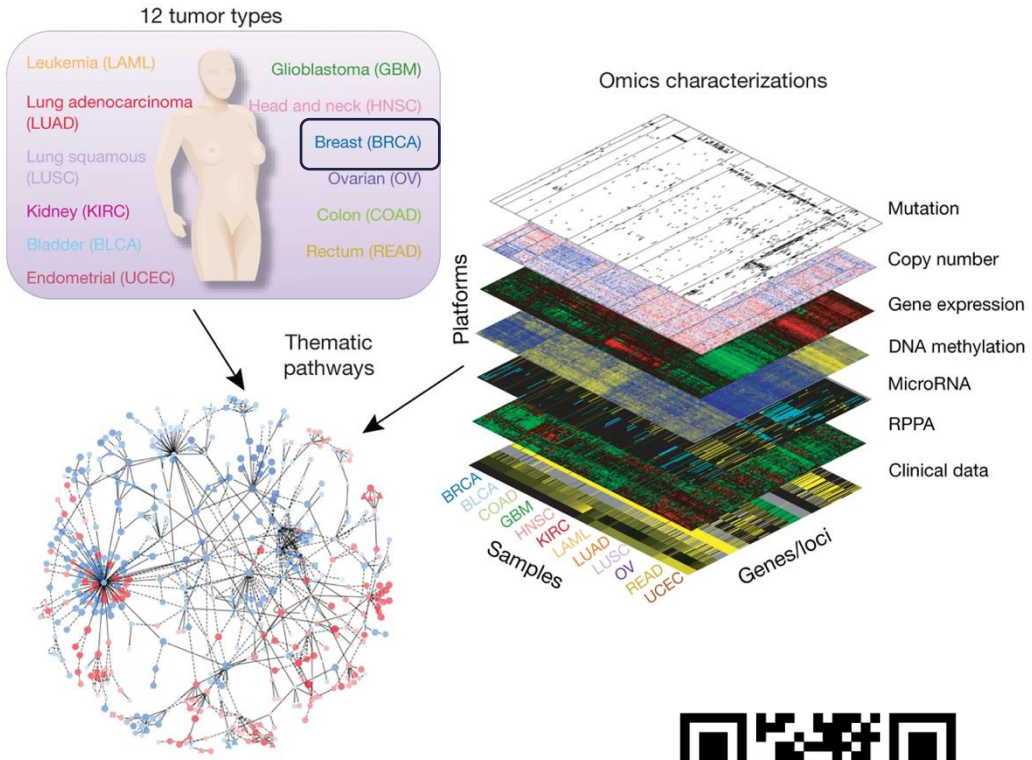


Access TCGA Data Here!

Our case study: The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA)

- 974 Breast cancer patients

Data available:
- Clinical data
- Mutational data
- Transcriptomic data
- Imaging data (day 4)



- We download this data with R, more specifically, the *curatedTCGAData* package

09:30–11:00

Omics data exploration: how genes drive clinical outcomes? Genomics

DNA alterations as sparse, interpretable features

Why start with DNA?

DNA is not always the complete explanation, but it is often the most accessible, interpretable, and cost-effective starting point.

Lower complexity

Higher complexity

DNA — Most foundational layer; changes can propagate upward.

↓

RNA / Gene expression — Translates genetic information into functional molecules.

↓

Proteins — Carry out cellular functions and signaling.

↓

Cells — Reflect altered behavior, state, and interactions.

↓

Tissues / Organs — Cells organized into structures with specialized functions.

↓

Phenotype / Clinical outcome — Observable traits, disease risk, treatment response, and survival.

= Base of the biological hierarchy

DNA assays are more accessible and interpretable

DNA sequencing (genomic tests)

- Relatively low cost
- Stable samples
- Broadly available
- Standardized workflows

Affordable, robust and widely available

RNA sequencing (transcriptomics)

- Higher cost
- Sensitive to tissue quality and timing
- Lower stability (RNA degrades)
- Greater biological variability

More expensive and context-dependent

Proteomics

- Higher cost
- Technically complex
- Incomplete proteome coverage
- Lower reproducibility

Costlier and technically challenging

Metabolomics

- Higher cost
- Highly dynamic
- Affected by diet, drugs, and time of day
- Harder to standardize

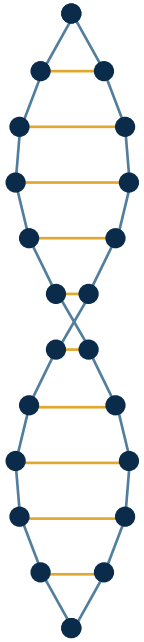
More variable and context-sensitive

Single-cell / Spatial omics

- Very high cost
- Complex workflows
- Requires specialized infrastructure
- Data analysis intensive

Powerful, but expensive and resource-intensive

What Does Genomics Measure?



DNA-level alterations

Single-nucleotide variants
Insertions/deletions

Copy-number gains/losses
Structural changes

Germline vs somatic context

Clinical meaning

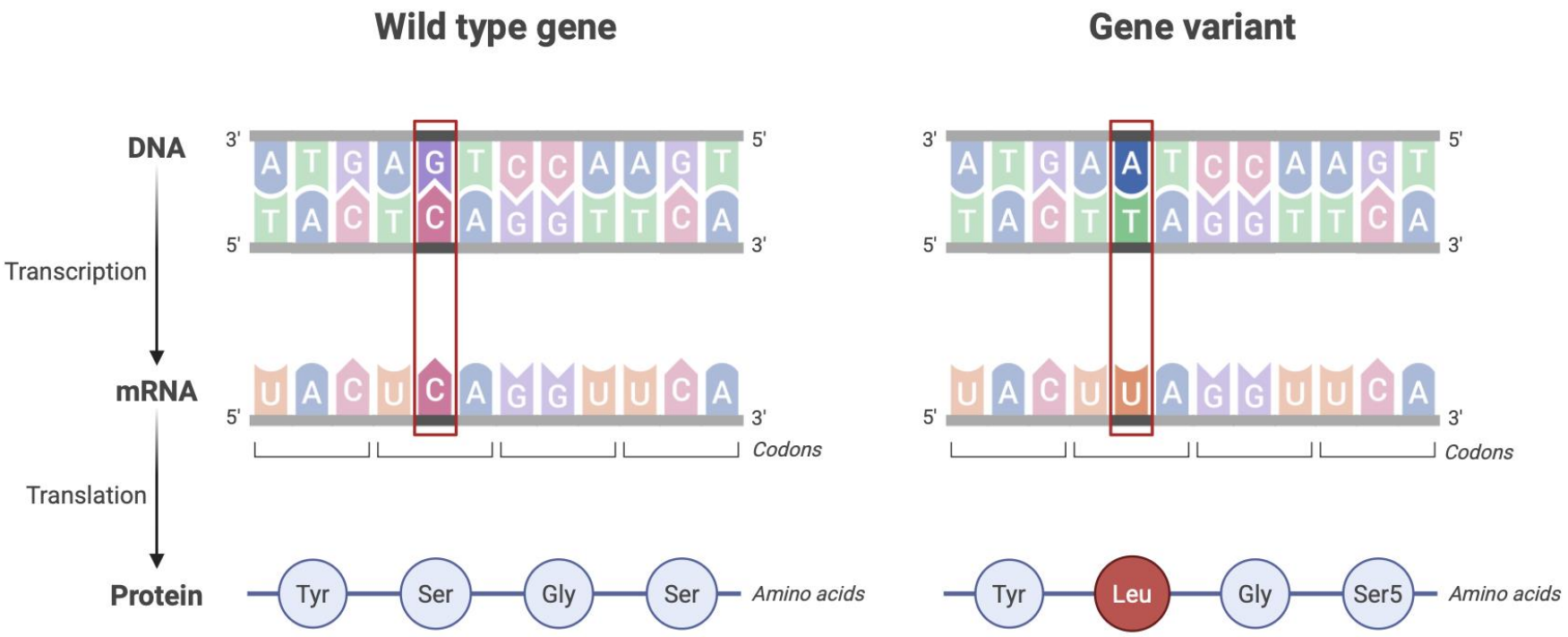
Can identify biomarkers

Can explain treatment sensitivity

Can define subtypes

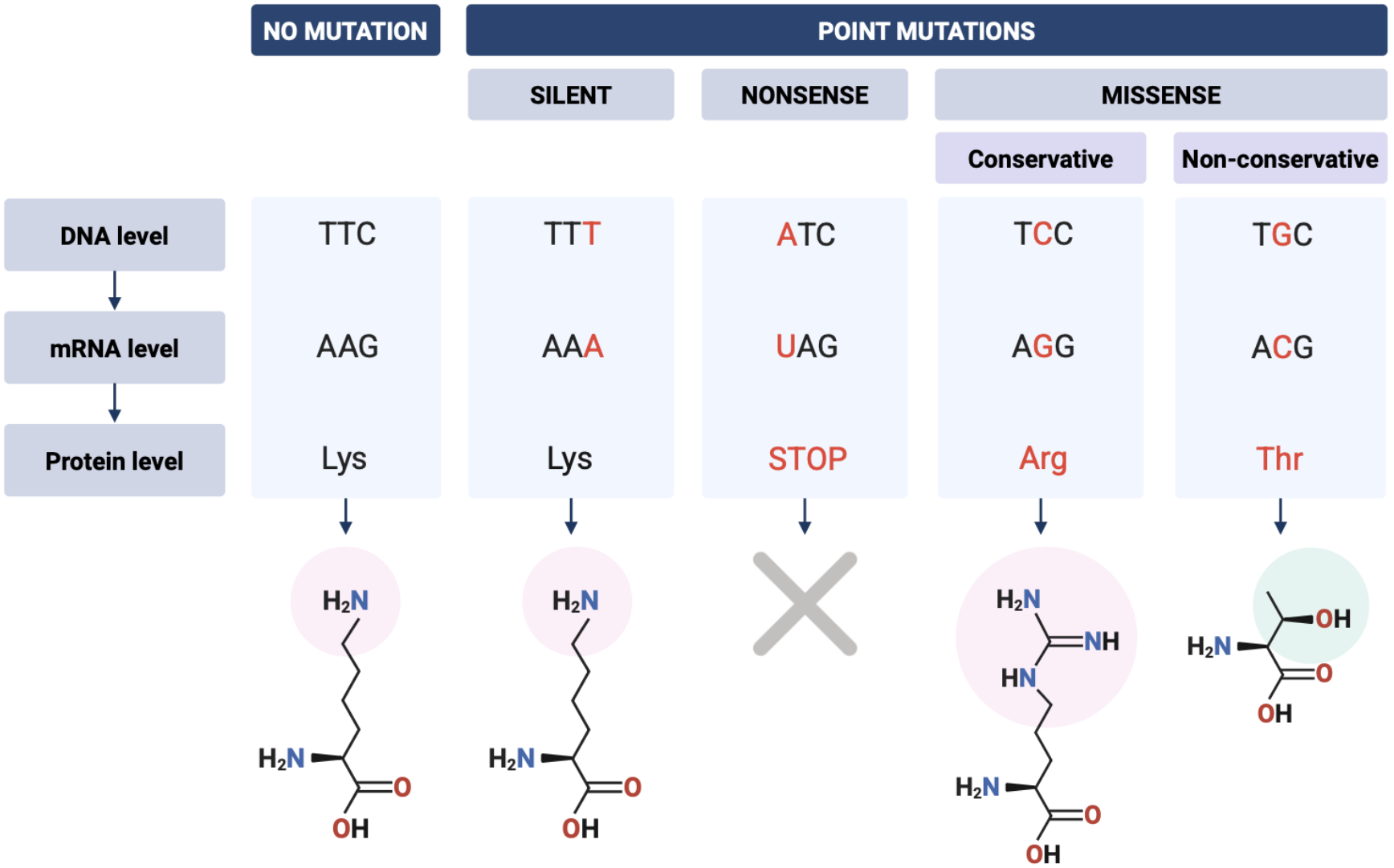
Can be absent in many patients

What is a variant?

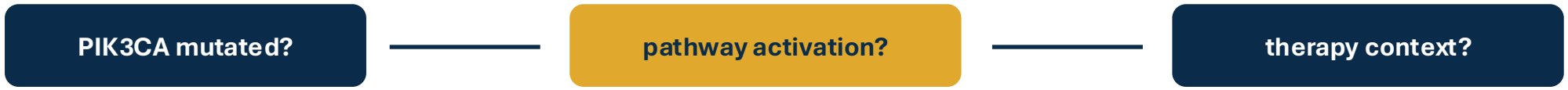


Variant types

Could be a way of prioritizing mutations



Example: Mutation As A Biomarker



Clinical reasoning

A mutation becomes a biomarker **only** when it has interpretable biological and clinical relevance in a specific context.

Mutation Data Are Sparse

Many zeros, few mutation events

ML implication

Because the omic space is so large, the feature matrix becomes sparse when each alteration is treated as a separate binary feature.

Sparse features can be interpretable
But rare events are hard to model
Small sample sizes limit conclusions
Grouping genes into pathways may help

Take-home

- Genomic features are often sparse: many patients do not carry a given alteration.
- It's common to have many features and few samples – typical "curse of dimensionality"

Genomics And ML: Why Genomics Can Still Be Useful

Advantages

Interpretable yes/no features

Close to mechanism

Some variants are clinically actionable

Feature count can be manageable

Risks

Rare events give few positive samples

Many mutations are just dragged by other alterations

“Not detected” may mean uncovered or below sensitivity

Cohort-specific signals may not generalize

Limitations: Not Every Mutation Is Actionable

Avoid overinterpretation

Mutation present ≠ driver

Driver ≠ druggable target

Association ≠ causality

Cohort bias matters

Ask better questions

Is the alteration common enough?

Is there biological rationale?

Is there external validation?

What is the clinical use case?

Mini Activity: Genomics Feature Choice

Feature	Interpretability	Risk
TP53 mutated	high	Broad, not always actionable
Total mutation count	medium	confounded by assay
Rare gene mutation	high	few samples
Pathway altered	medium-high	definition choices

= any detected functional-class alteration in genes assigned to that pathway, with missingness handled explicitly

Prompt

Which feature would you include in a clinical ML model, and what would you check before trusting it?

Mini Activity: Genomics Feature Choice (with discussion)

Feature	Interpretability	Risk	Discussion
TP53 mutated	high	Broad, not always actionable	Good simple feature but not clinically specific
Total mutation count	medium	confounded by assay	Useful summary, but strongly affected by assay type
Rare gene mutation	high	few samples	Biologically interesting but hard to model due to few positives
Pathway altered	medium-high	definition choices	Often a good compromise

- ML can use altered features without knowing the mechanism; interpretation requires checking recurrence, missingness, assay sensitivity, and validation
- Let the model find useful alteration patterns, but validate that the signal is not just cohort structure, assay sensitivity, or missing data

Hands on time!

Please open the genomics session notebook (should be in your email)

11:00–11:30

Break

Next: transcriptomics

11:30–13:00

Omics data exploration: how genes drive clinical outcomes? Transcriptomics

RNA activity, expression matrices and biological programs

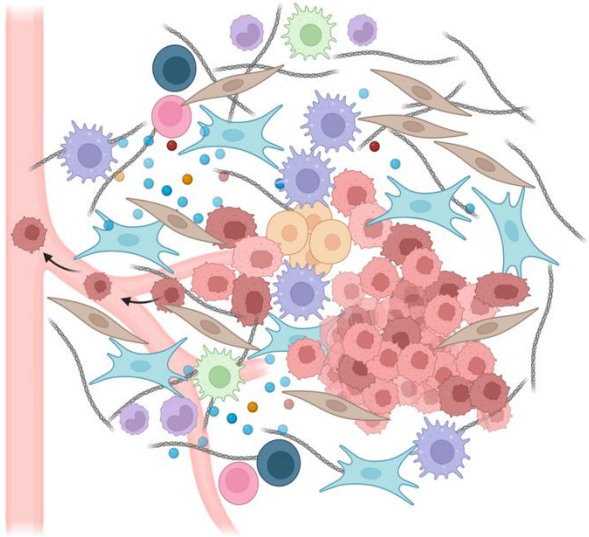
Before moving on

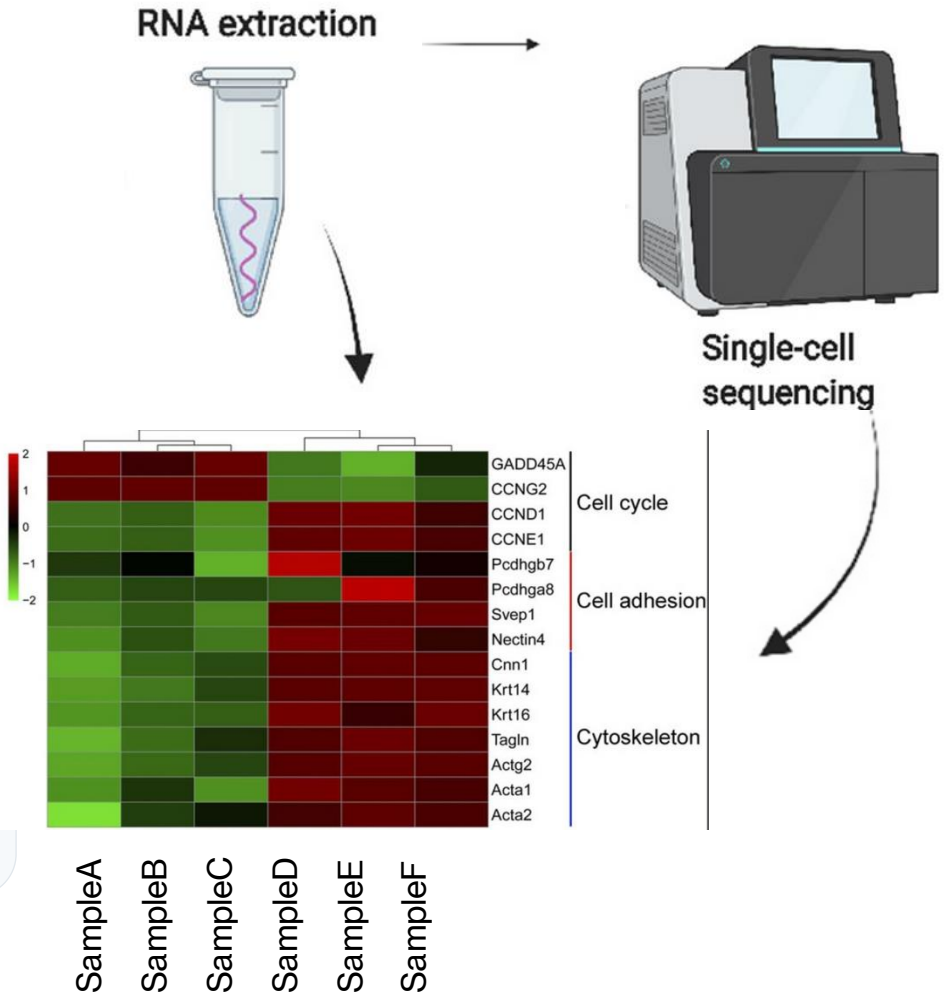
Please open the transcriptomics session notebook and click on “Execute all" – we’ll get back to it later

What Problem Does RNA-seq Solve?

Which genes are active?

Which biological pathways are dysregulated in cancer?



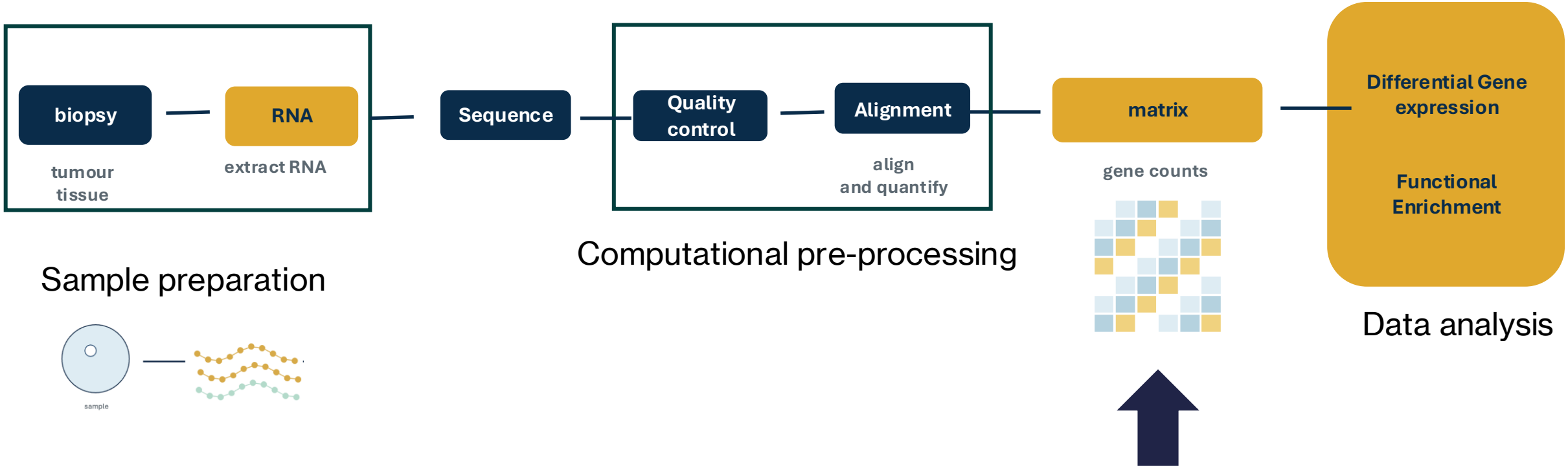


Plain language

RNA-seq measures the organism active biological programs at the time the sample was collected.

Note: In Bulk-RNA gene expression is sample wide, information about single cell is lost

RNA-seq In One Slide: Tissue To Expression Matrix

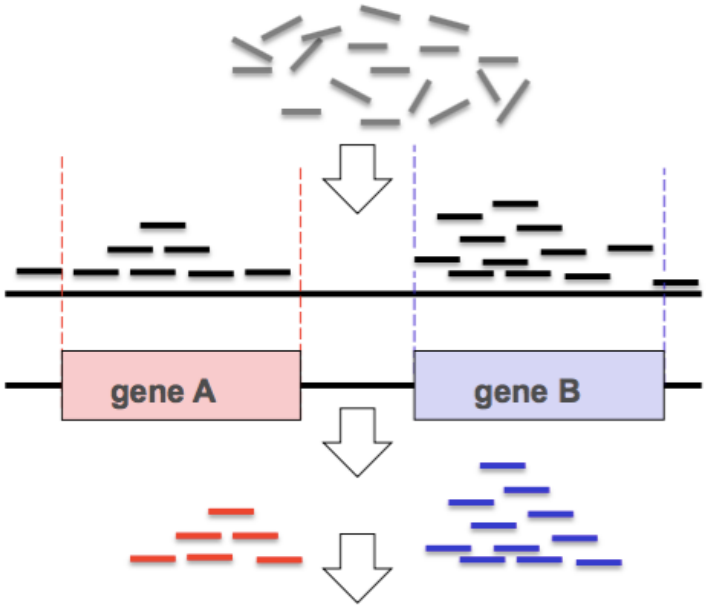


For today

We start after the computational pipeline: the expression matrix is already prepared. But understanding where it comes from helps interpret quality and limitations.

RNA-seq: Tissue To Expression Matrix

RNA is fragmented and amplified during sample preparation



	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3
Gene A	6	5	7	17	10	11
Gene B	11	11	10	3	4	2
Gene C	200	150	355	50	1	3
Gene D	0	1	0	2	0	1

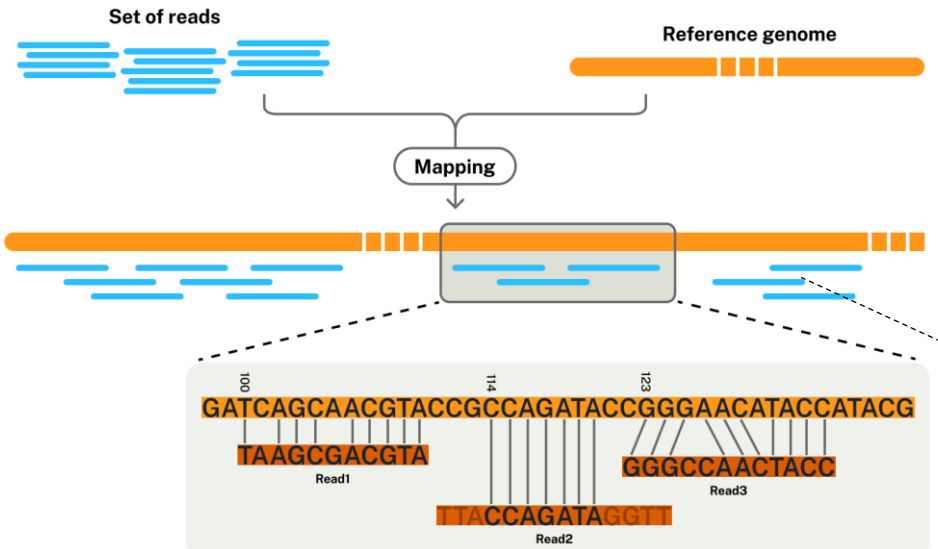
How do we know from which transcripts/genes the reads came from?

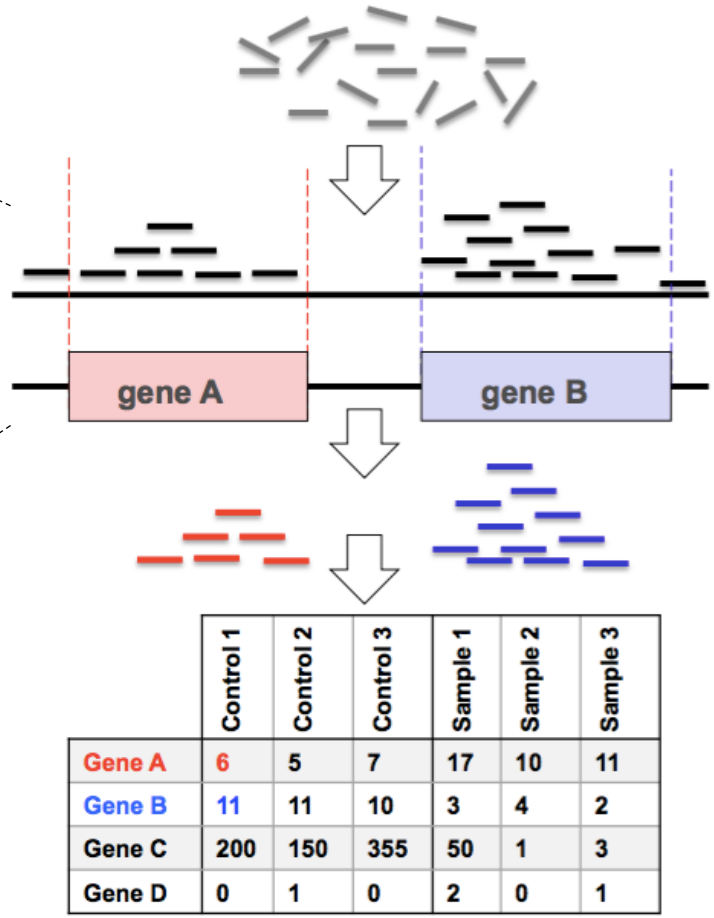
RNA-seq: Tissue To Expression Matrix

Alignment: From where did the read originate from?

- Alignment is like a **puzzle** where **we have individual pieces (reads)** and **know the final image (Reference genome)**

- Alignment tools solve this puzzle



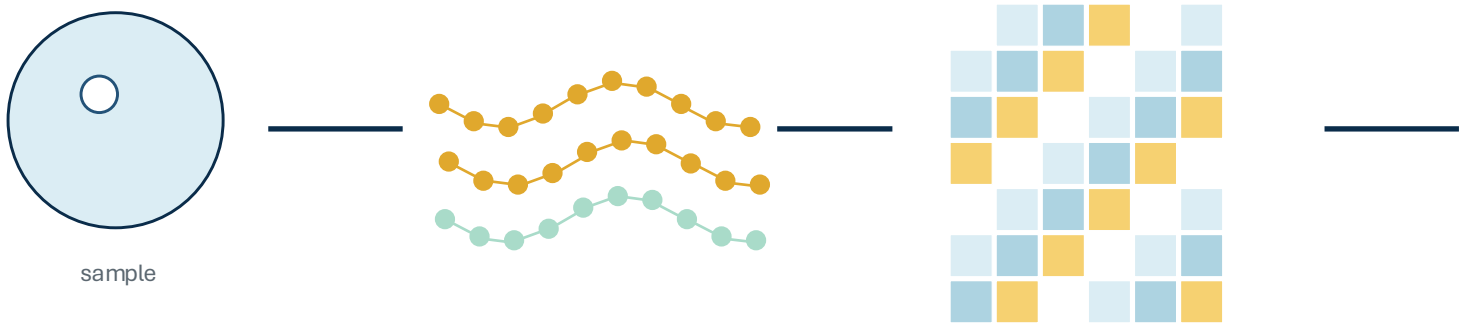


	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3
Gene A	6	5	7	17	10	11
Gene B	11	11	10	3	4	2
Gene C	200	150	355	50	1	3
Gene D	0	1	0	2	0	1

Hands on time!

Please open the transcriptomics session notebook

1 What does the expression matrix represent?



Meaning

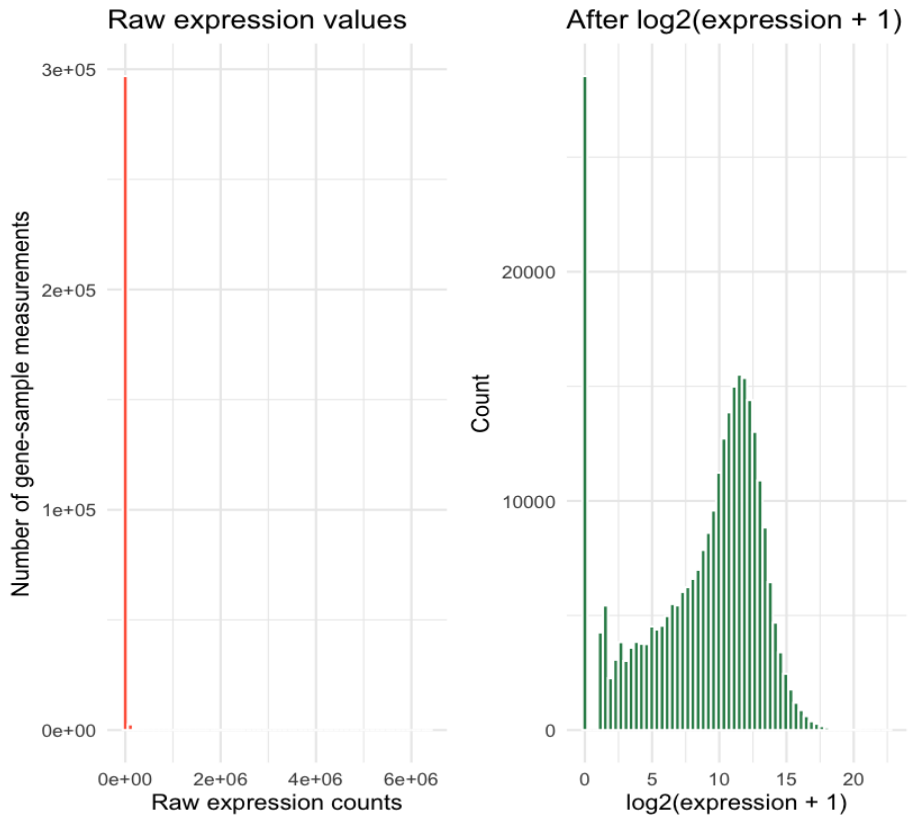
Rows = genes
Columns = samples
Values = expression estimates

Notebook checkpoint

Load the RDS object and inspect genes x samples.

	TCGA-A1-A0SB-01A-11R-A144-07	TCGA-A1-A0SD-01A-11R-A115-07	TCGA-A1-A0SE-01A-11R-A084-07	TCGA-A1-A0SF-01A-11R-A144-07	TCGA-A1-A0SG-01A-11R-A144-07	TCGA-A1-A0SH-01A-11R-A084-07	TCGA-A1-A0SI-01A-11R-A144-07	TCGA-A1-A0SJ-01A-11R-A084-07	TCGA-A1-A0SK-01A-12R-A084-07	TCGA-A1-A0SM-01A-11R-A084-07	...
A1BG	164	546	1341	836	512	1126	449	1433	626	2154	...
A1CF	0	0	0	1	3	1	2	0	1	4	...
A2BP1	22	1	2	0	25	4	5	1	1	4	...
A2LD1	127	331	498	526	451	838	203	831	2948	733	...
A2ML1	94	144	114	77	76	127	121	119	367	96	...
A2M	102123	107181	101192	50316	45826	81857	39132	128591	27972	60189	...
A4GALT	890	1409	1711	1142	1751	1997	846	3802	514	2031	...
A4GNT	6	5	2	2	2	7	6	2	4	8	...
AAA1	0	0	1	0	4	0	0	0	0	0	...
AAAS	2139	2219	4294	3052	2450	3178	2247	2999	6992	2898	...
AACSL	2930	3	4	1	0	3	1	18	5	3	...
AACS	6533	3102	5271	1248	5695	4732	3295	4900	16066	8966	...
AADACL2	0	0	1	0	0	1	2	0	52	0	...
AADACL3	0	0	0	0	0	1	1	0	0	0	...
AADACL4	1	0	0	0	0	2	0	0	2	0	...
AADAC	42	4	13	4	16	17	12	29	31	26	...
AADAT	694	226	200	67	130	544	159	355	2952	253	...

Counts vs Normalized Expression: Why It Matters



Counts: how gene distribution looks like

Integer-like values

Negative bimodal distribution:

- **Many genes** have **low counts**
- **Few genes** have **high counts**

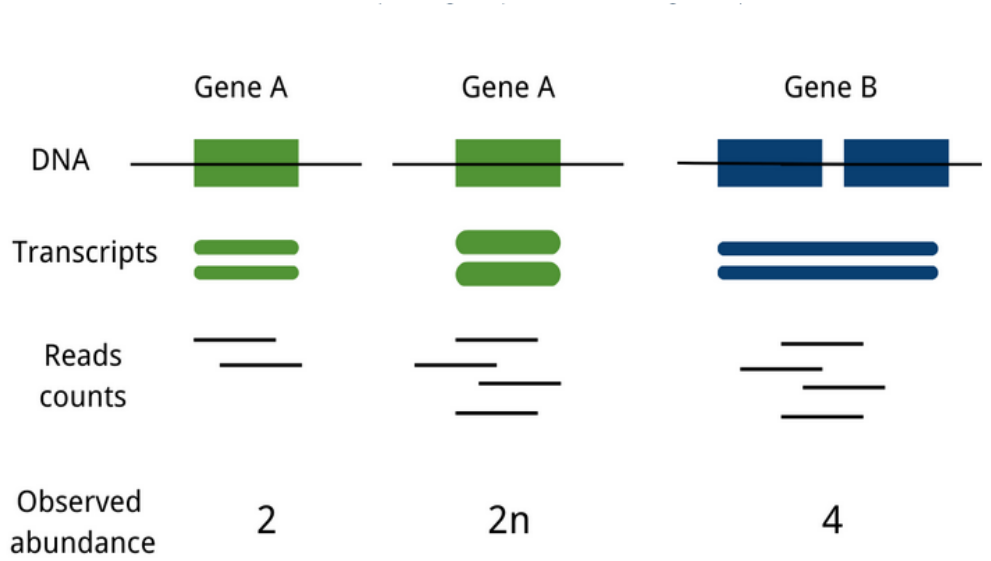
Are these genes comparable?

Need **normalization** or modeling for fair comparison

Normalization strategies

If one million fragments were sequenced, how many come from geneA?

Strategy: **TPM** (Transcripts per Million)



Corrects for: **Gene length** bias

Longer genes will have more reads regardless of expression

Genes are relative to their sample!

Great for **exploratory analysis** but not for group differences analysis

Exploratory analysis: How can we look into this data?

After normalization… Such as TPM

~700 sample columns

~20,000 gene rows

	TCGA-A1-A0SB-01A-11R-A144-07	TCGA-A1-A0SD-01A-11R-A115-07	TCGA-A1-A0SE-01A-11R-A084-07	TCGA-A1-A0SF-01A-11R-A144-07	TCGA-A1-A0SG-01A-11R-A144-07	TCGA-A1-A0SH-01A-11R-A084-07	TCGA-A1-A0SI-01A-11R-A144-07	TCGA-A1-A0SJ-01A-11R-A084-07	TCGA-A1-A0SK-01A-12R-A084-07	TCGA-A1-A0SM-01A-11R-A084-07	...
A1BG	164	546	1341	836	512	1126	449	1433	626	2154	...
A1CF	0	0	0	1	3	1	2	0	1	4	...
A2BP1	22	1	2	0	25	4	5	1	1	4	...
A2LD1	127	331	498	526	451	838	203	831	2948	733	...
A2ML1	94	144	114	77	76	127	121	119	367	96	...
A2M	102123	107181	101192	50316	45826	81857	39132	128591	27972	60189	...
A4GALT	890	1409	1711	1142	1751	1997	846	3802	514	2031	...
A4GNT	6	5	2	2	2	7	6	2	4	8	...
AAA1	0	0	1	0	4	0	0	0	0	0	...
AAAS	2139	2219	4294	3052	2450	3178	2247	2999	6992	2898	...
AACSL	2930	3	4	1	0	3	1	18	5	3	...
AACS	6533	3102	5271	1248	5695	4732	3295	4900	16066	8966	...
AADACL2	0	0	1	0	0	1	2	0	52	0	...
AADACL3	0	0	0	0	0	1	1	0	0	0	...
AADACL4	1	0	0	0	0	2	0	0	2	0	...
AADAC	42	4	13	4	16	17	12	29	31	26	...
AADAT	694	226	200	67	130	544	159	355	2952	253	...

Potential problems

- Too many variables for few patients
- Many genes measure related biology
- Noise can look like signal
- Interpretation becomes hard

→ Classic high dimensionality problem!

We want to use dimensionality reduction techniques

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

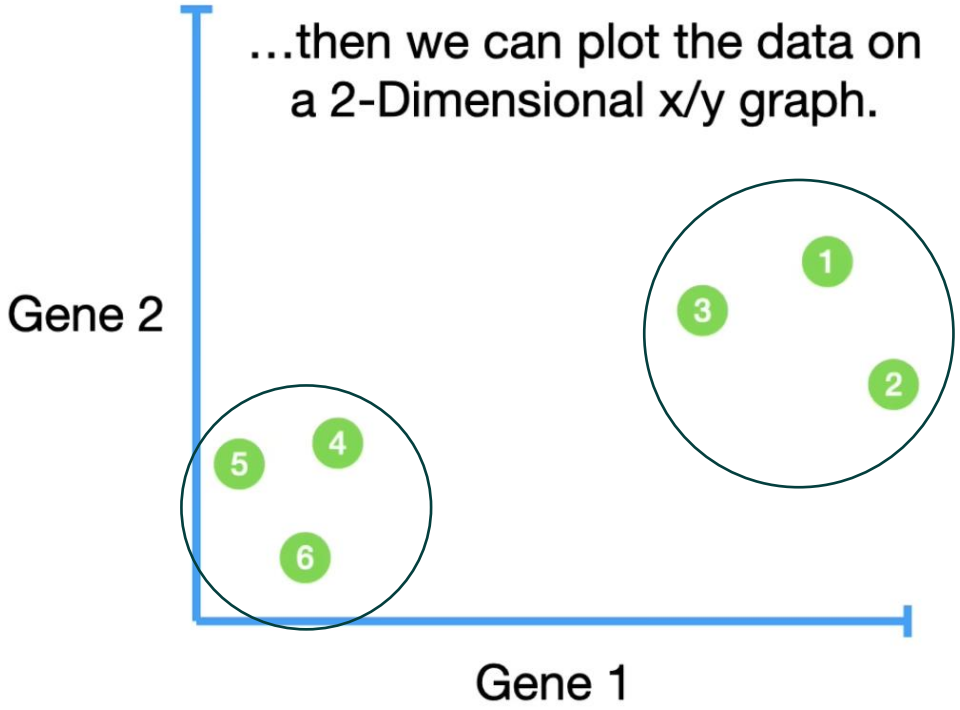
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	1	2

If we only measure 1 gene, we can plot the data on a number line…

Principal Component Analysis (PCA)

If we measured two genes…

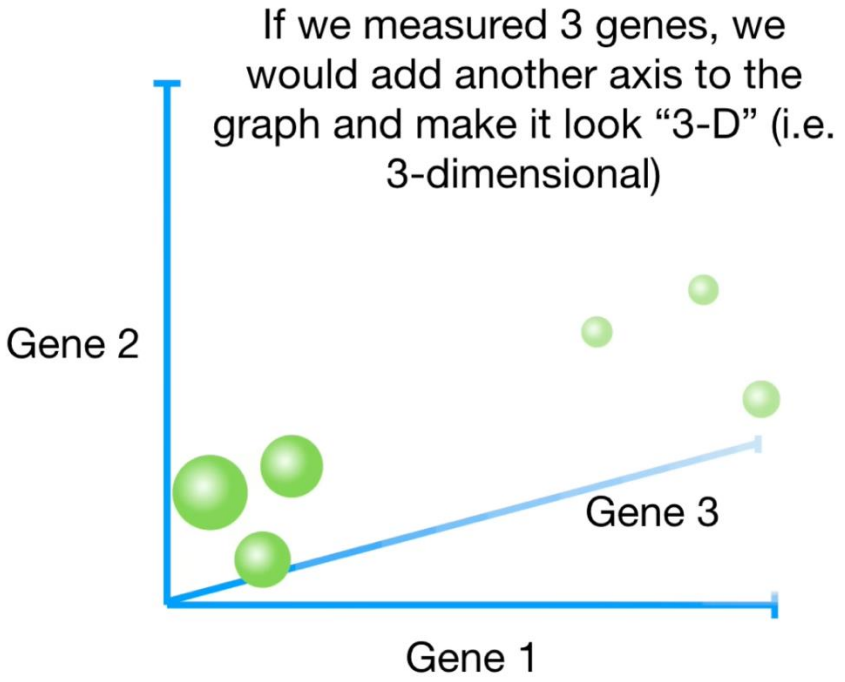
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	1	2
Gene 2	6	4	5	3	2.8	1



Principal Component Analysis (PCA)

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2

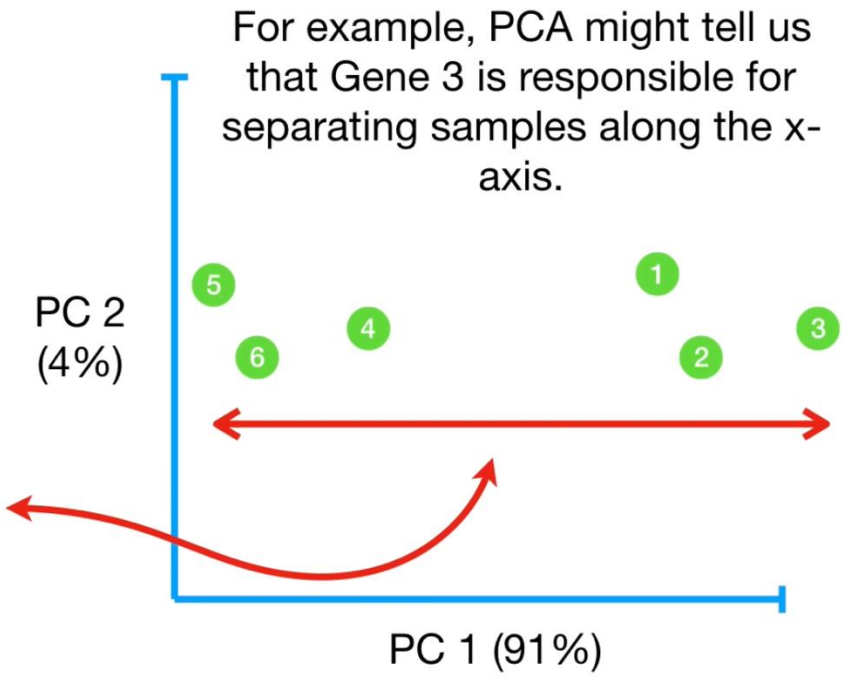
...



We can no longer plot the data from this dimension ☹

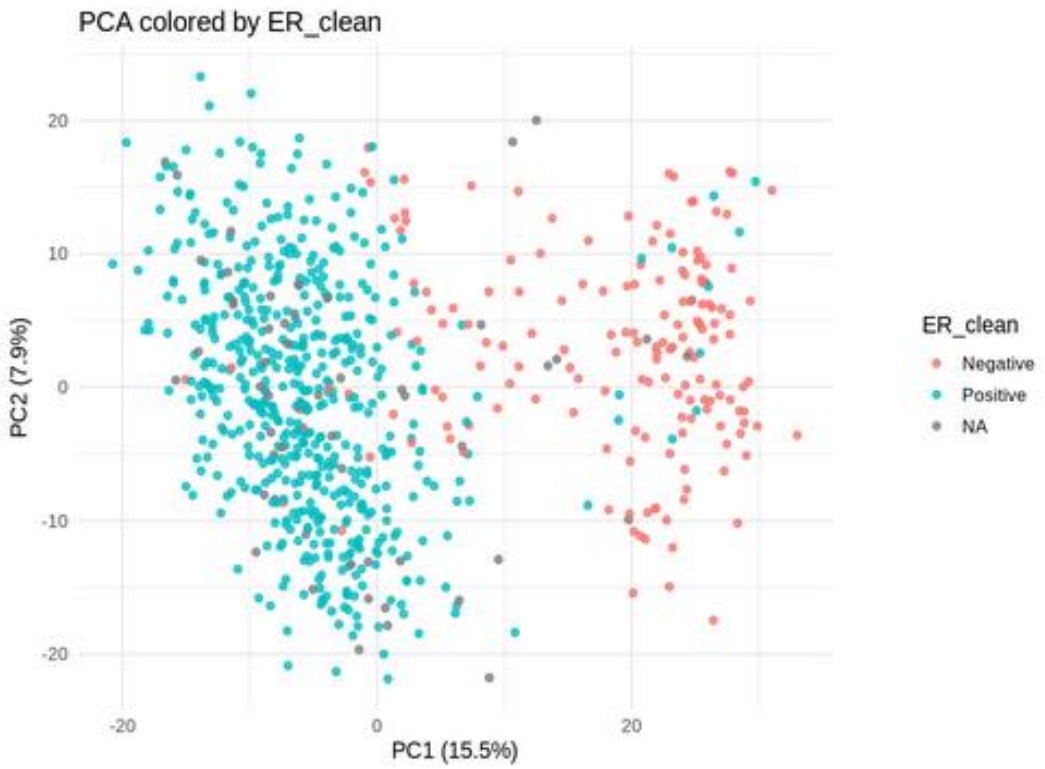
Principal Component Analysis (PCA)

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7



PCA: A Map Of Molecular Similarity

A first look before modeling



How to interpret

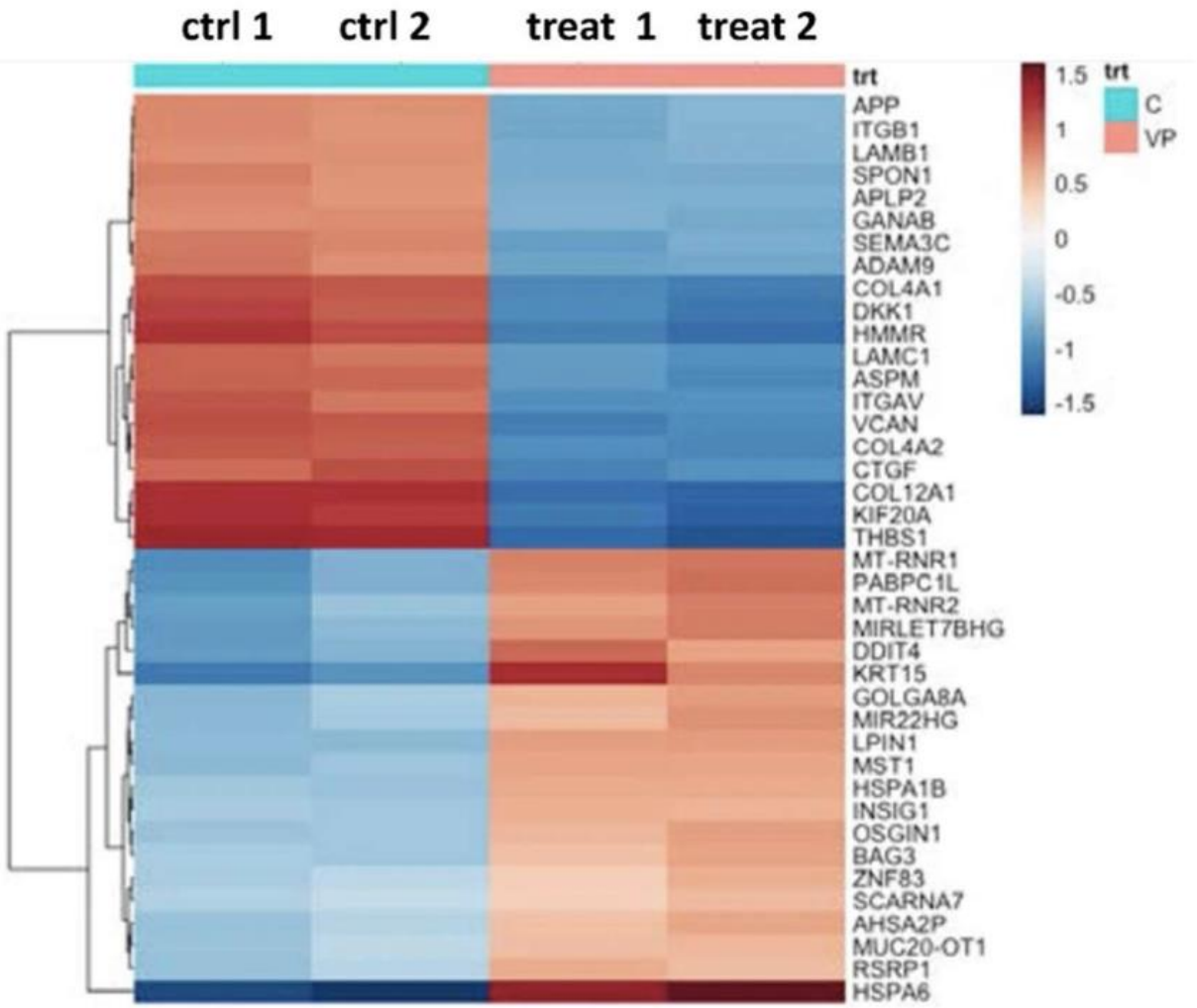
PCA does not use clinical labels

We color samples afterwards

Separation suggests molecular similarity follows that label

No separation does not mean the clinical variable is unimportant

How to find gene expression differences across groups?



How to compare genes across samples?

What genes differ between groups, i.e: metastasis formation vs none?

Differential expression analysis (DEA)!

3

Differential expression analysis (DEA)

Differential expression

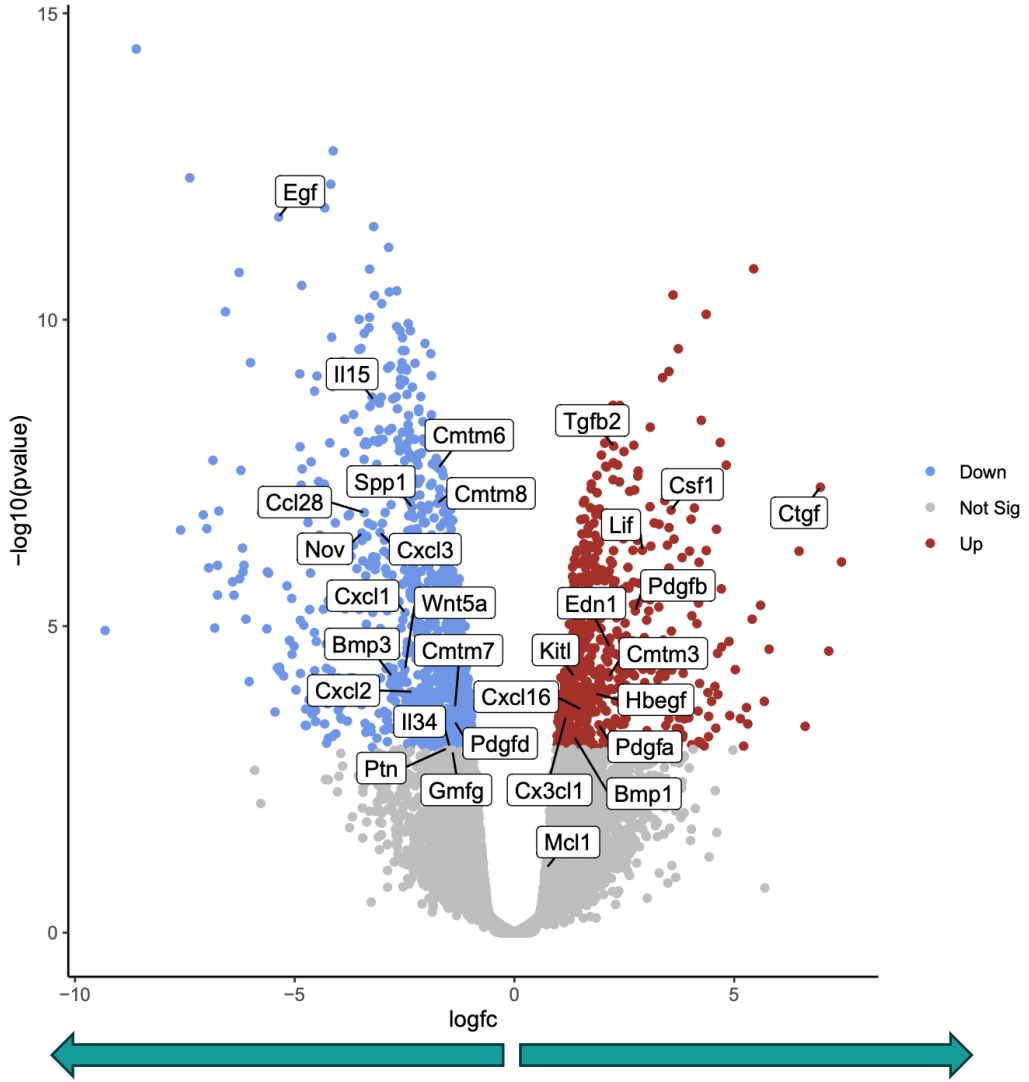
- Question: **which genes differ between groups?**
- Output: **fold-change** and **adjusted p-value**
- Goal: **biological interpretation**
- **Not automatically a predictive model**

Machine learning

- Question: **can we predict in new patients?**
- Output: **AUC, sensitivity, specificity, calibration**
- Goal: **generalizable prediction**
- **Needs validation and leakage control**

DEA Output: Volcano Plot - Size Of Difference + Confidence

Statistical significance



Magnitude of change between groups

Reading the plot

Far left/right = larger group difference

Higher = stronger statistical support

Pick a gene and ask: does it make biological sense?

3

DESEq2: p-value correction

Problem:
Pvalue = 0.05 (5% change a gene is a false positive)
With 20k genes, that is 1k FP!

Benjamini-Hochberg correction:
Raises the bar – the more genes compared, the stronger the correction

From Gene List To Biological Hypothesis

Gene	Direction	Evidence	Hypothesis
ESR1	Higher in group A	small padj	hormone signaling
FOXA1	Higher in group A	small padj	luminal identity
MKI67	Higher in group A	small padj	proliferation

Proposed task

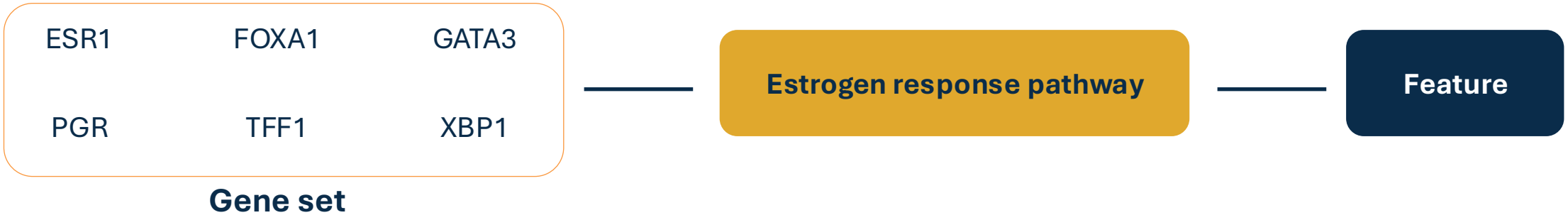
Choose one top gene, open **GeneCards**, and write one plain-language **hypothesis**. Example: “This gene may be higher because ER-positive tumours activate hormone-response programs.”

Pathways: from individual genes to biological programs

How to go from Gene's LOGFC value to a biological hypothesis ?

Gene set enrichment analysis (GSEA)!

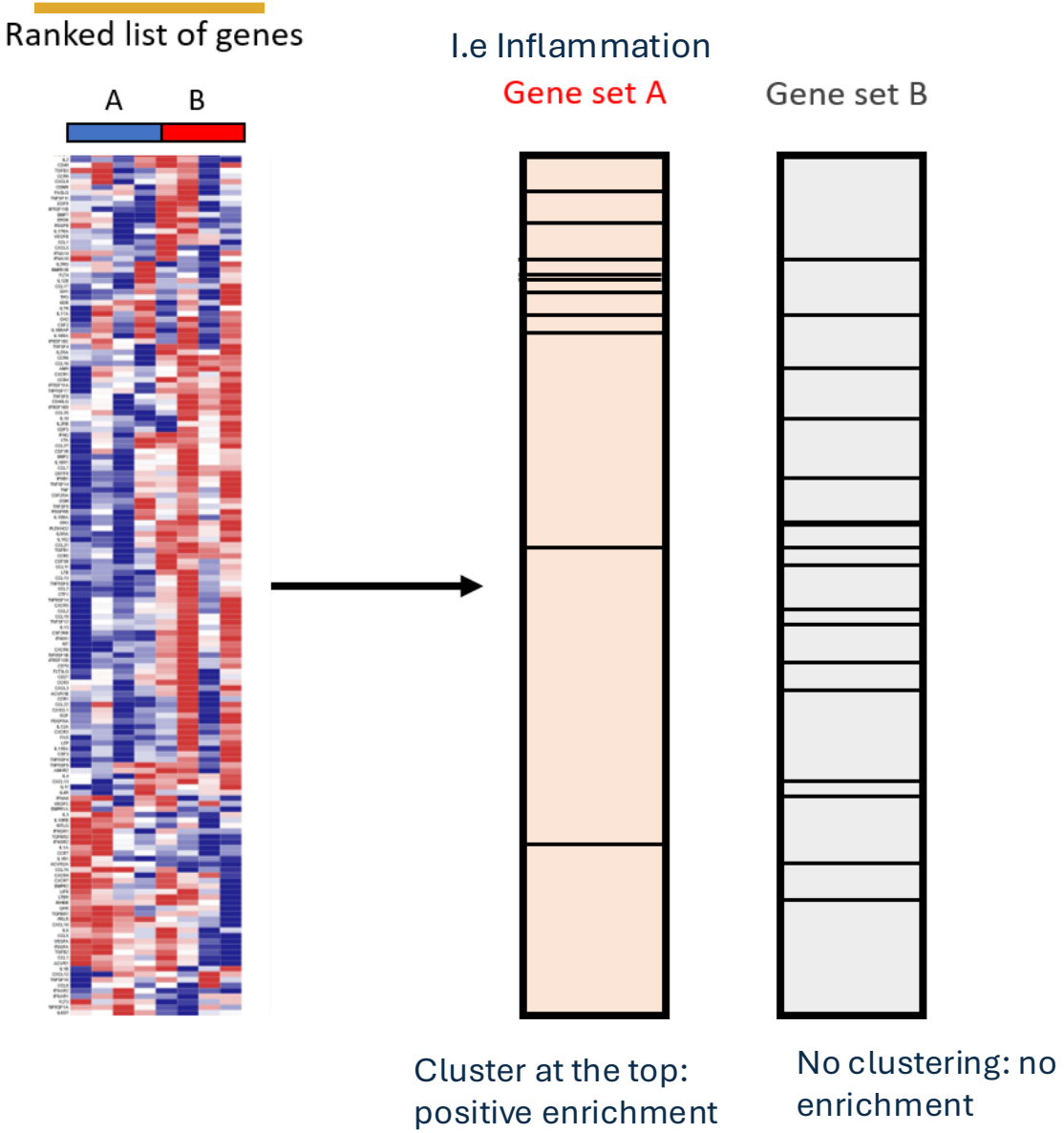
4 Pathways: from individual genes to biological programs



Tip: Why this helps ML

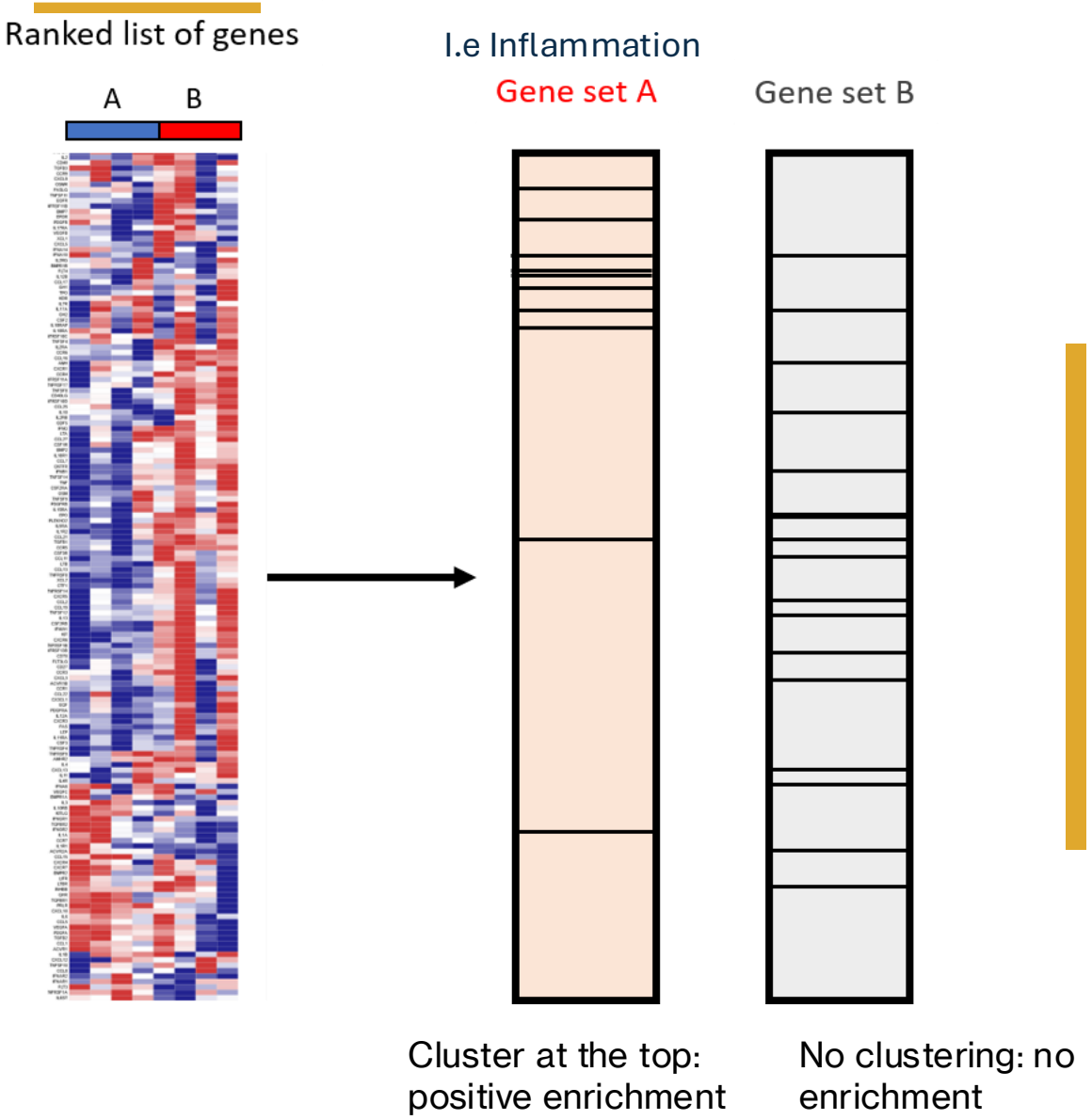
Pathway scores **summarize** coordinated biology and **reduce dimensionality**

GSEA: Biological pathways differ between groups?



- **Genes are ranked by LOGFC**
- **Pathways** are **list of genes (gene sets)** , i.e, 'Inflammation pathways' is composed of genes involved in this process
- Question: **When looking at the list of genes, where in the rank do they appear? This is enrichment**

GSEA: Biological pathways differ between groups?



Interpretation

- **If pathway genes cluster at the top**: pathway is **enriched** in one group
- If they cluster at the bottom: enriched in the other group
- **The result is easier to interpret clinically than a long gene list**

Transcriptomics Take-Home

What we learned

- **Expression matrices** are high-dimensional snapshots of (tumour) biology.
- **PCA** helps explore whether clinical labels align with molecular patterns.
- **Differential expression** highlights group-associated genes.
- **Pathways summarize** coordinated biology and is useful for ML features.

13:00–14:00

Lunch

After lunch: from many genes to useful ML features

14:00–17:00

From many genes to a few genes: Useful statistics and feature preparation for ML models

Feature reduction, leakage control and ML-ready tables

Why Raw Transcriptomics Is Hard For ML

20,000 genes

~700 patients

High overfitting risk

Problems

- Correlated variables
- Noisy/redundant features
- Hard interpretation
- Few patients per subgroup

Goal

- Reduce features without losing useful biology
- Keep interpretations clinically meaningful
- Validate honestly

Four Ways To Go From Many Genes To Fewer Features

Variance filtering

Keep genes that vary across patients

Biology-driven

Summarize known pathways/gene sets (GSEA)

Unsupervised

PCA or latent representations

Supervised

Select genes using labels carefully

15:40–16:00

Break

Final hour: feature table and transition to Day 3

Day 2 Take-Home Messages

What to remember

- **Omics** can reveal **biological mechanisms** behind clinical phenotypes.
- **Genomics is often sparse and interpretable; transcriptomics is dense and high-dimensional.**
- **Exploration** comes before modeling.
- Differential expression and GSEA support biological interpretation, not direct prediction alone.
- **Pathway scores can turn thousands of genes into interpretable ML features.**
- Good study design means careful labels, train/test separation and leakage prevention.

Bridge To Day 3: Machine Learning

Today we turned genomics and transcriptomics into interpretable features. Tomorrow we ask whether those features predict outcomes in new patients.

Final question

Would you build a model using clinical variables, genes, pathways, or a combination? Ask Laura tomorrow ;)

End of module 1

Thank you!

If you have any remaining question(s), don’t hesitate to contact us

Sofia Torres

Email: sofia-torres@medicina.ulisboa.pt

Raquel Romão

Email: raquel.romao@medicina.ulisboa.pt

Elias Barreira

Email: eli@edu.ulisboa.pt