Mahidol University
Wisdom of the Land

Transcriptomics: Gene Expression Analysis

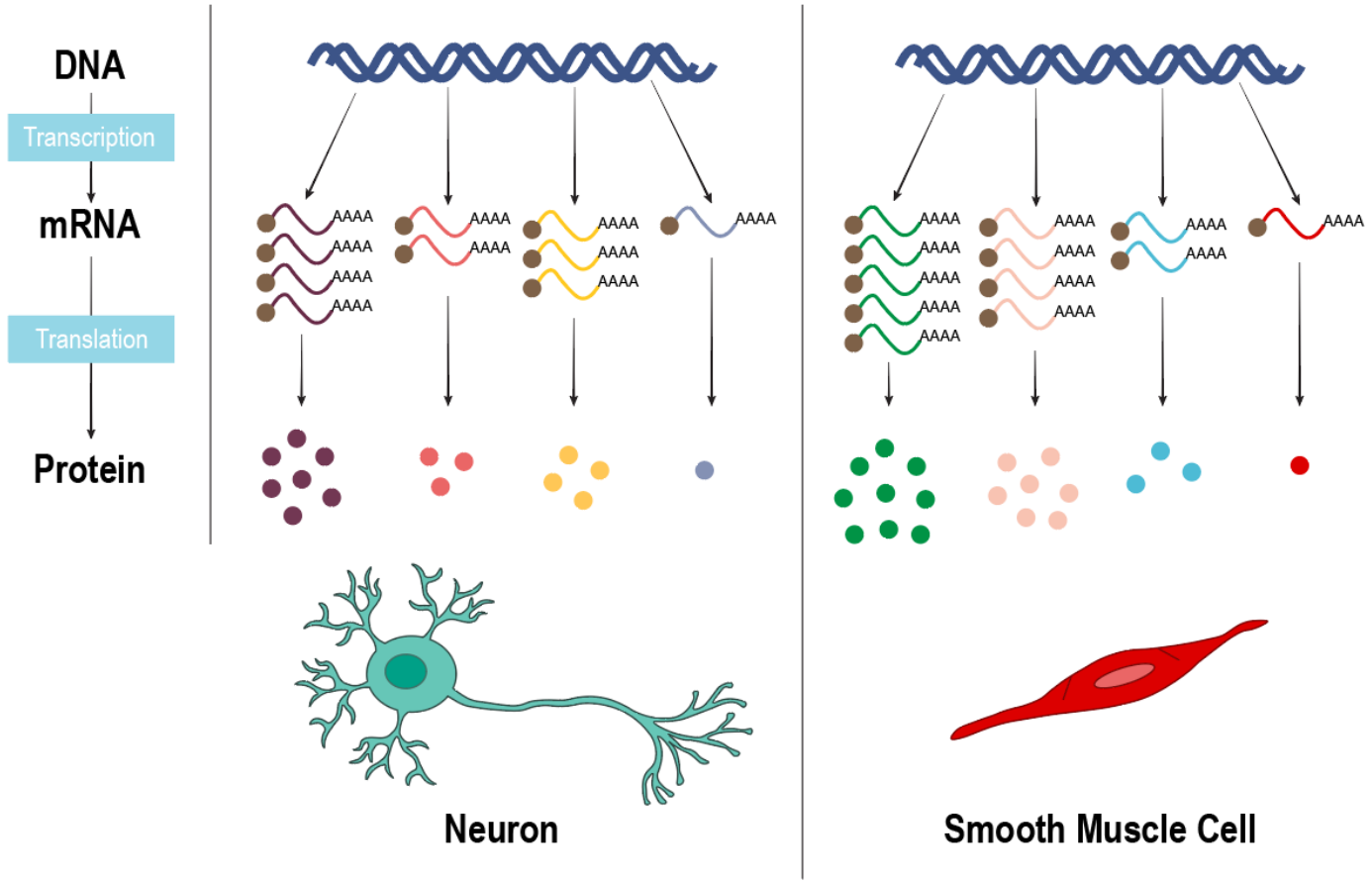
SCBM 373 – Bioinformatics

Waradon Sungnak, PhD

13/02/2026



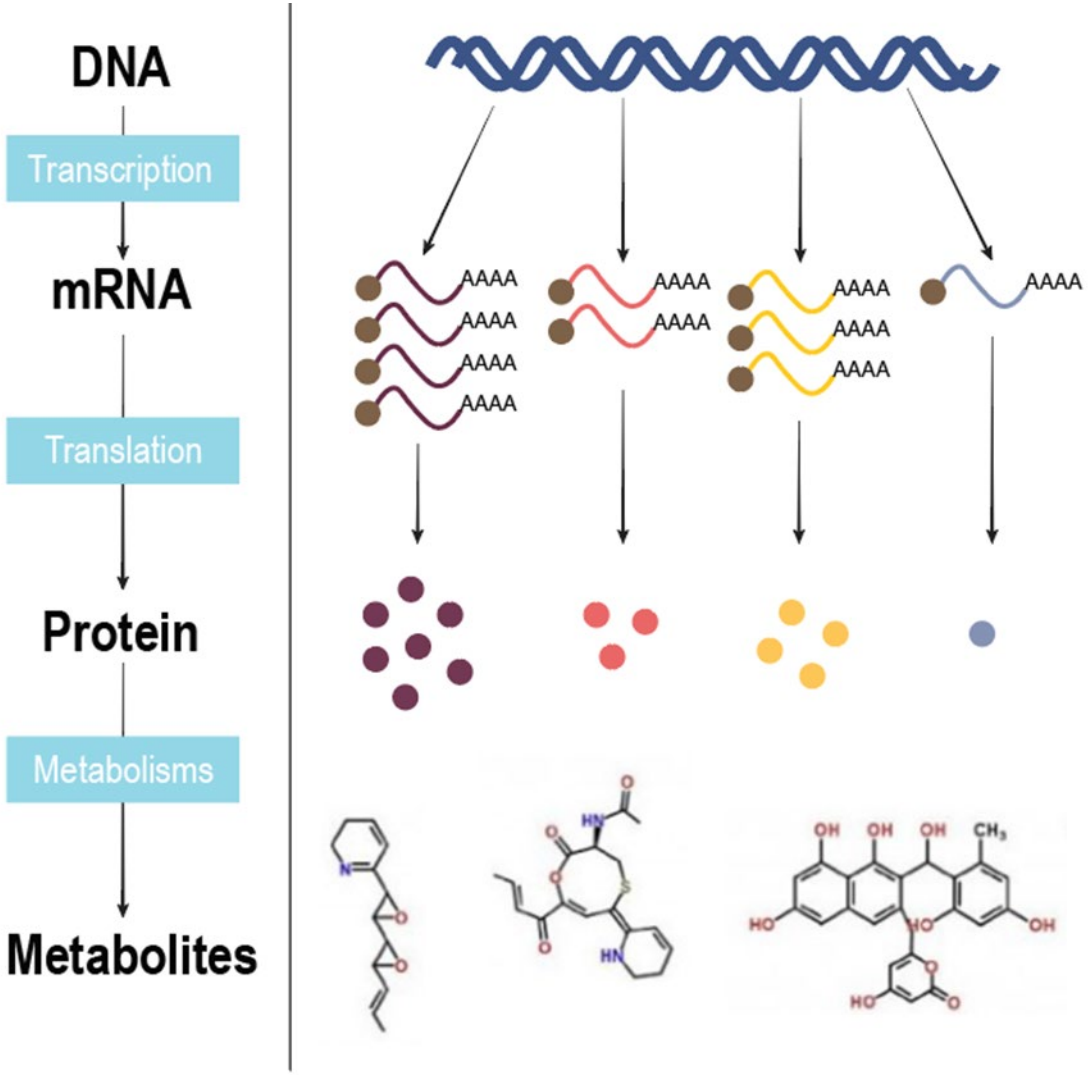
Why studying mRNA/transcriptomes?

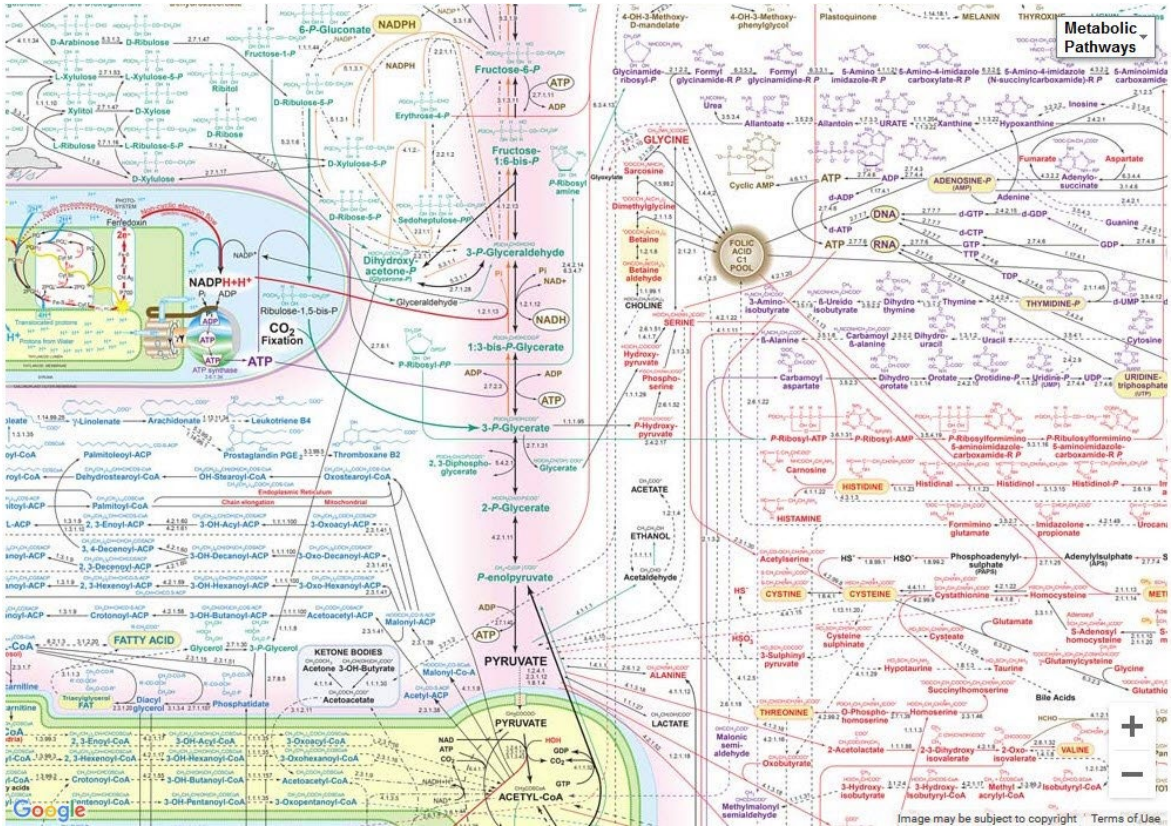


Created by Waradon Sungnak

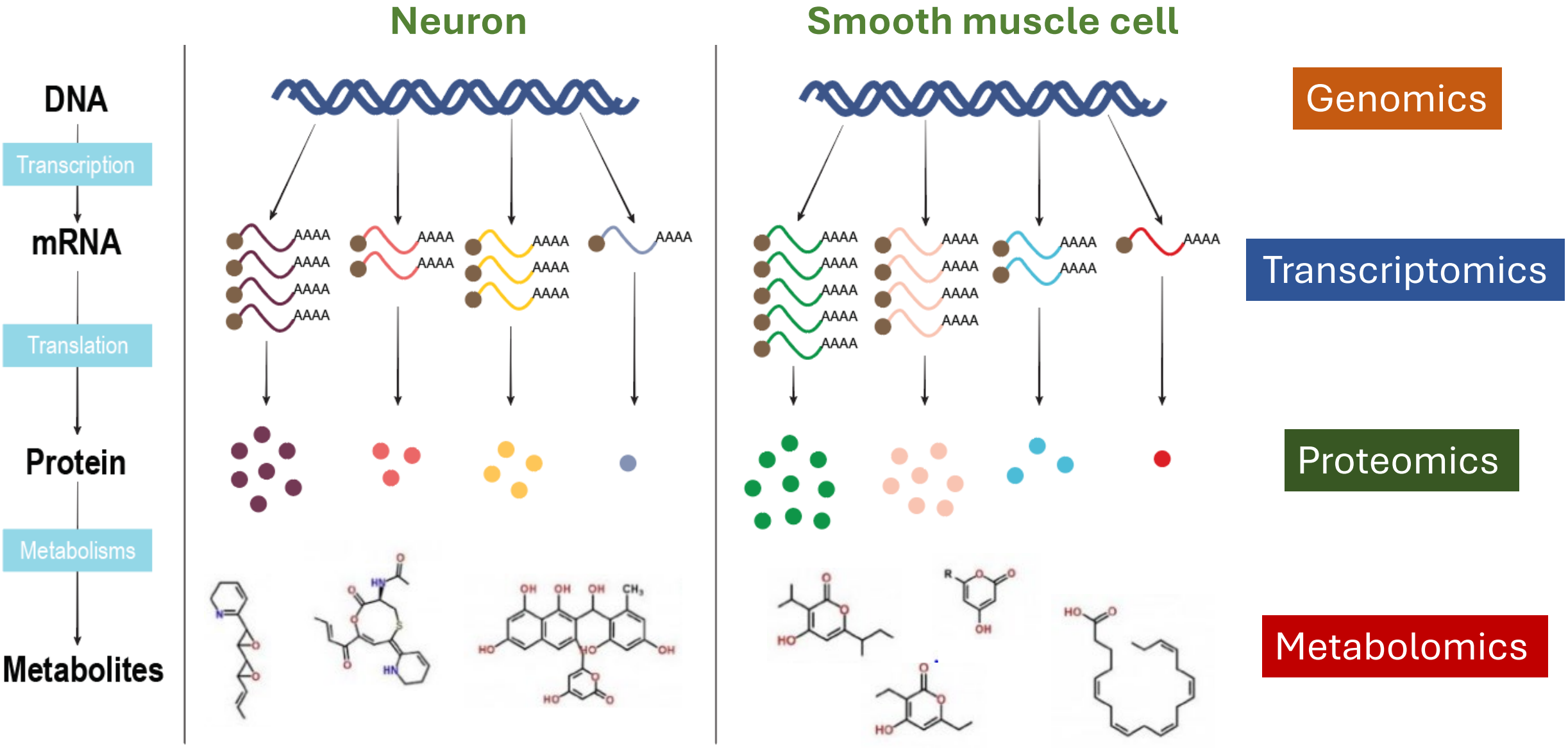
Transcriptomics/proteomics dictates cellular phenotype

The central dogma

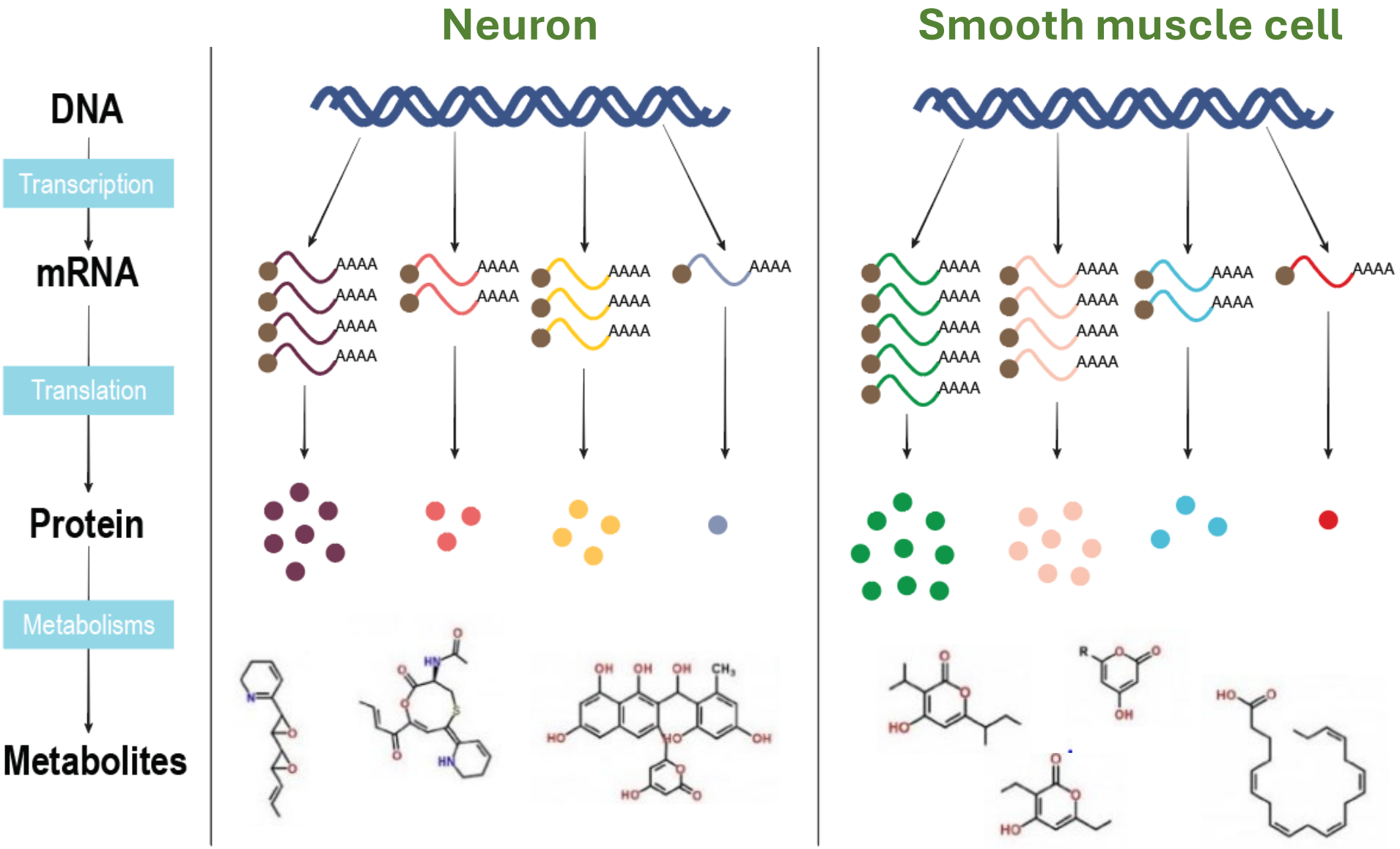




Comprehensive studies of the central dogma at different levels



Comprehensive studies of the central dogma at different levels



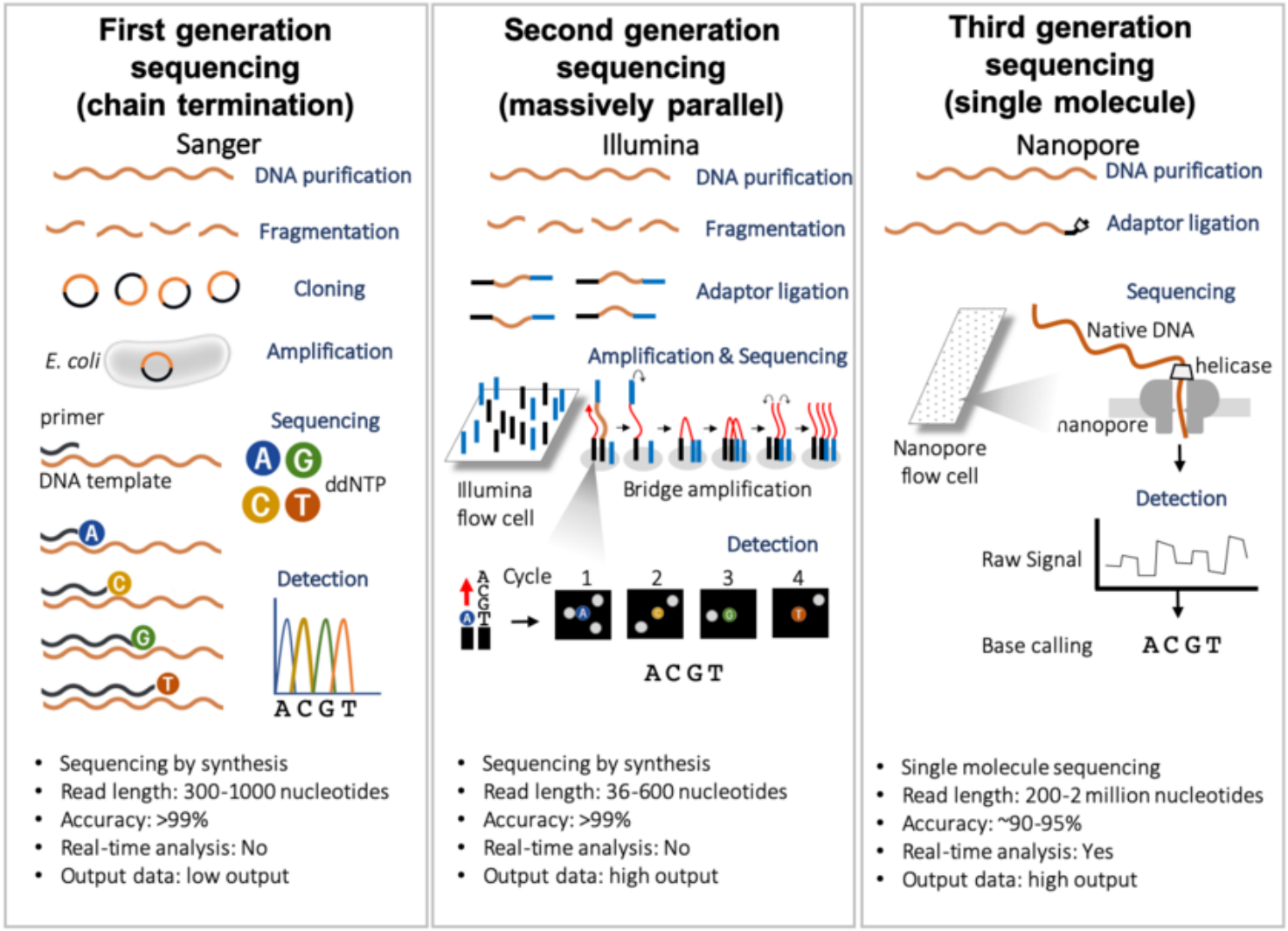
Next-generation sequencing

Mass spectrometry

Recap: 3 generations of DNA sequencing

1st Gen	2nd Gen	3rd Gen

Recap: 3 generations of DNA sequencing



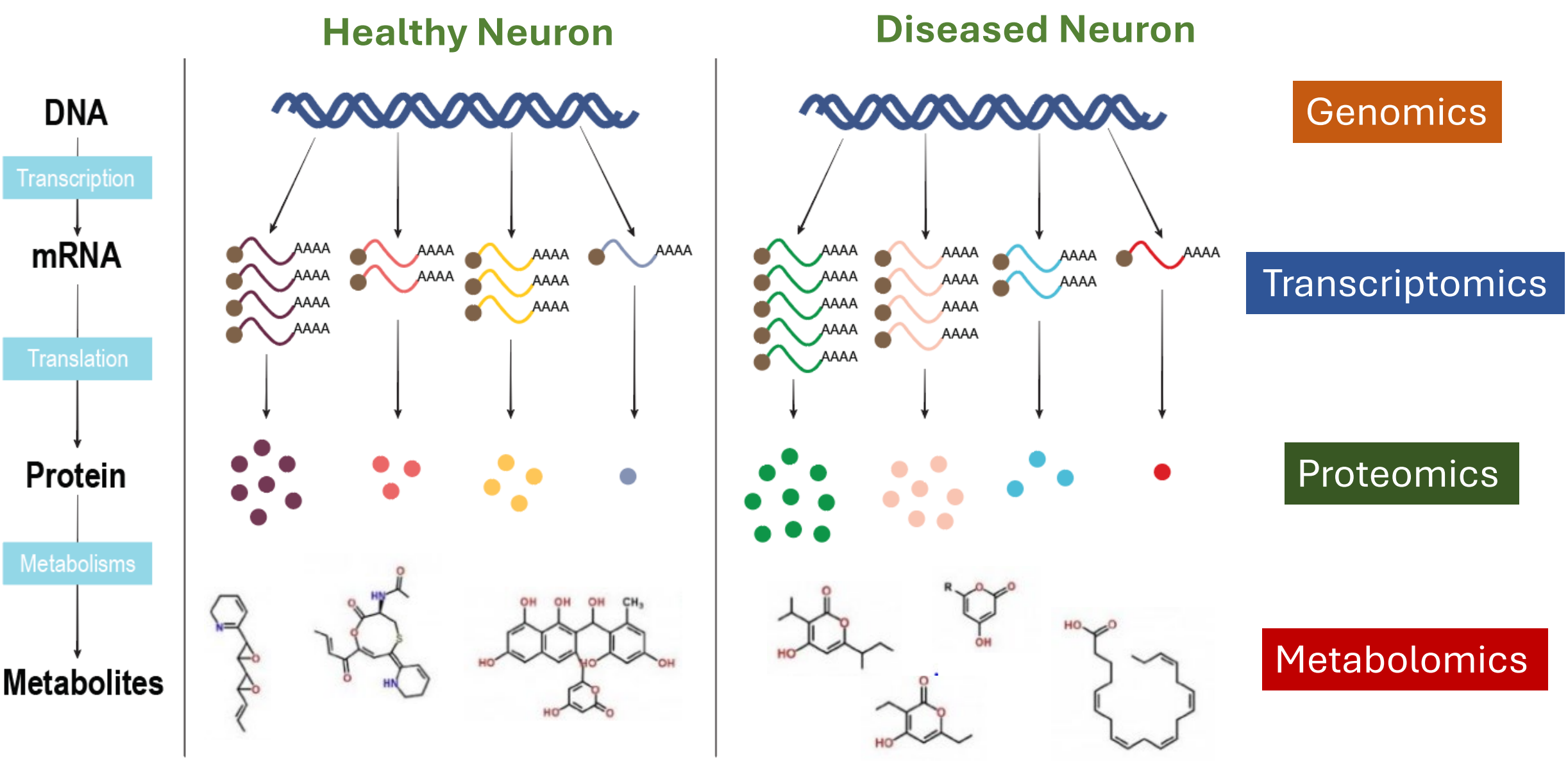
- Sequencing by synthesis
- Read length: 300-1000 nucleotides
- Accuracy: >99%
- Real-time analysis: No
- Output data: low output

- Sequencing by synthesis
- Read length: 36-600 nucleotides
- Accuracy: >99%
- Real-time analysis: No
- Output data: high output

- Single molecule sequencing
- Read length: 200-2 million nucleotides
- Accuracy: ~90-95%
- Real-time analysis: Yes
- Output data: high output

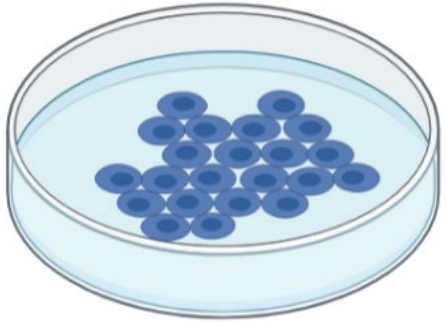
- 3 generations
  - 1st gen: Sanger
  - 2nd gen: Illumina
  - 3rd gen: Napopore, PacBio
- 2nd and 3rd gens are commonly called **next-generation sequencing (NGS)**
- Throughput: how many DNA species able to be sequenced at one time
  - Sanger: one or a few
  - NGS: a lot; millions to billions
- Illumina: short reads but the highest accuracy
- 3rd gen platforms: long reads but a bit lower accuracy

Comprehensive studies of the central dogma at different levels



Comprehensive studies of the central dogma at different levels







You have a new drug candidate, and you would like to check its effects *in vitro* and *in vivo*...



How do track phenotypic changes in a given condition/treatment?

You can measure in your selected “system”:

- Survival/morphological/clinical changes
- Changes in your gene(s)/protein(s) of interest by:

These methods require you to decide which gene(s)/protein(s) to detect…

* qRT-PCR (RNA level)
* Western blot/flow cytometry/ELISA (protein level)
* Immunofluorescence (protein level *in situ*)

How do track phenotypic changes in a given condition/treatment?

- But what if you don't know which gene/protein target(s) you should look at to track the potential change(s)?

- This is where transcriptomics/proteomics may help.

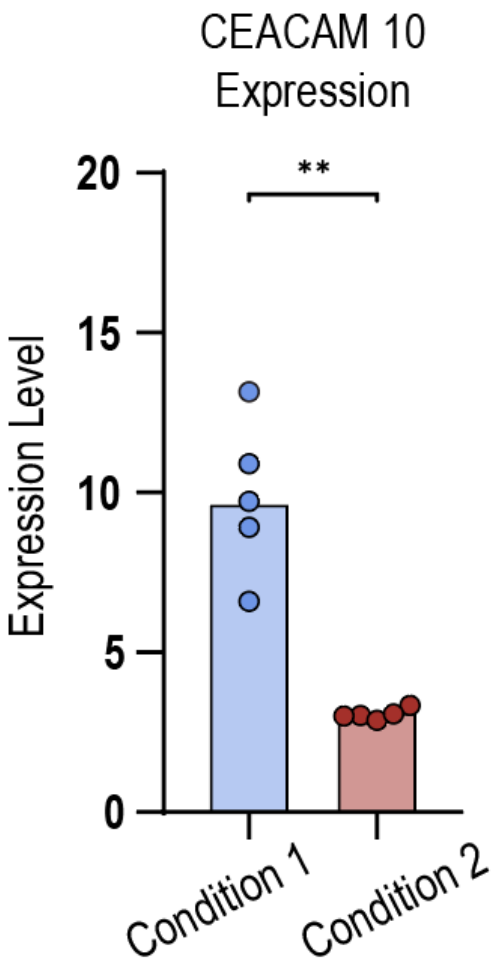
- Instead of checking changes in a selected target, you check the levels of **ALL mRNAs/proteins** in your samples.

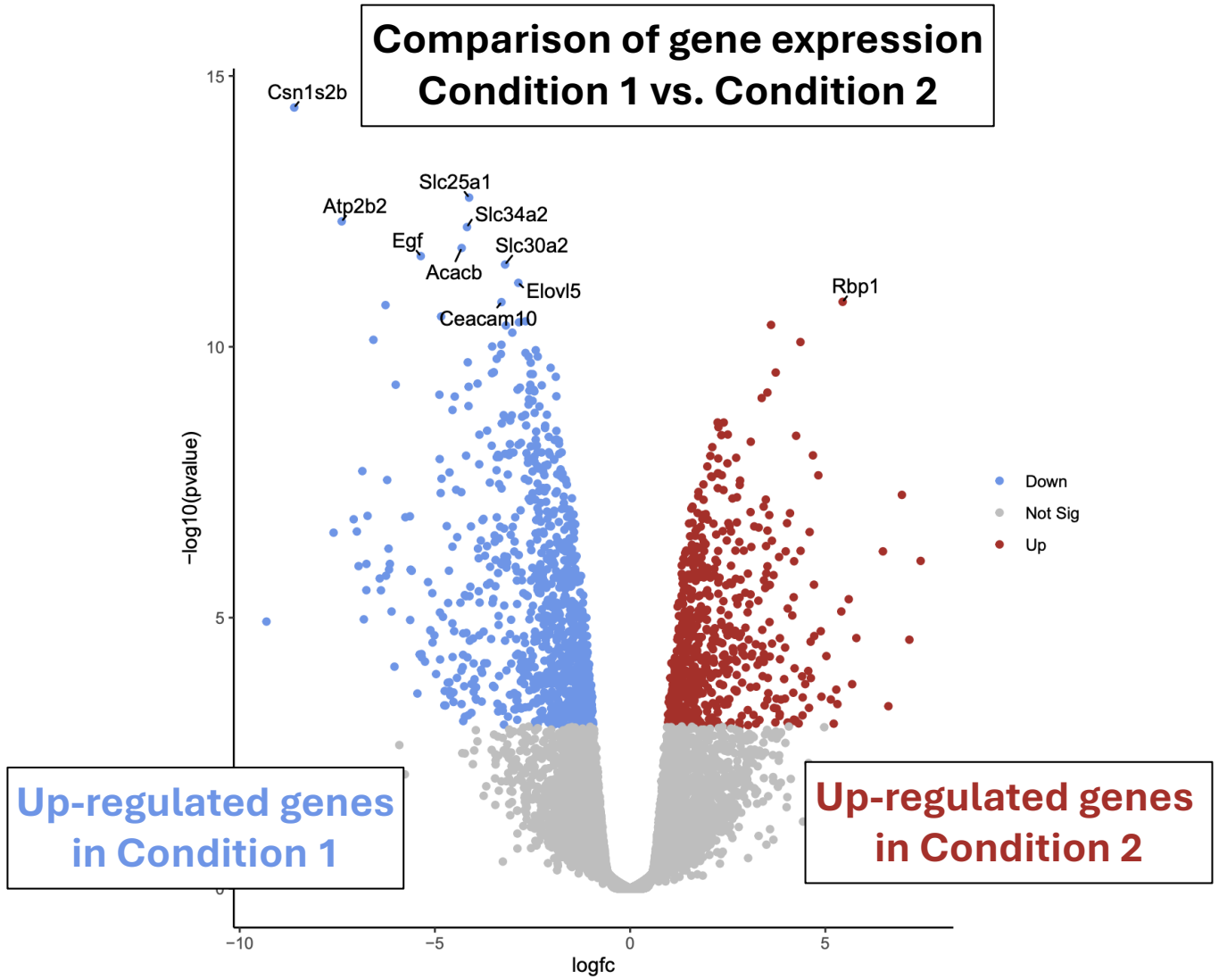
- You will see the landscape of RNA expression before and after the treatment. And you can see which genes are differentially expressed to be investigated further.

What could transcriptomics do?

- Systemic phenotyping: providing comprehensive, “unbiased” information on phenotypic changes at the mRNA level

Single-gene level vs. transcriptomic level detection





Overview of RNA detection approaches

- **Low-throughput:** one, a few, to a dozen genes at a time
  - qRT-PCR (or real-time RT-PCR)
  - Northern blot (barely used anymore)
- **Medium-throughput:** a few hundred genes at a time
  - NanoString nCounter Gene Expression
- **High-throughput:** thousands of genes at a time
  - Microarray
  - RNA-seq

Working with RNA is technically more demanding!

- Chemical stability of RNA is low
- RNases are everywhere in the environment!

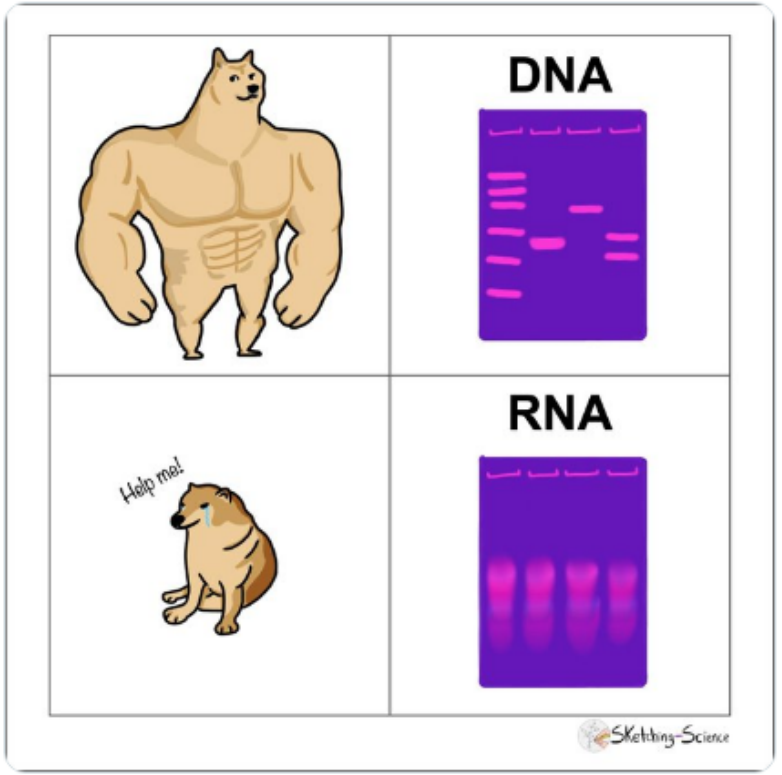


← Tweet

c0ncOrdance
@c0ncOrdance

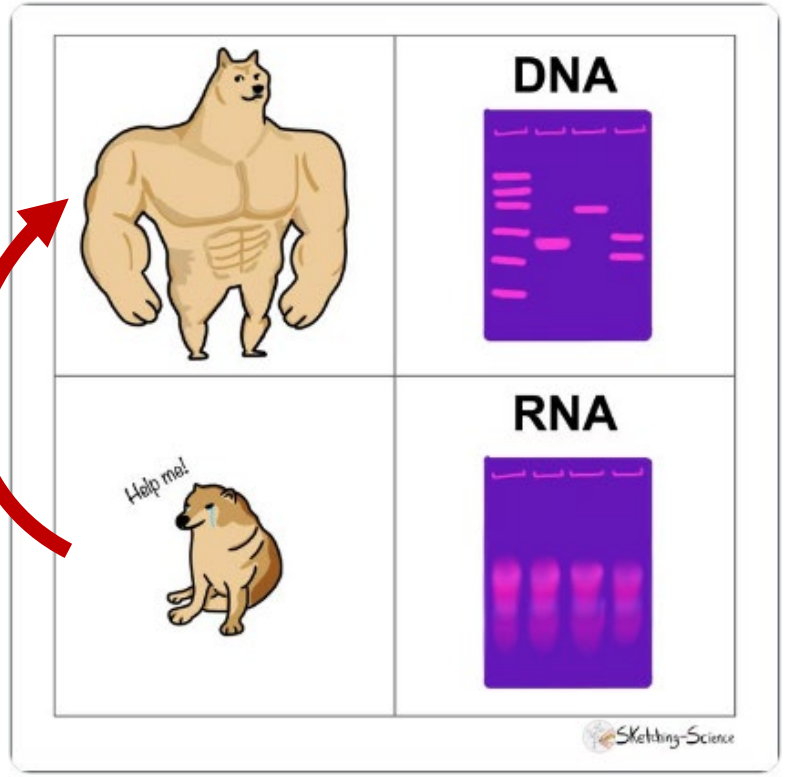
The only solution to RNase contamination:

1. Take a giant bathtub.
2. Fill it with RNaseZAP or similar RNase removal solution.
3. Put everything in the lab in the bathtub.
4. Get in the bathtub.
5. You live & work there now.

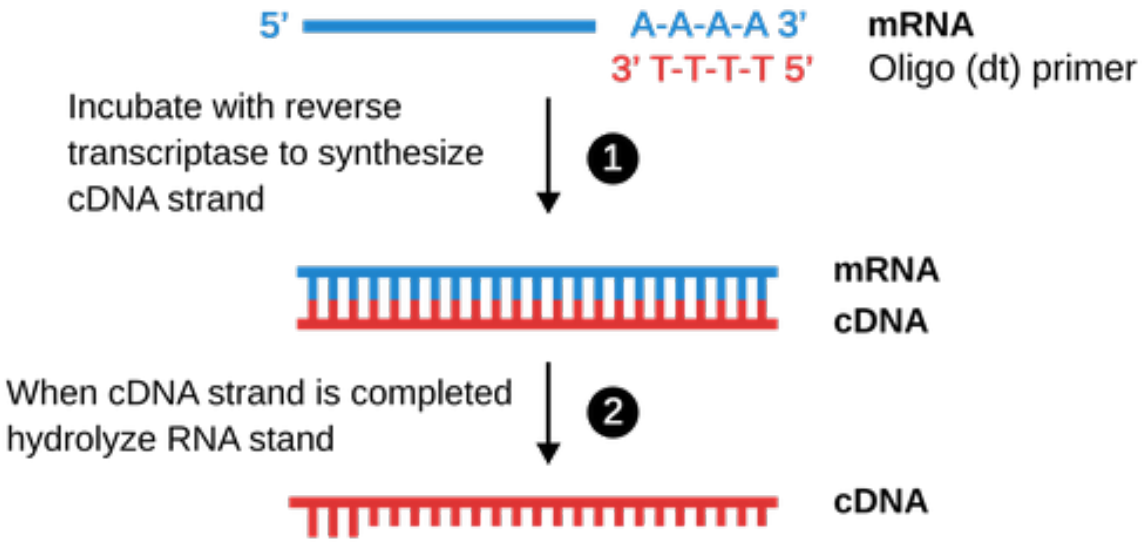


7:25 AM · Dec 9, 2020 · Twitter Web App

Reverse transcription (RT) is a key step for most of the RNA-based studies

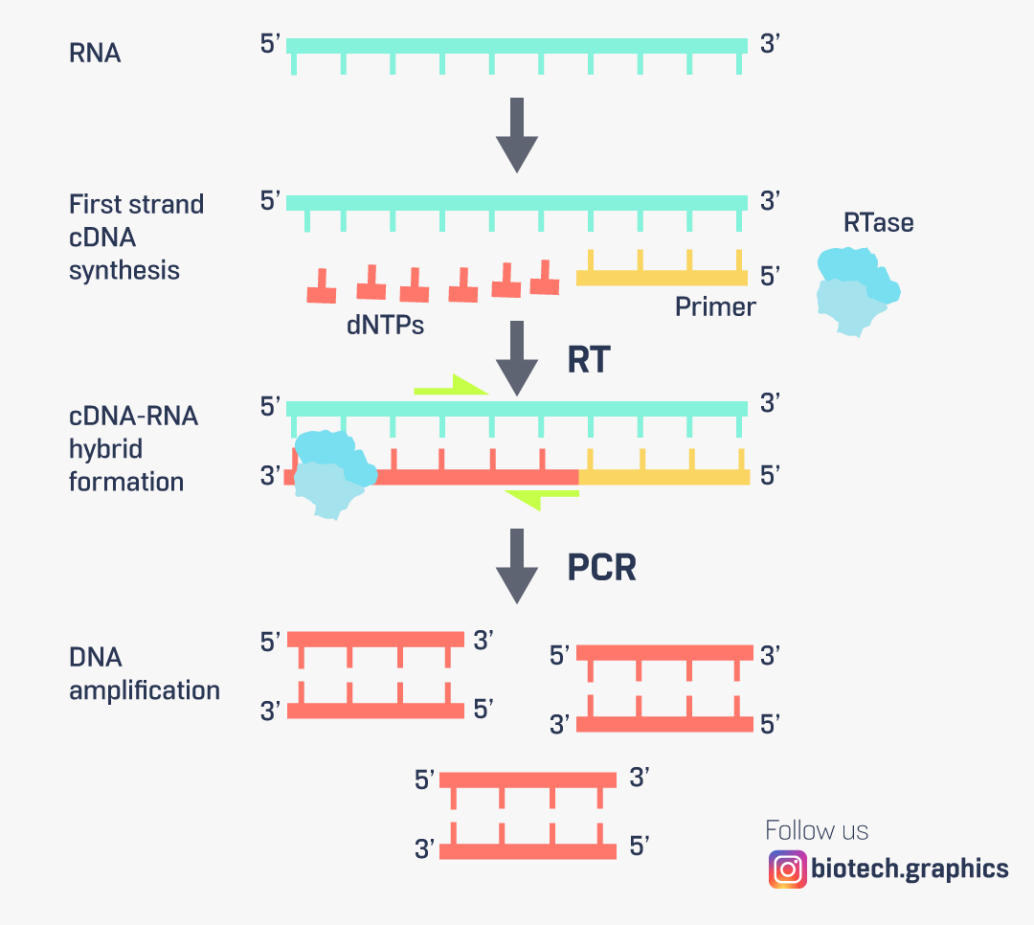


7:25 AM · Dec 9, 2020 · Twitter Web App



From mRNA to detectable signal

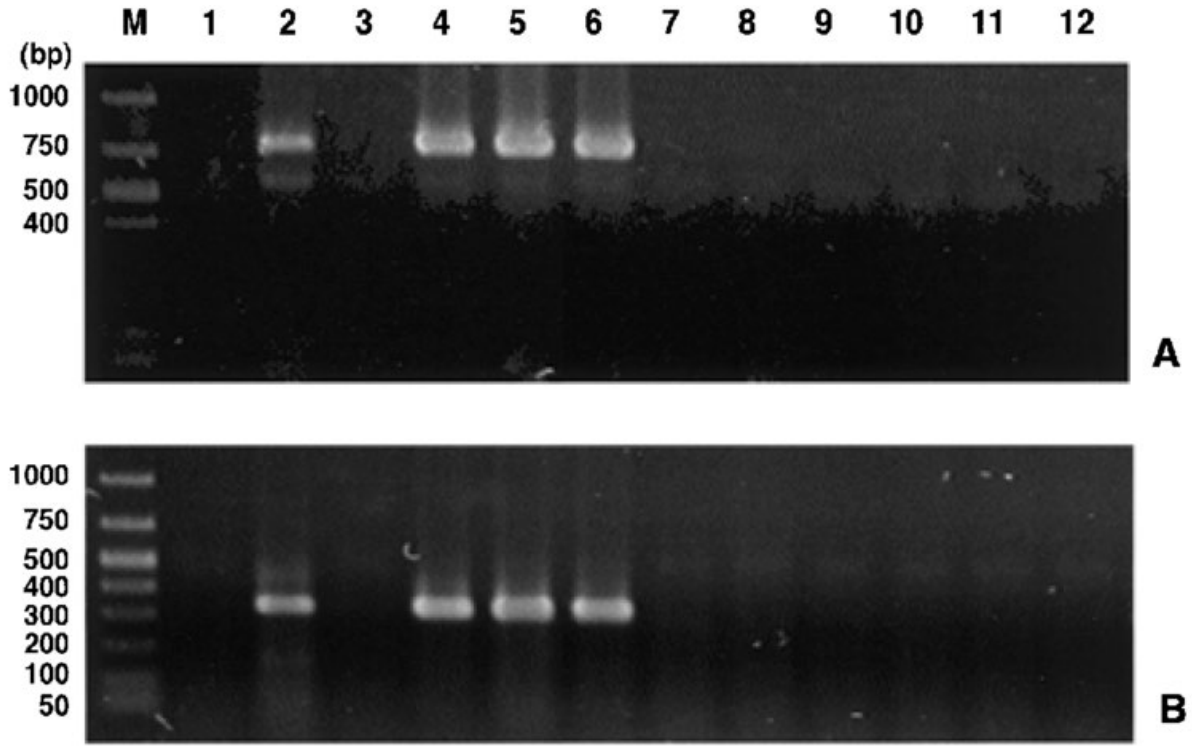
- RT converts **ALL mRNA** into cDNA
- It is a 1-to-1 reaction, so the **amount of cDNA remains very small** → **DIFFICULT to detect**
- With **PCR**, we can **amplify the cDNA**
- However, to perform PCR, we must **decide which gene we want to amplify**, as this requires designing specific primer pairs



With RT-PCR, if a gene of interest is expressed, you will have an amplification you can detect

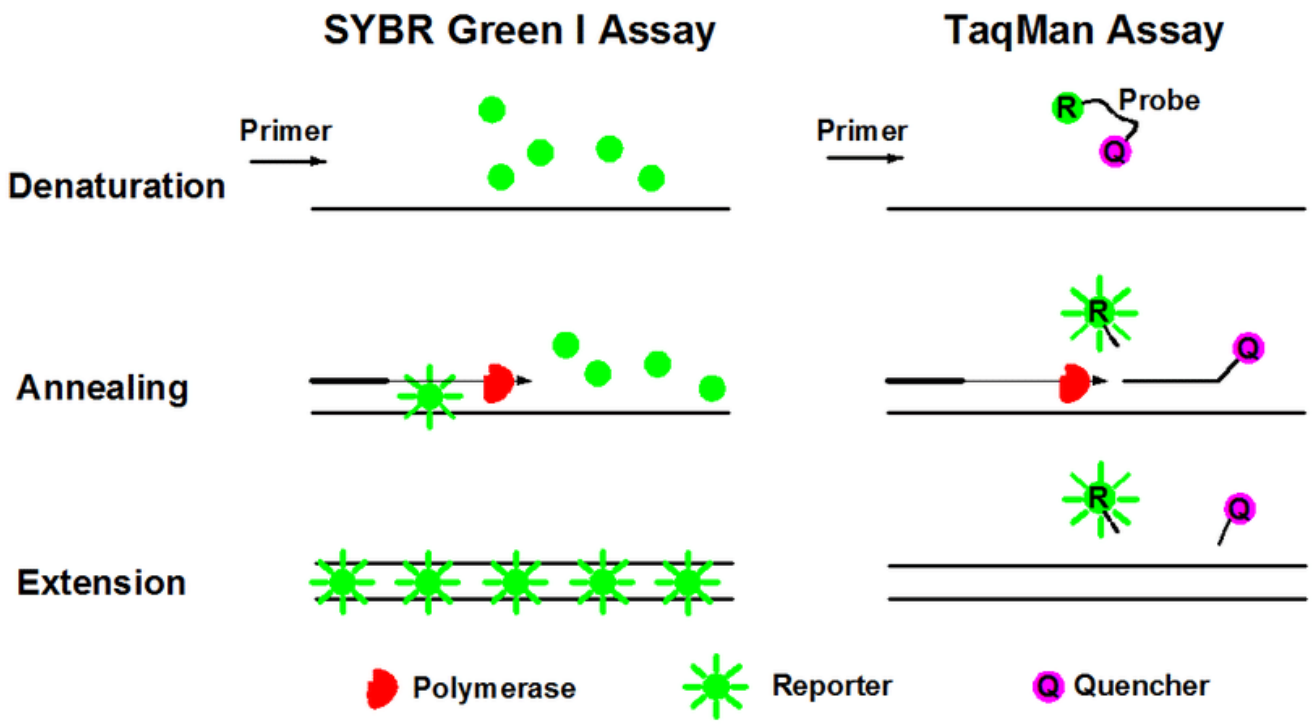
How do you detect such amplification?

Old school approach: run PCR products on gel electrophoresis



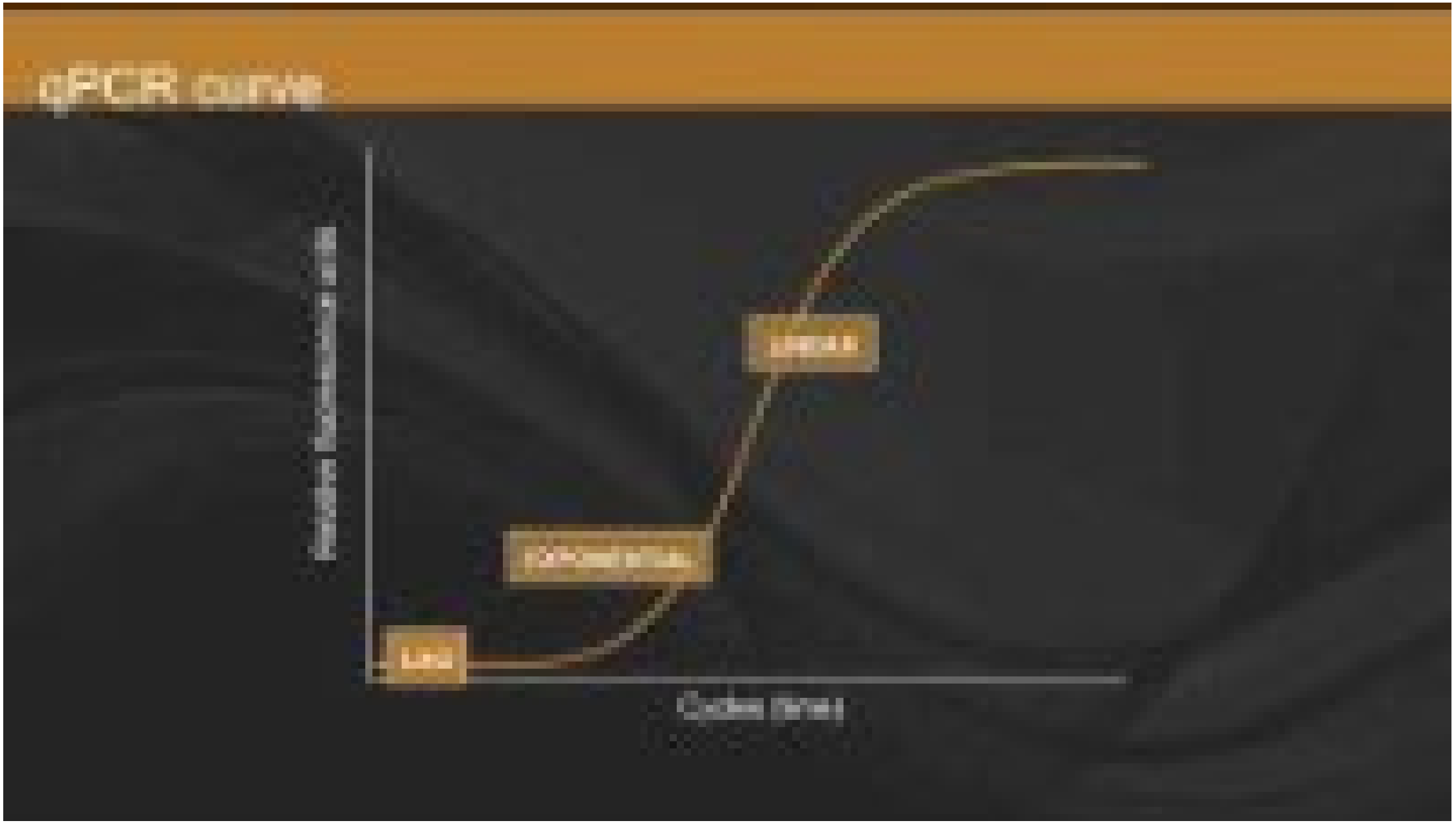
With the primers you design, you know the specific size of the PCR targets if there is any amplification

qRT-PCR: using fluorescence reporter to detect PCR amplification



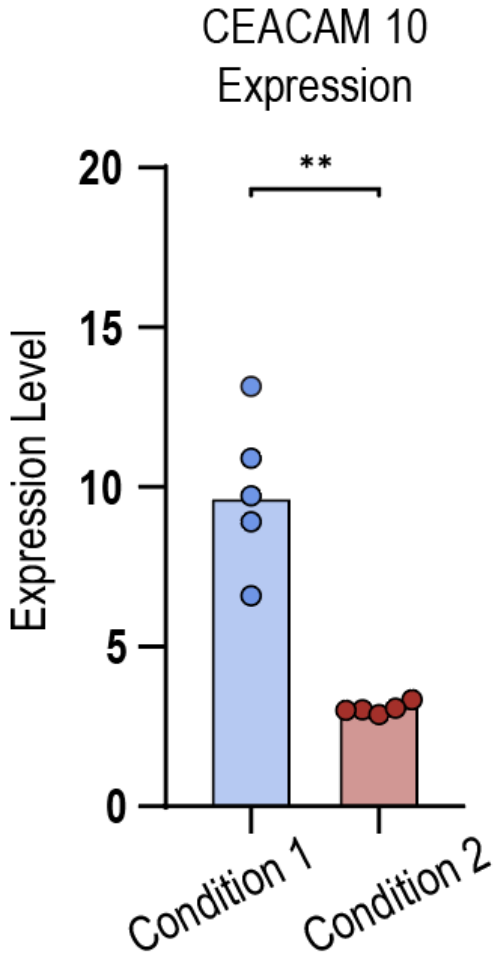
Cao et al. *Molecules*. 2020

- For SYBR green, you probably need to design/optimize primers (unless you rely on previous literature)
- For TaqMan, there are validated primer/probe for specific genes of interest for commonly used species



Overview of qPCR: https://www.youtube.com/watch?v=1kvy17ugI4w

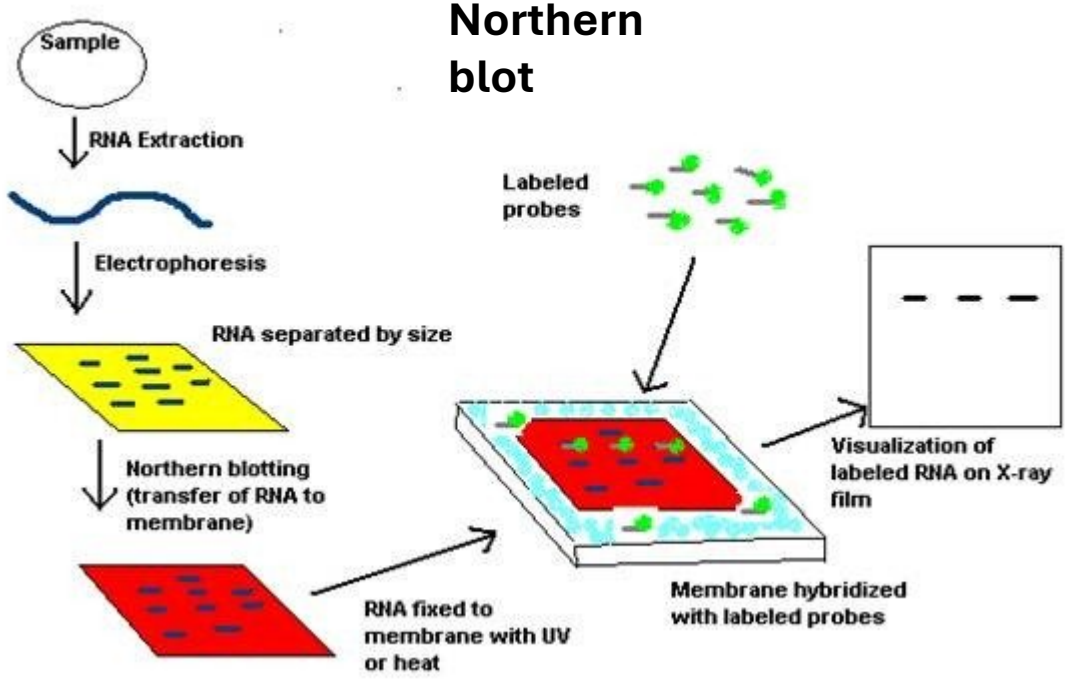
qRT-PCR results: detection of one or a few genes of interest at a time…



Northern blot: RNA detection without RT

- Low-throughput: one, a few, to a dozen genes at a time
  - qRT-PCR (or real-time RT-PCR)
  - Northern blot (barely used anymore)
- Medium-throughput (a few hundred genes at a time)
  - NanoString nCounter
- High-throughput (thousands of genes at a time)
  - Microarray
  - RNA-seq

Northern blot



***** EXTREMELY DIFFICULT TO PERFORM**

- You work directly on RNA. No RT to cDNA

- No amplification involved

- Samples exposed to ambient RNases a lot

https://www.mybiosource.com/learn/testing-procedures/northern-blotting/

Overview of RNA detection approaches

- Low-throughput: one, a few, to a dozen genes at a time
  - qRT-PCR (or real-time RT-PCR)
  - Northern blot (barely used anymore)
- Medium-throughput (a few hundred genes at a time)
  - NanoString nCounter
- High-throughput (thousands of genes at a time)
  - Microarray
  - RNA-seq

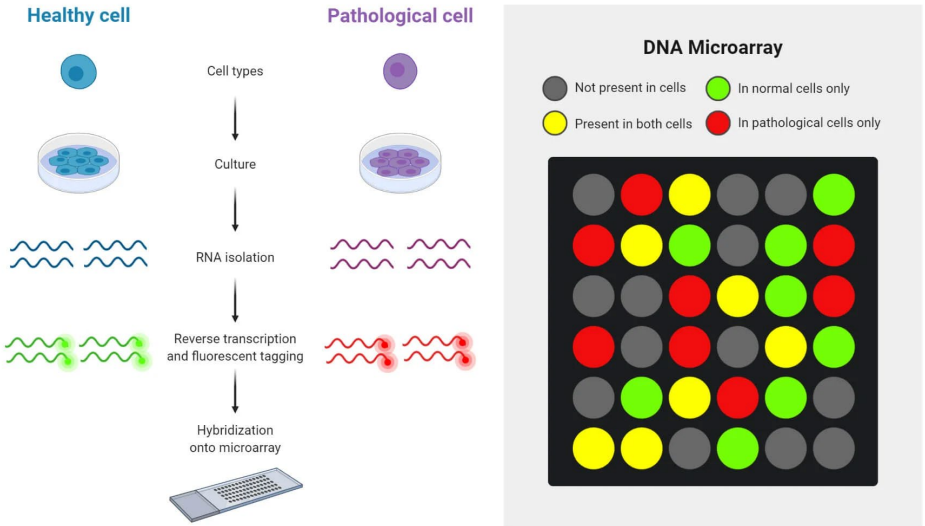
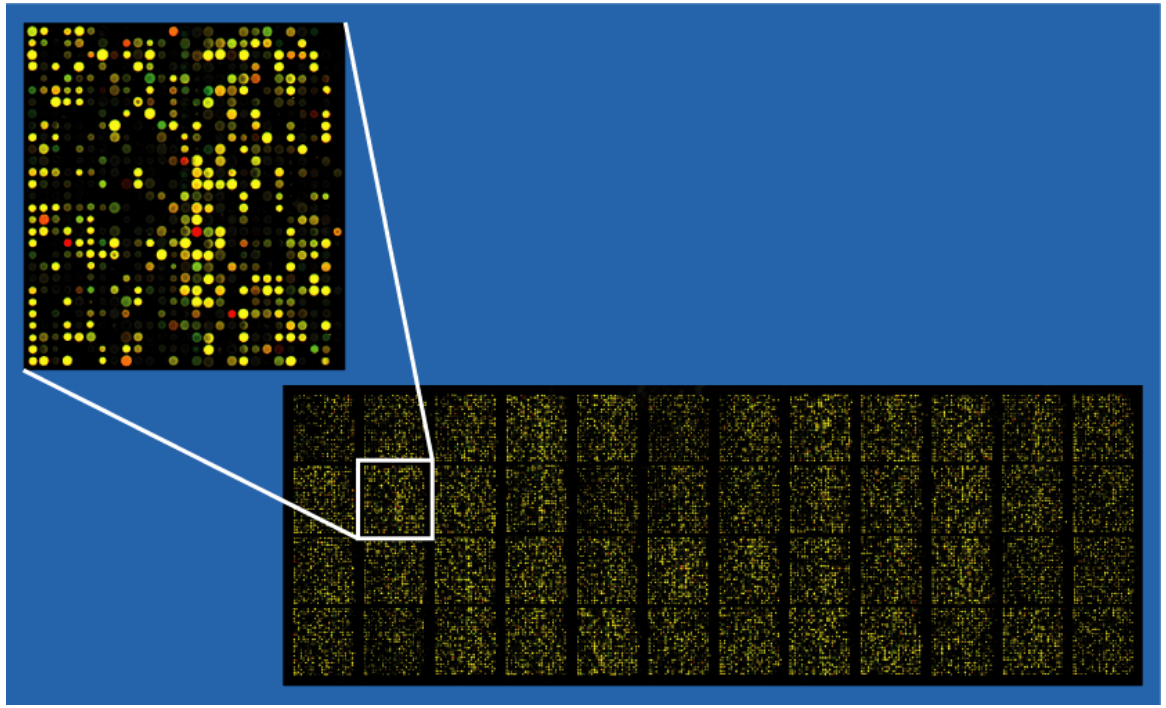
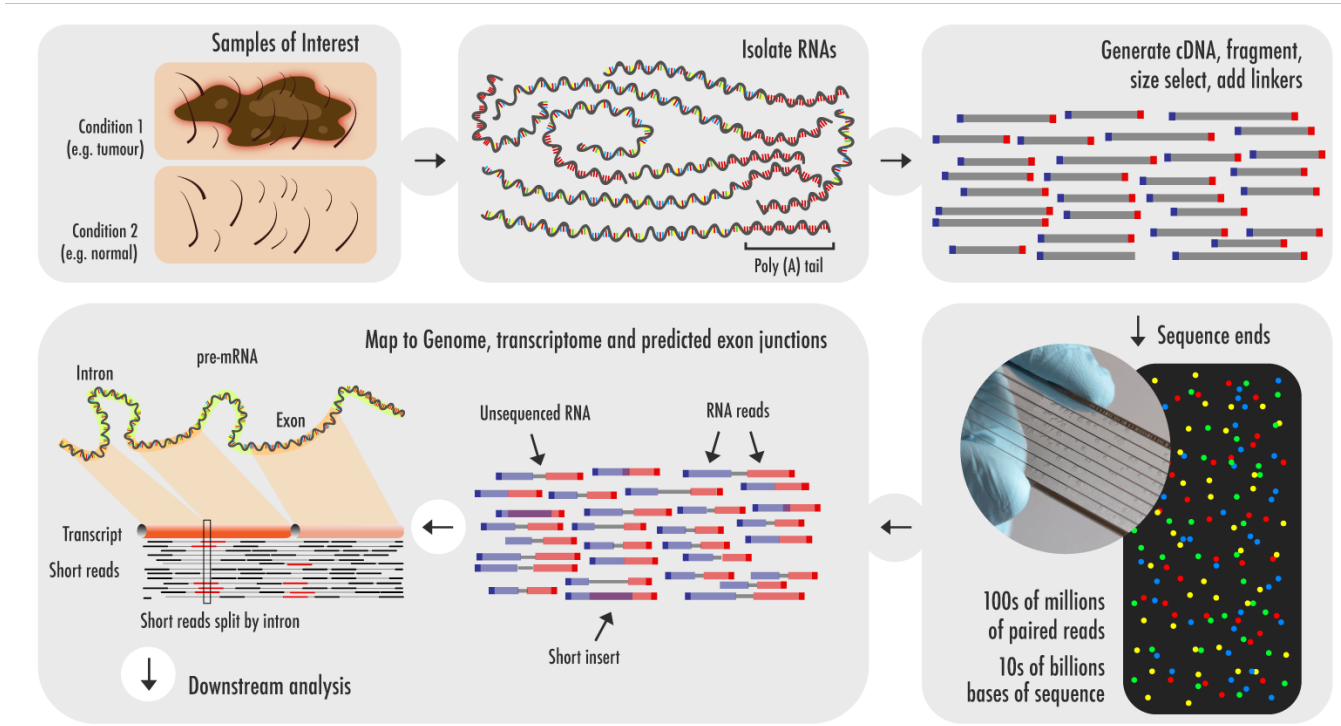


Image By Sagar Aryal, created using biorender.com



RNA-seq: the gold standard of transcriptomic detection

- Low-throughput: one, a few, to a dozen genes at a time
  - qRT-PCR (or real-time RT-PCR)
  - Northern blot (barely used anymore)
- Medium-throughput (a few hundred genes at a time)
  - NanoString
- High-throughput (>thousands of genes at a time)
  - Microarray
  - RNA-seq



Source: https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461

Review Quiz

Your lab has found that a gene called *TBX56* is up-regulated in multiple diseases.

You have liver cell samples from hepatitis patients (along with those from healthy controls). You'd like to check if *TBX56* mRNA level is higher in the liver cell samples of the patients.

Which RNA detection technique would you use? Why?



Review Quiz

You are interested in a rare brain disease your lab has just discovered.

You are tasked to find potential drug targets to treat the disease. You believe the targets should be genes/proteins highly expressed in neurons of the diseased patients, but not in healthy people.

You managed to get neural biopsy samples from those patients (and healthy controls).

Which RNA detection technique would you use to search for such targets? Why?



Take-home messages

- Cellular phenotype can be reflected by gene expression i.e. mRNA expression
- If you know a target gene to detect, you can use a technique that is based on gene-specific primers/probes following RT-PCR/qRT-PCR

- Transcriptomics – comprehensive information on gene expression (-omics ~ study of ALL)
- **Detection of transcriptomes ~ RNA-seq**

A generic RNA-seq workflow

1. RNA extraction
2. Reverse transcription into cDNA
3. Adapter ligation
4. Amplification
5. Sequencing

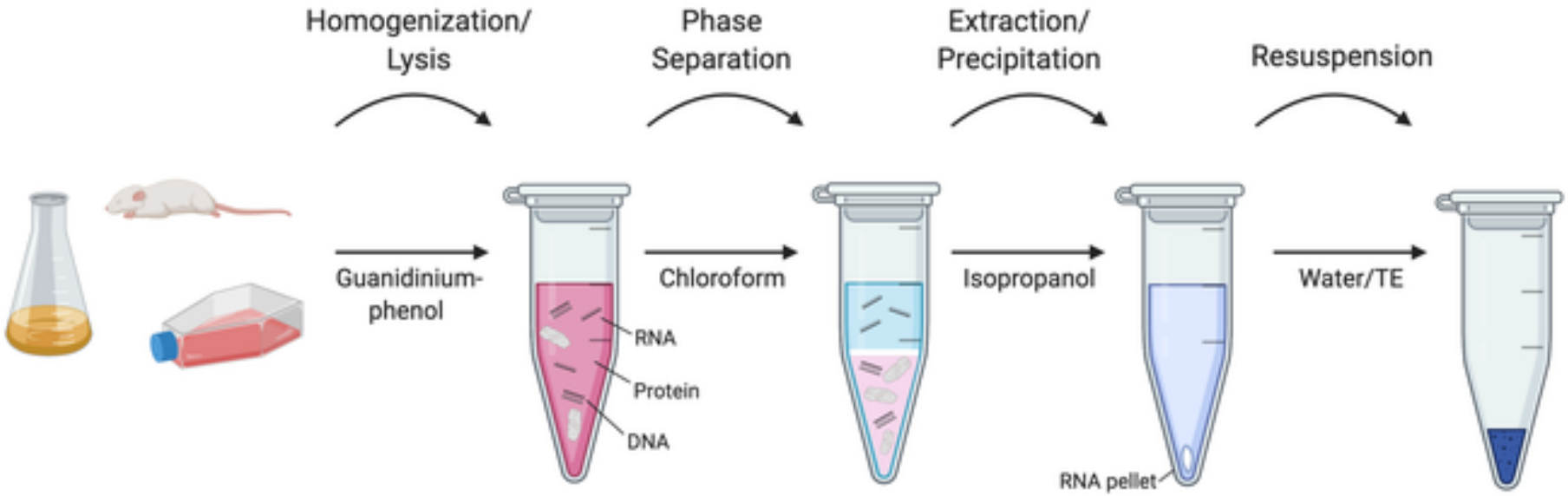
***Steps 2-5 can be variable, depending on different kits

Currently, Illumina sequencing is the most popular approach for a regular RNAseq experiment

A generic RNA-seq workflow

1. RNA extraction
2. Reverse transcription into cDNA
3. Adapter ligation
4. Amplification
5. Sequencing

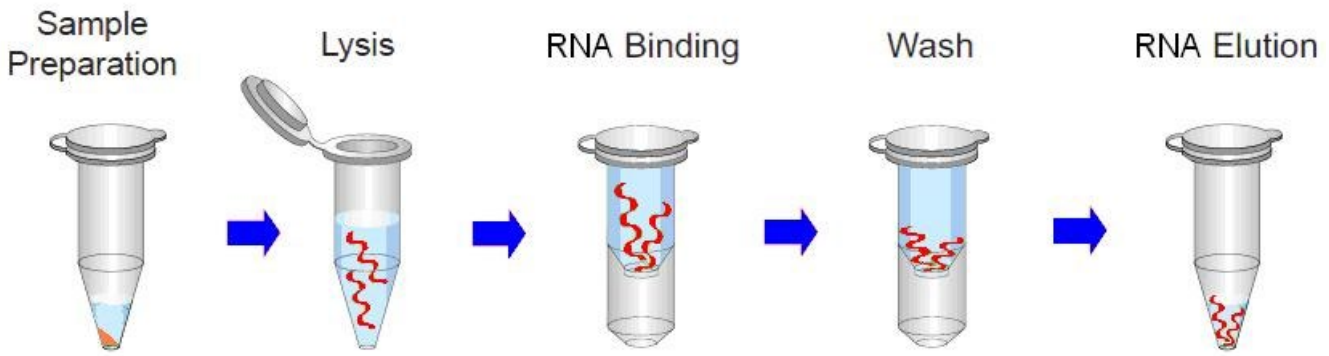
RNA extraction: phenol-chloroform-based/TRIzol-based approach



https://www.addgene.org/protocols/kit-free-rna-extraction/

RNA extraction: column-based



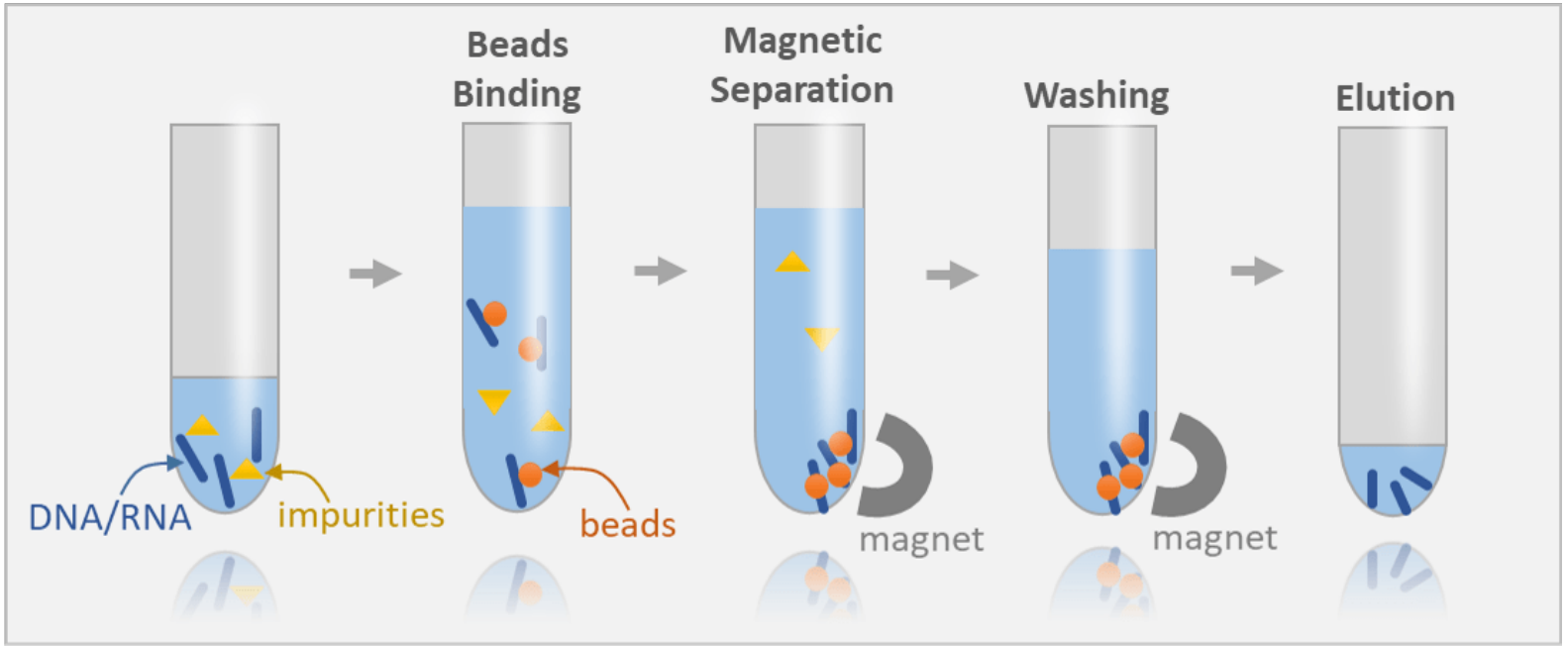


https://www.fairbiotech.com/shopping_show-2a14901a591a2a.html

RNA extraction: magnetic beads-based







https://biodynami.com/product/magnetic-beads-dna-rna-purification/

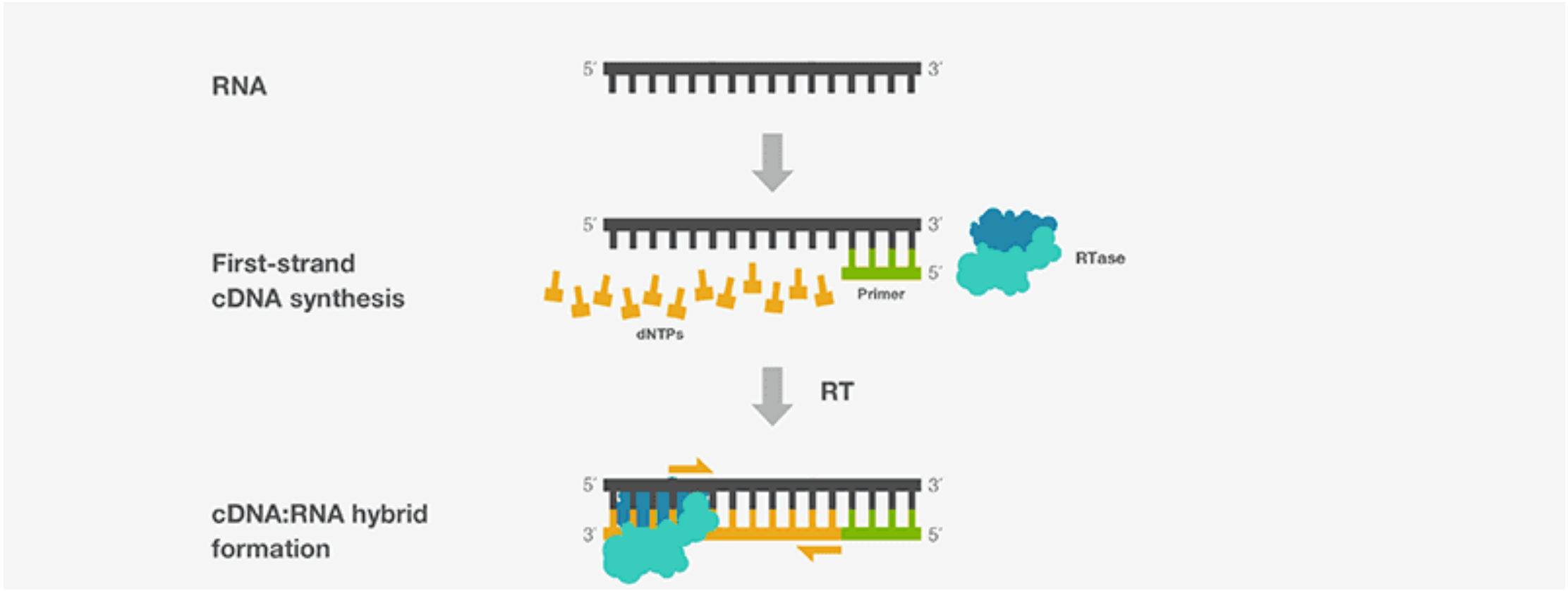
Summary: RNA extraction technique examples

1. Phenol-chloroform/TRIzol: cheapest
2. Column-based: convenient and cost-effective
3. Magnetic beads-based: super sensitive but expensive

A generic RNA-seq workflow

1. RNA extraction
2. Reverse transcription into cDNA
3. Adapter ligation
4. Amplification
5. Sequencing

Reverse transcription



If we'd like to synthesize cDNA from mRNA, what kind of RT primer should we use?

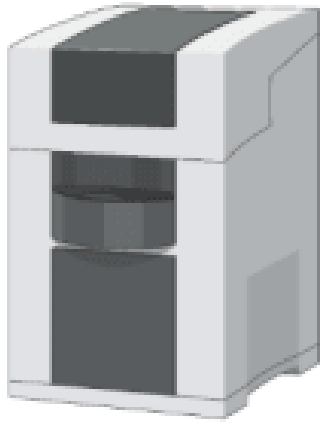
A generic RNA-seq workflow

1. RNA extraction
2. Reverse transcription into cDNA
3. Adapter ligation
4. Amplification
5. Sequencing

Recap

DNA Sequencing Sequencing Tools

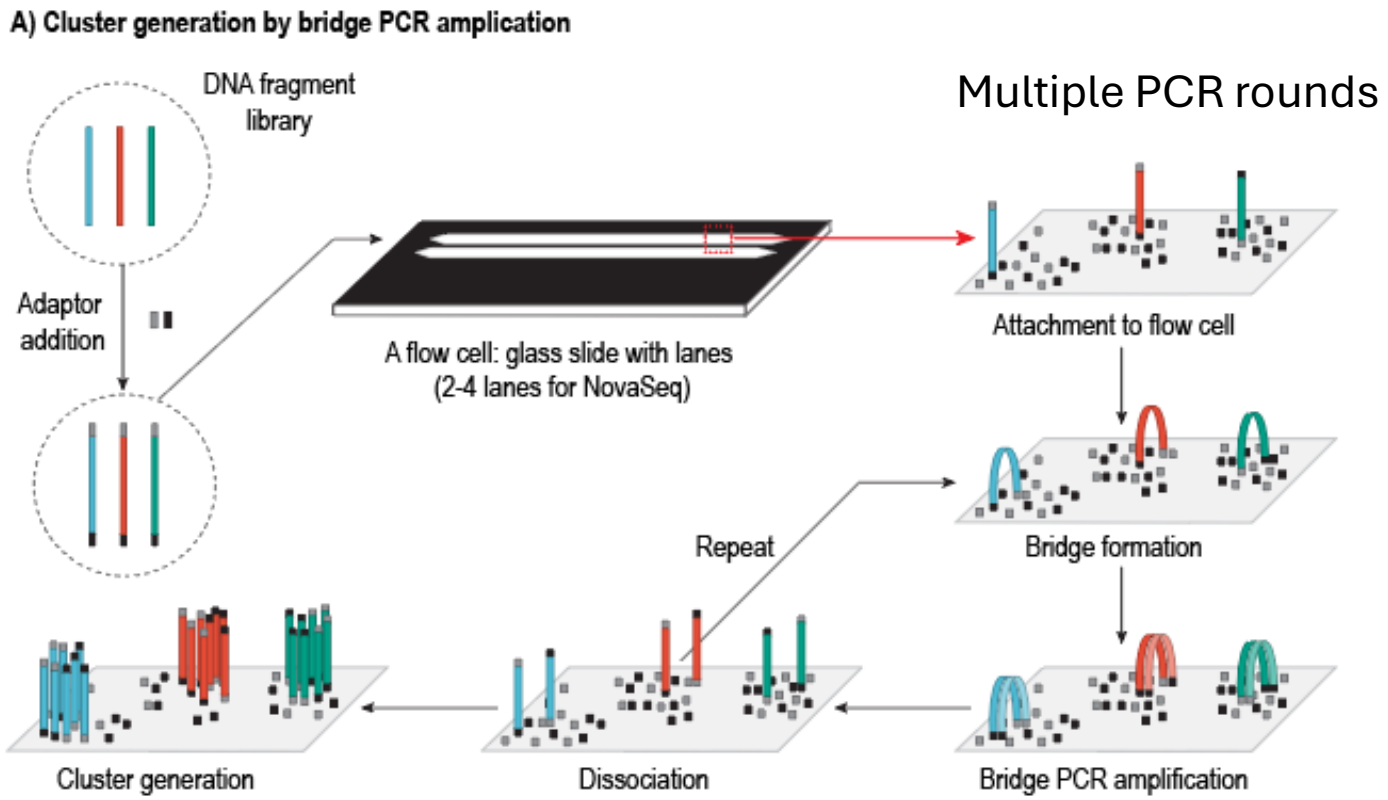
1st Generation	2nd Generation	3rd Generation
Sanger Sequencing	**Illumina**	**Oxford Nanopore Technology**
500-1000bp	50-500bp	**Pacific Biosciences**
		10s of kb







Illumina sequencing: recap

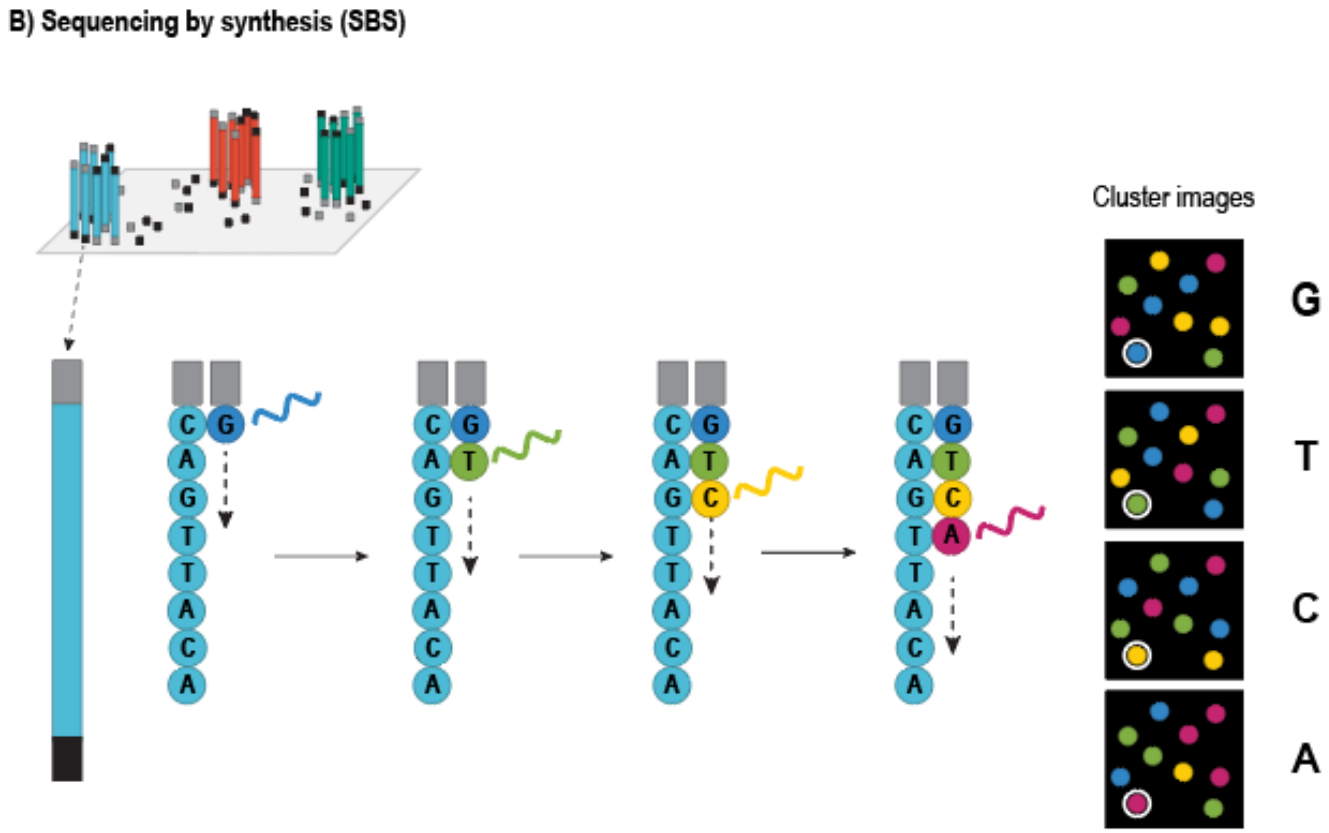


Multiple PCR rounds



Created by Waradon Sungnak

Illumina sequencing



- Illumina can only sequence short DNA, with the length of 50-300 bps
- For transcriptomics, the recommended length is 150bps BUT performed TWICE from each end...

Created by Waradon Sungnak

Benchtop Sequencers | Production-Scale Sequencers


iSeq 100


MiniSeq


MiSeq Series


NextSeq 550 Series


NextSeq 1000 & 2000

Popular Applications & Methods	iSeq 100 Key Application	MiniSeq Key Application	MiSeq Series Key Application	NextSeq 550 Series Key Application	NextSeq 1000 & 2000 Key Application
Large Whole-Genome Sequencing (human, plant, animal)					
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●	●
Exome & Large Panel Sequencing (enrichment-based)				●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)				●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)				●	●
Targeted Gene Expression Profiling	●	●	●	●	●
miRNA & Small RNA Analysis	●	●	●	●	●
DNA-Protein Interaction Analysis (ChIP-Seq)			●	●	●
Methylation Sequencing				●	●
16S Metagenomic Sequencing		●	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)				●	●
Cell-Free Sequencing & Liquid Biopsy Analysis				●	●

Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11-48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	360 Gb *
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.2 billion *
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 300 bp

Benchtop Sequencers | Production-Scale Sequencers



NextSeq 1000 & 2000



NovaSeq 6000 Series



NovaSeq X Series

Popular Applications & Methods	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)		●	●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●
Exome & Large Panel Sequencing (enrichment-based)	●	●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	●	●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	●	●	●
Chromatin Analysis (ATAC-Seq, ChIP-Seq)	●	●	●
Methylation Sequencing	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)	●	●	●
Cell-Free Sequencing & Liquid Biopsy Analysis	●	●	●

Run Time	11-48 hours	~13–38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)	~15–21 hours (1.5B flow cells‡) ~18–25 hours (10B flow cells‡) ~48 hours (25B flow cells‡)
Maximum Output	360 Gb *	6000 Gb	16 Tb
Maximum Reads Per Run	1.2 billion *	20 billion	26 billion (single flow cells) 52 billion (dual flow cells)
Maximum Read Length	2 × 150 bp	2 x 250 bp**	2 × 150 bp

Overview of RNA detection approaches

- Low-throughput: one, a few, to a dozen genes at a time
  - **qRT-PCR (or real-time RT-PCR)**
  - Northern blot (barely used anymore)
- Medium-throughput (a few hundred genes at a time)
  - NanoString
- High-throughput (thousands of genes at a time)
  - Microarray
  - **RNA-seq**



You can basically detect ALL expressed mRNA aka. “transcriptome”

As of now you can assume:

Transcriptomics ~ RNA-seq

Quiz: what is the major disadvantage of Illumina sequencing?

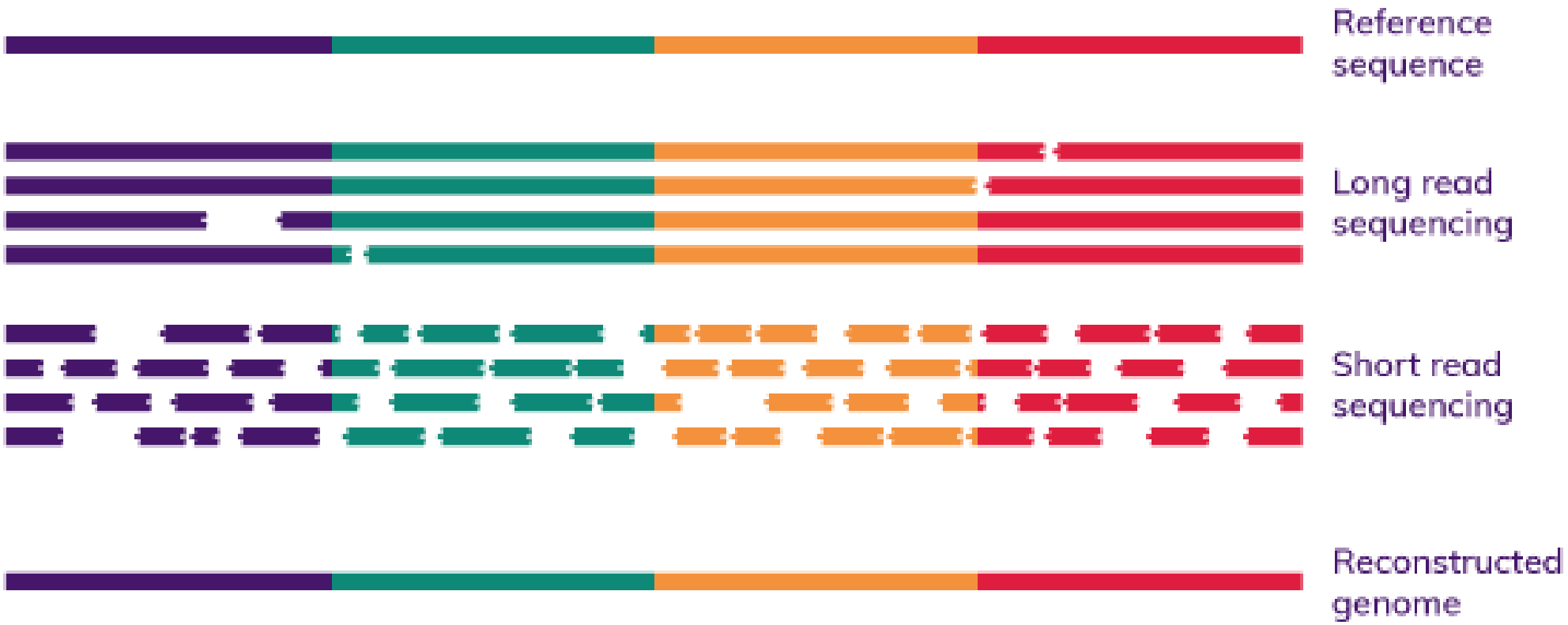
Recap on Illumina sequencing

- You can only sequence SHORT DNA
- For Illumina sequencing, the currently recommended length is 50-75 bps; preferably sequenced TWICE from each end of a DNA fragment, typically shorter than genome sequencing (why is that?)

- What we get out from a sequencer is called READs, which should be 50-75 bps

Again!

A “read” = a DNA fragment with a known sequence (e.g. an output from sequencer)



Sequencers





“Reads”

What does a “read” look like?

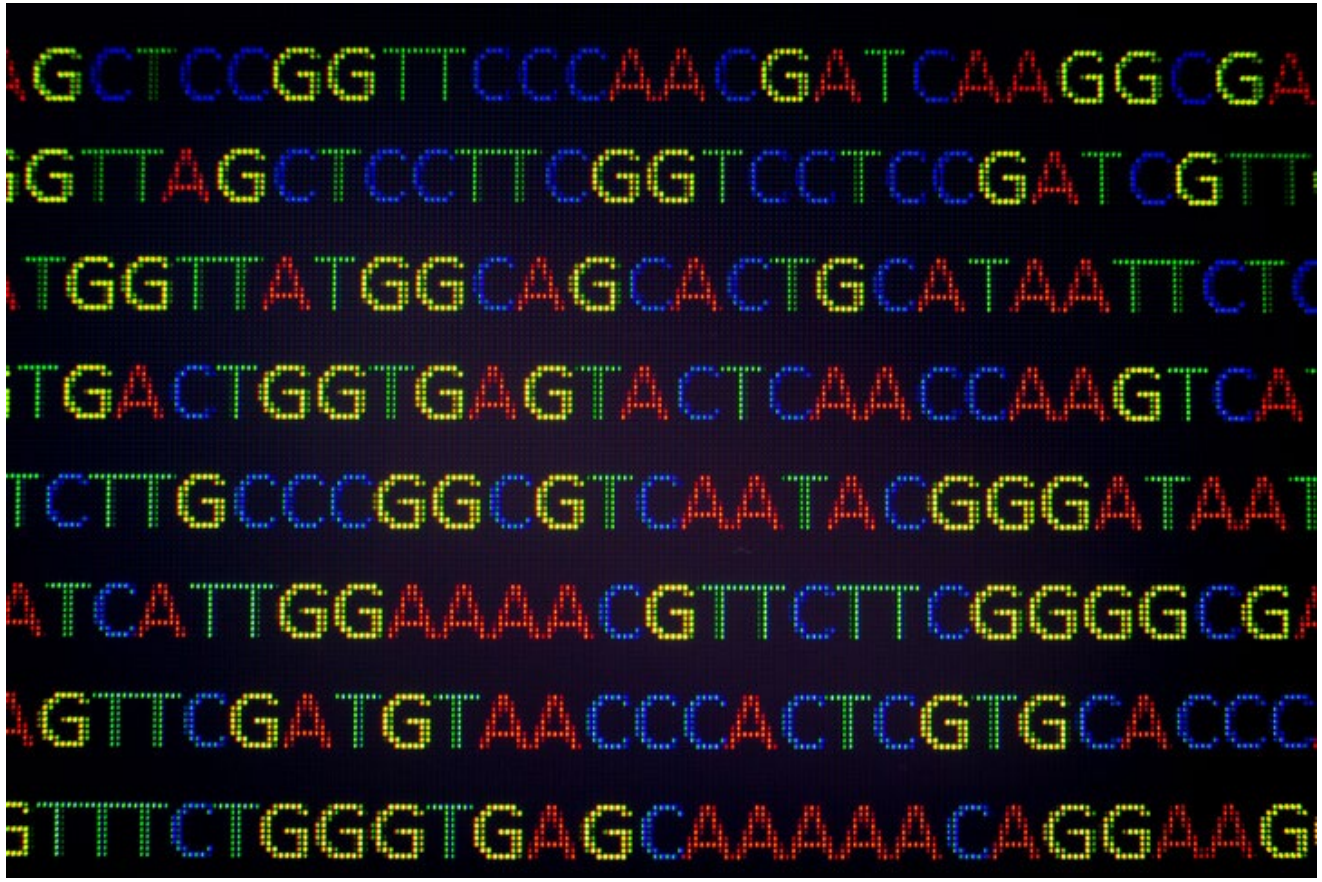
A read sequence in FASTQ format

Header Sequence Quality

```
@HWI-ST227:389:C4WA2ACXX:7:1204:2272:59979
GGAGGAAGGTCCTCGCTCCTCTTTCATATAAGGGAAATGGCTGAAT
+
FFFFHHHHHHJIJJJJJJJJIJJJIGIGIGGIJJIJIJJJJJJIII
@HWI-ST227:389:C4WA2ACXX:7:1205:15214:42893
GAGGATCCCAGGGAGGAAGGTCCTCGCTCCTCTTTCATCTAAGGGA
+
12BAFB?A:3<AE1@<FF;1*@EG*)?0?DBD>9BF9B*?######
@HWI-ST227:389:C4WA2ACXX:8:2208:2467:44624
AAAGAGGAGAGAGGACCATCCTCCCTGGGATCCTCAGAAGTCTACT
+
BDDA:DB?2AA@FC>F?EEGC<FED>GFD;?GBB?<?F99*/9?9?
```

Now what? How do we make sense of these?

How do we make sense of these biological sequences?



You'd want to make sense of the "reads" by comparing them to known "reference" sequences

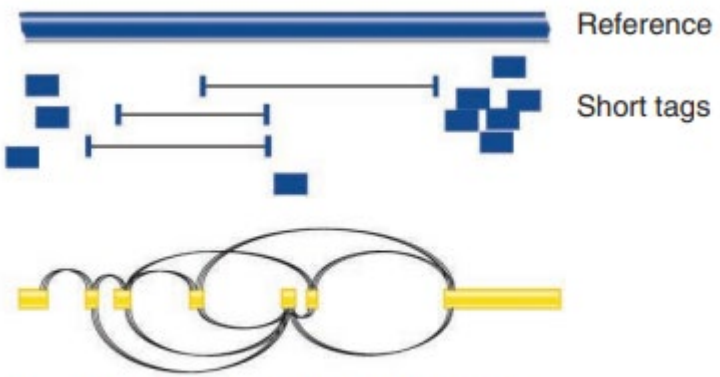
What is this process called?

Read alignment/mapping

- You can make sense of a read by comparing it to known reference sequences
- "Alignment" and "mapping" are often used interchangeably, as many tools would do both for us
- However, there are some subtle differences:
  - When we **align** a read, we're asking for not just where it likely came in the genome, but the exact base-to-base correspondence
  - When we **map** a read, we're just asking, "where did it come from?" We don't necessarily care about the exact alignment between the read and where it came from

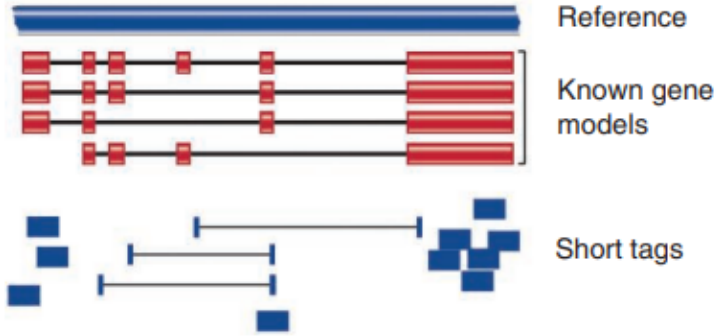
Three RNA-seq alignment strategies

Align to reference genome



Infer possible transcripts and abundance

Align to transcriptome



Use known and/or predicted gene models to examine individual features

De novo assembly



Assemble transcripts from overlapping tags



Optional: align to genome to get exon structure

Cloonan and Grimmond. *Nat Methods*. 2010

Which alignment strategy is the best?

- Default: **alignment to a reference genome (splice-aware if relevant)**
- Alignment to transcriptome
  - If you have very short reads (<50 bps)
  - If you are short on computer resources
- *De novo* assembly
  - If a reference genome does not exist for the species you study
  - If complex polymorphisms might be missed by comparing to the reference genome
  - Longer reads are preferred

What alignment/mapping tool(s) do you know?

Limitations of BLAST in Transcriptomics

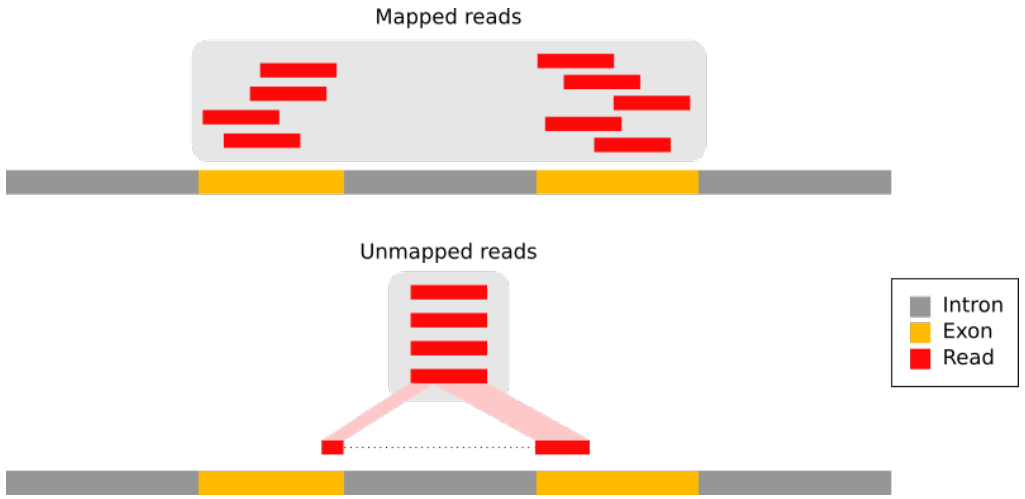
- High computational intensity when aligning millions/billions of short reads
- BLAST is more suitable for finding similarities between longer sequences, whereas transcriptomics often involves analyzing millions of short reads.

More specialized alignment tools for transcriptomics

Designed to efficiently map millions/billions of short reads to a reference genome/transcriptome

Example: **Bowtie2**

RNA-seq mapping challenges

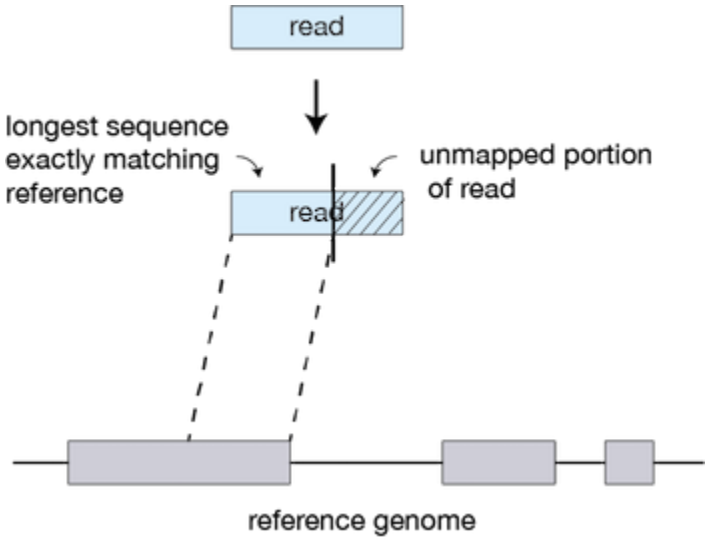


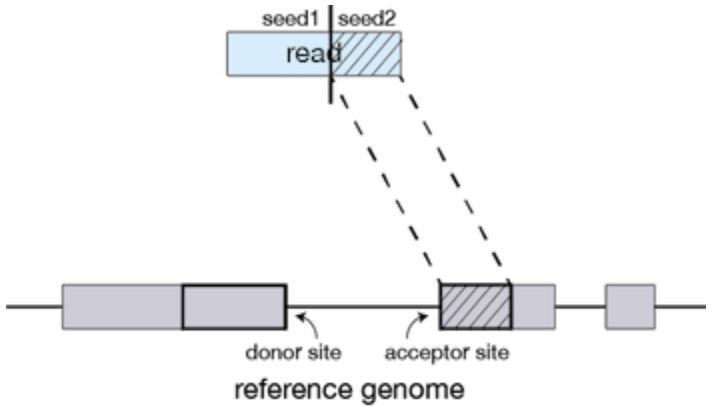
The majority of mRNA derived from eukaryotes is the result of splicing together discontinuous exons, and this creates specific challenges for the alignment of RNA-seq data.

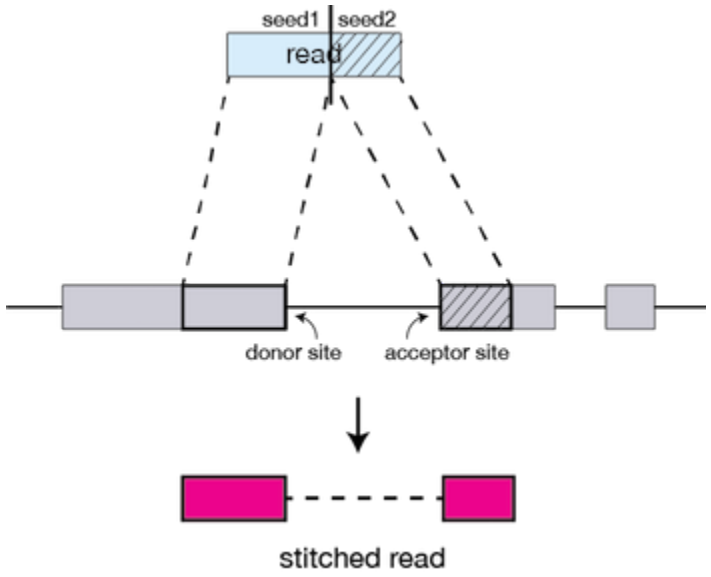
Solution: splice-aware alignment tools
Example: ***STAR***, Tophat2

https://bioinformatics.ccr.cancer.gov/docs/b4b/RNASeq_Overview/05.Alignment/

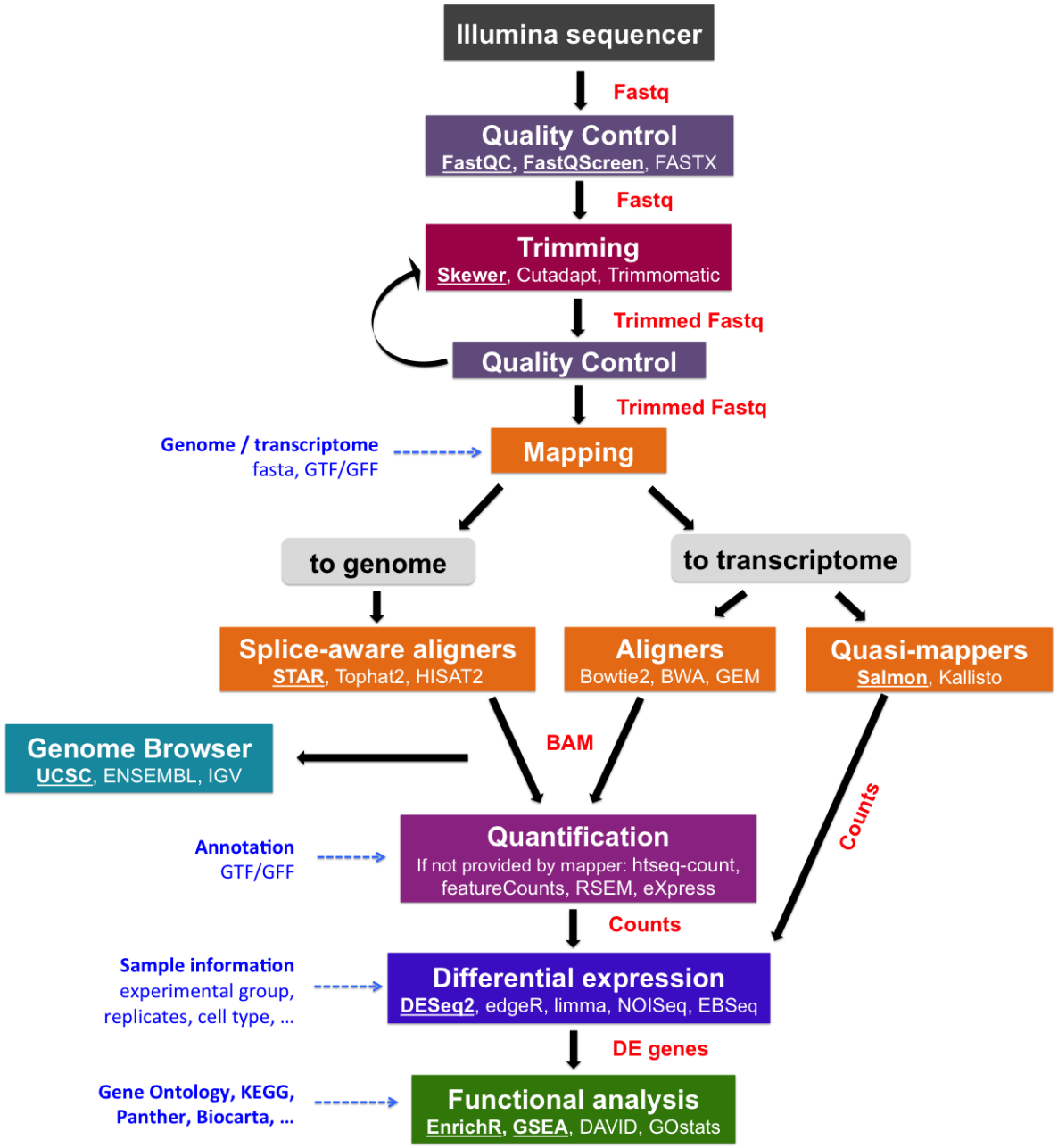
Splice-aware genome alignment: STAR





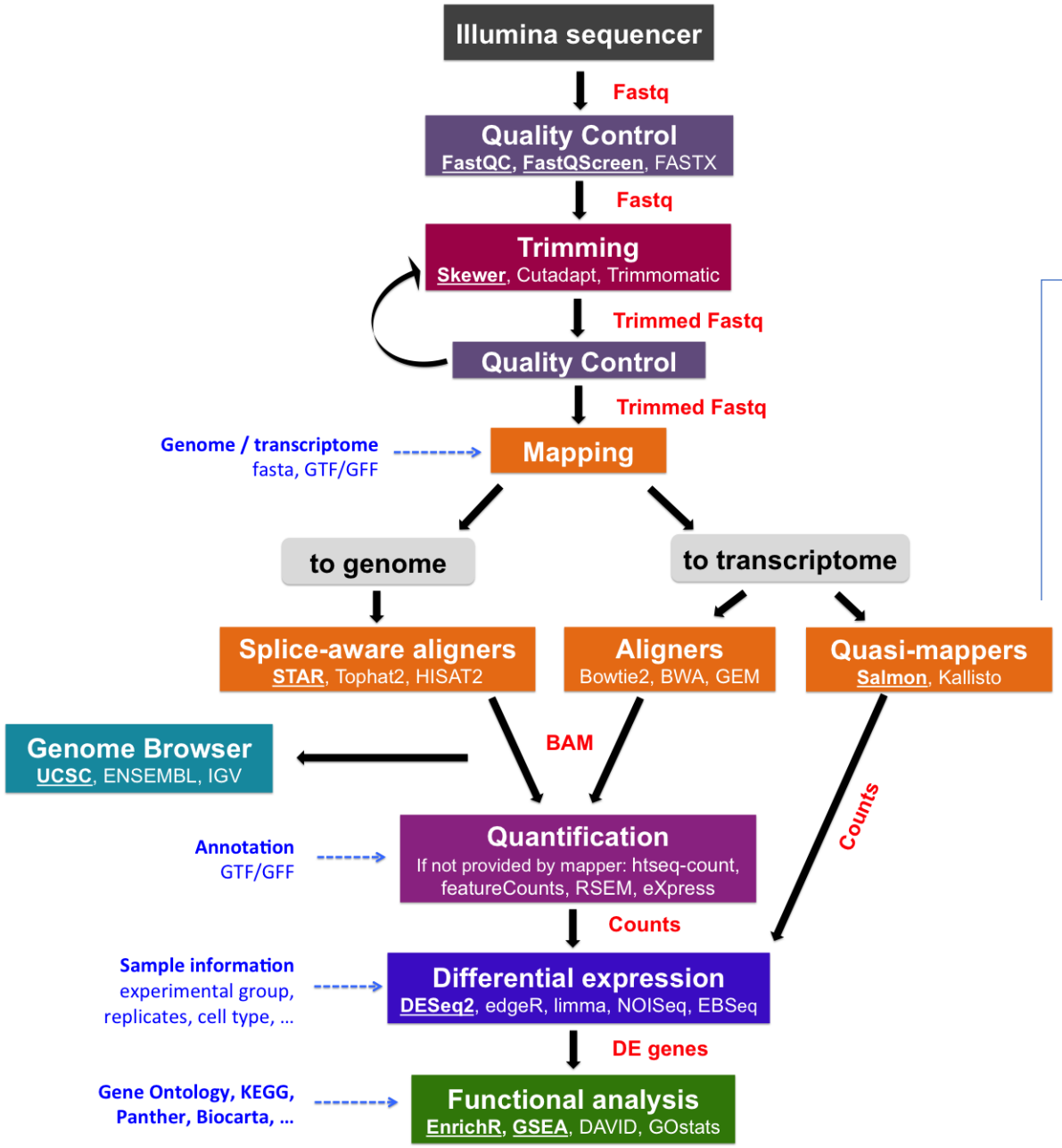


https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html



An example pipeline to process transcriptomic reads

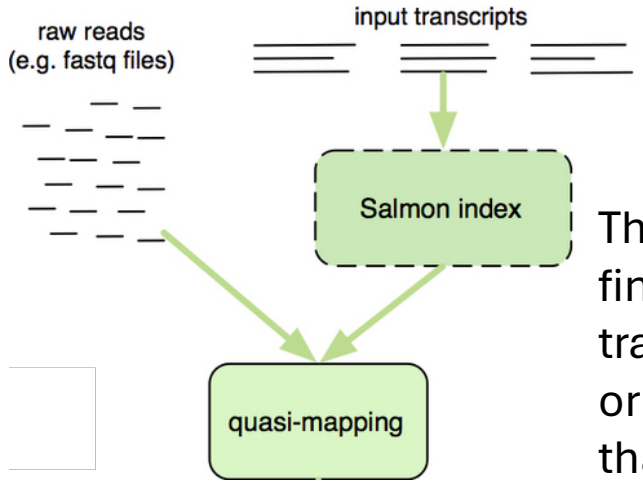
https://biocorecrg.github.io/RNAseq_course_2019/salmon.html



"Quasi-mappers"

"Wicked-fast" and lightweight

This is based on the idea that when quantifying reads, the necessary information is not **where** in a transcript a read aligns, but only **which** transcripts could have generated the read. This idea renders base-level alignment, which has been the computational bottleneck until now, unnecessary for this analysis.



They cannot be used for finding new genes, transcripts or splice forms, or for any analysis other than quantification.

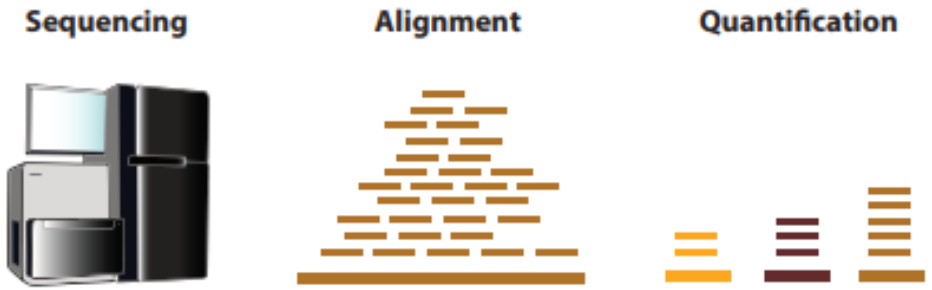
https://biocorecrg.github.io/RNAseq_course_2019/salmon.html

Recap on alignment/mapping tool examples

- BLAST
- Bowtie2: more computationally efficient in aligning short reads than BLAST
- STAR: splice-aware genome aligner
- Salmon: quasi-mapper – fast and lightweight

What you get after alignment and quantification?

count matrix
(count table, expression matrix/table)



	sample A	sample B
gene 1	50	40
gene 2	50	40
...	...	...
gene 99	50	40
gene 100	10	1000

reads

count matrix

What you get after alignment and quantification?

count matrix
(count table, expression matrix/table)

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Gene A	0	2	1	100	85	72
Gene B	43	52	36	29	72	35
…	…	…	…	…	…	…
Gene N	98	73	65	12	9	4

How many rows do you expect to see for human data?

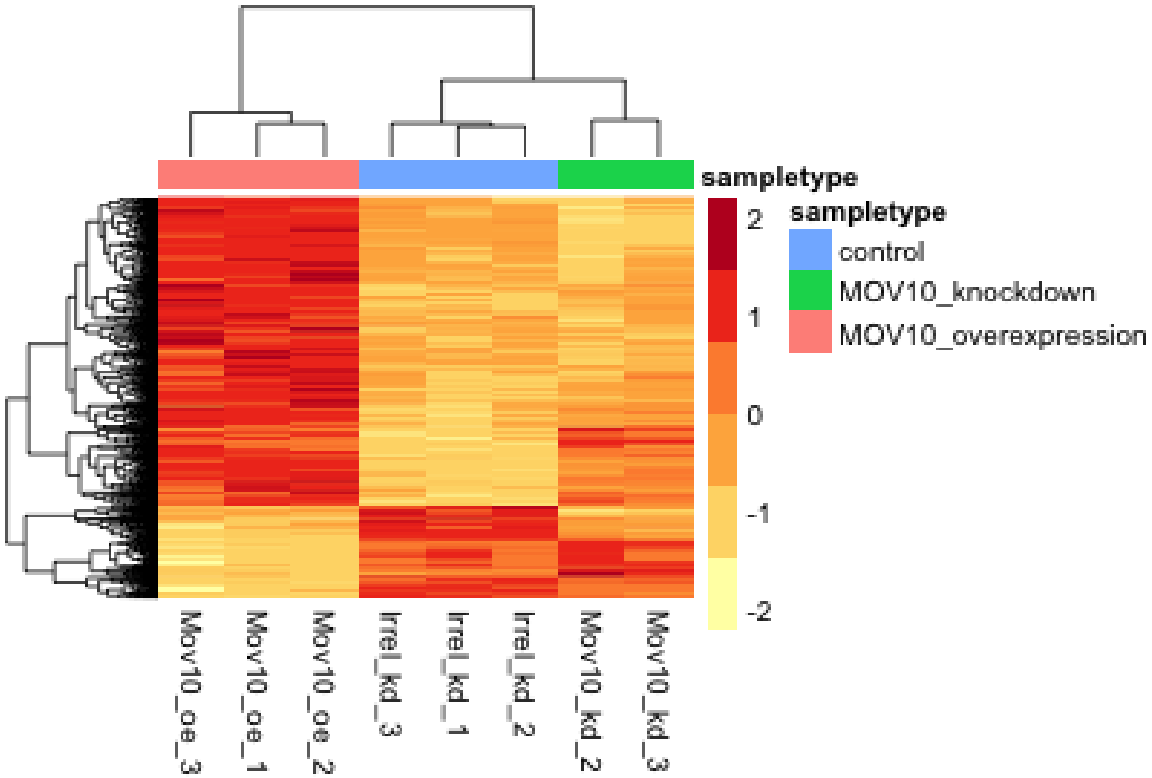
Instead of analyzing all genes, we only show genes that are "differentially expressed" between condition A vs condition B, which should explain differences between the two conditions (aka. informative genes)

	Condition A			Condition B		
	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**Sample 6**
Gene A	0	2	1	100	85	72
Gene B	43	52	36	29	72	35
…	…	…	…	…	…	…
Gene N	98	73	65	12	9	4

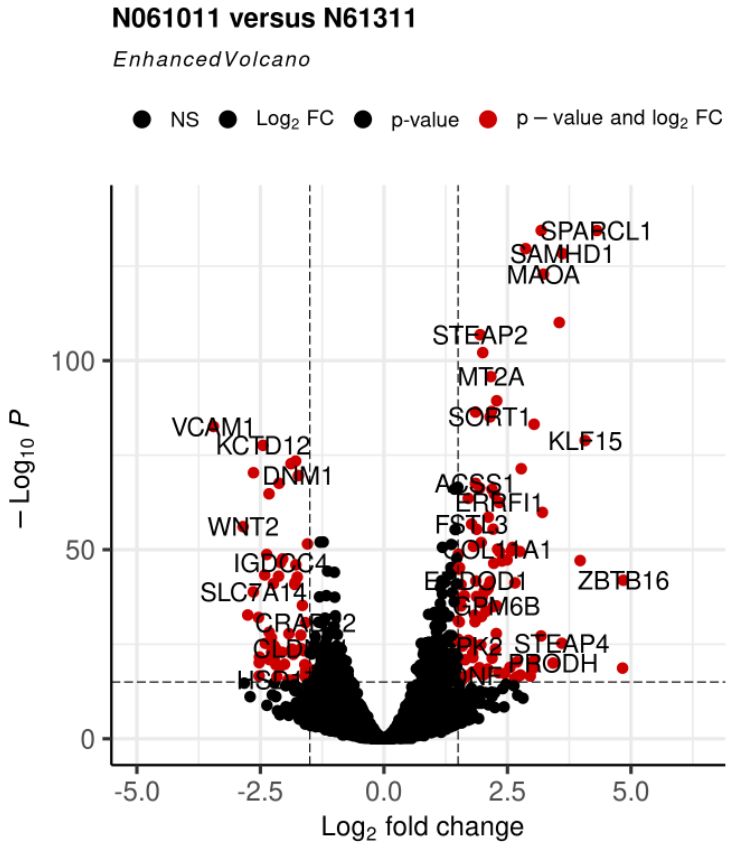
Gene A and Gene N, but not Gene B, appear to be differentially expressed
The genes that are differentially expressed are called "**DE genes**" or "**DEGs**."
The analysis is known as differential gene expression (**DGE**)

How people typically present DGE results...

Heatmap

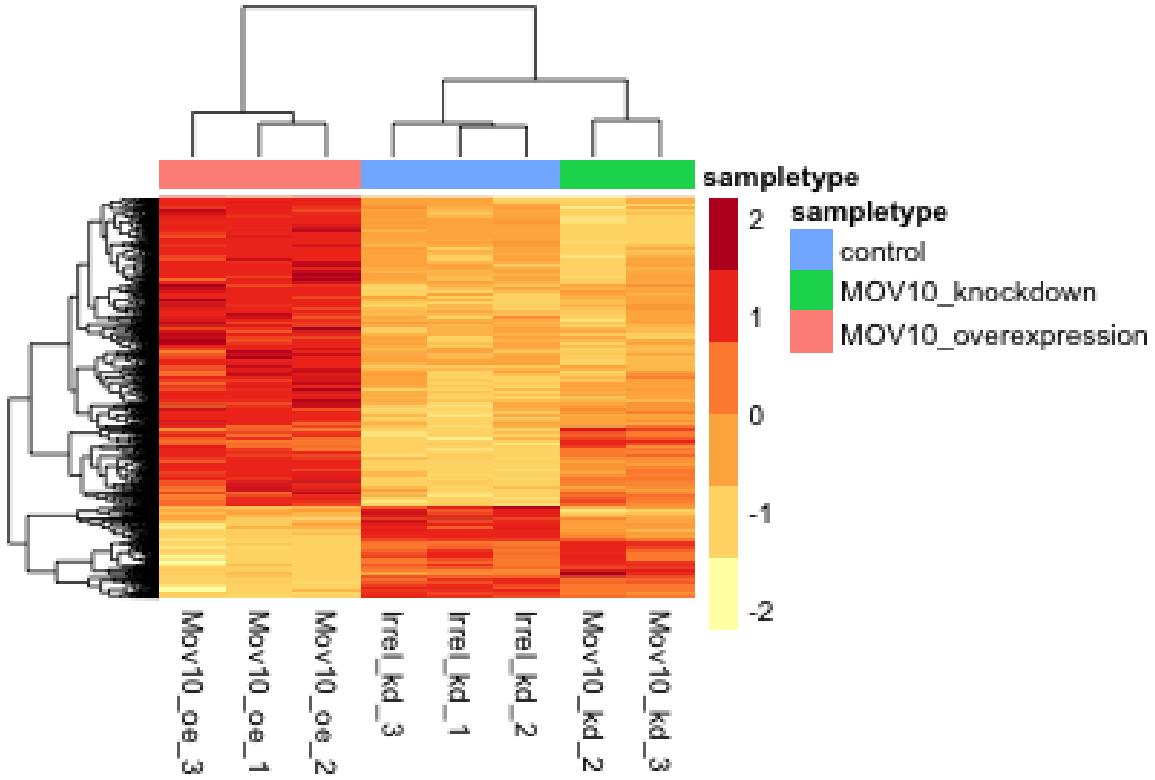


Volcano plot



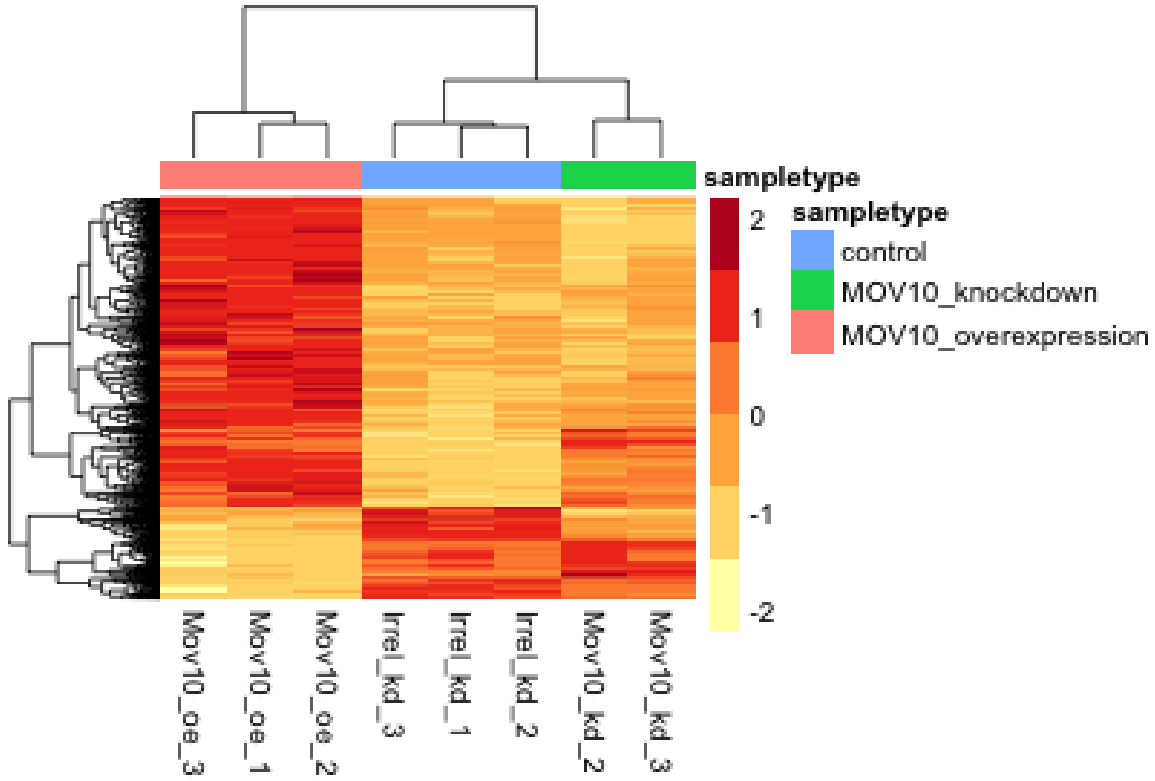
Heatmap:

- expression levels represented by colors
- genes are grouped by similar expression patterns; samples by experimental groups



- Each row is a gene
- Each column is a cell sample, typically grouped according to experimental conditions

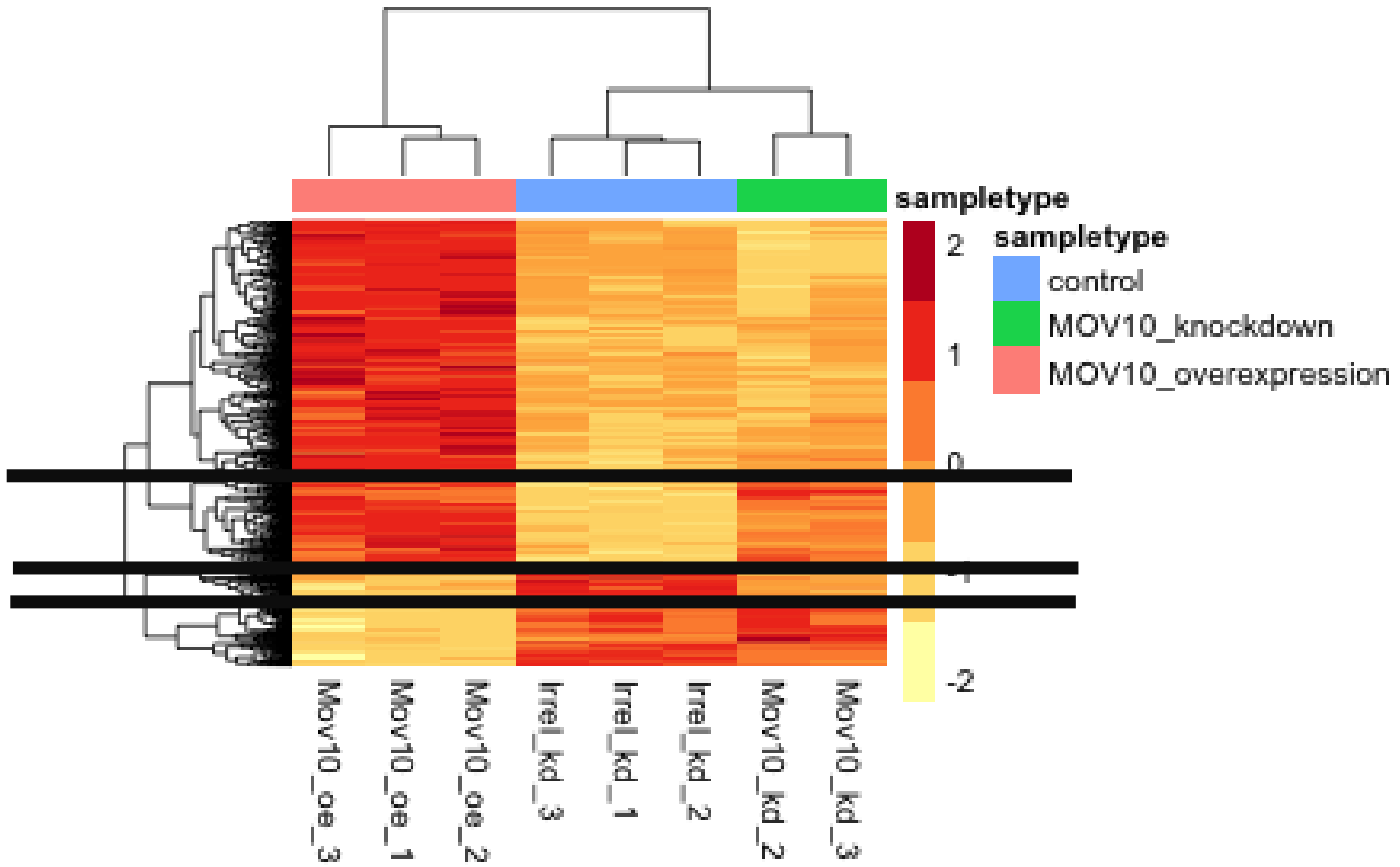
Heatmap allows you to spot gene expression patterns with respect to experimental groups



How many samples?

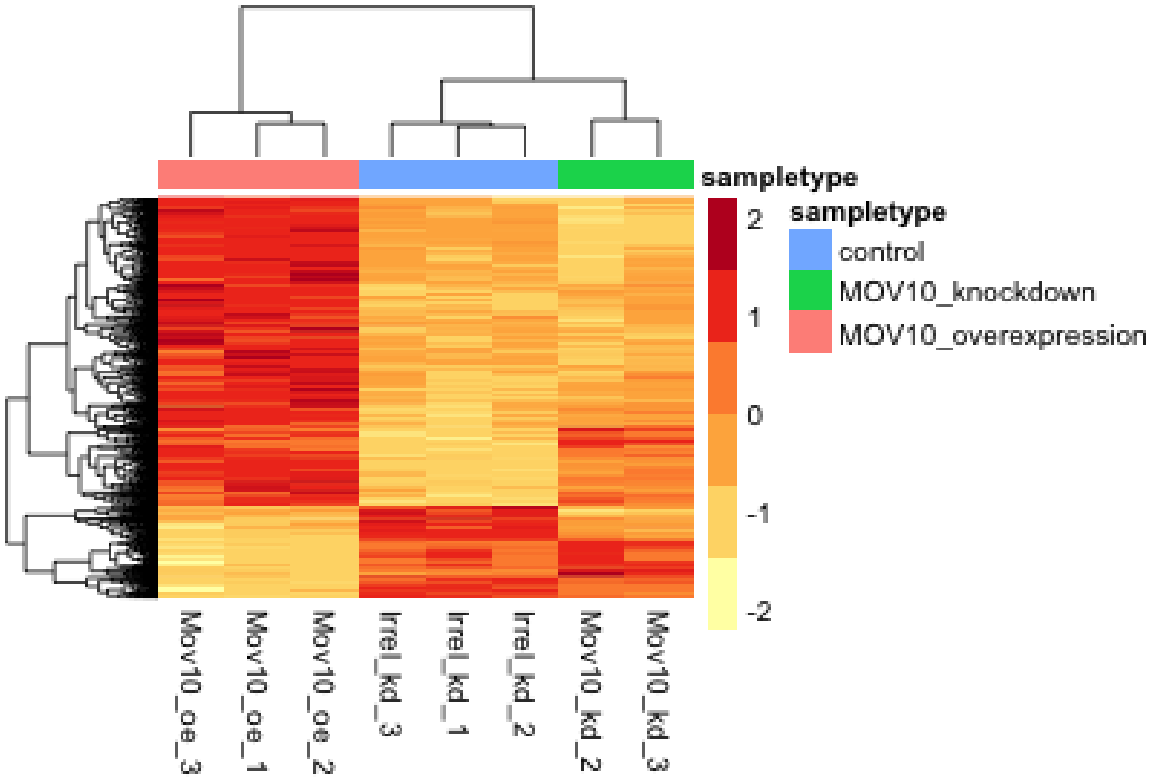
How many conditions?

How many gene expression “patterns” could you spot here?

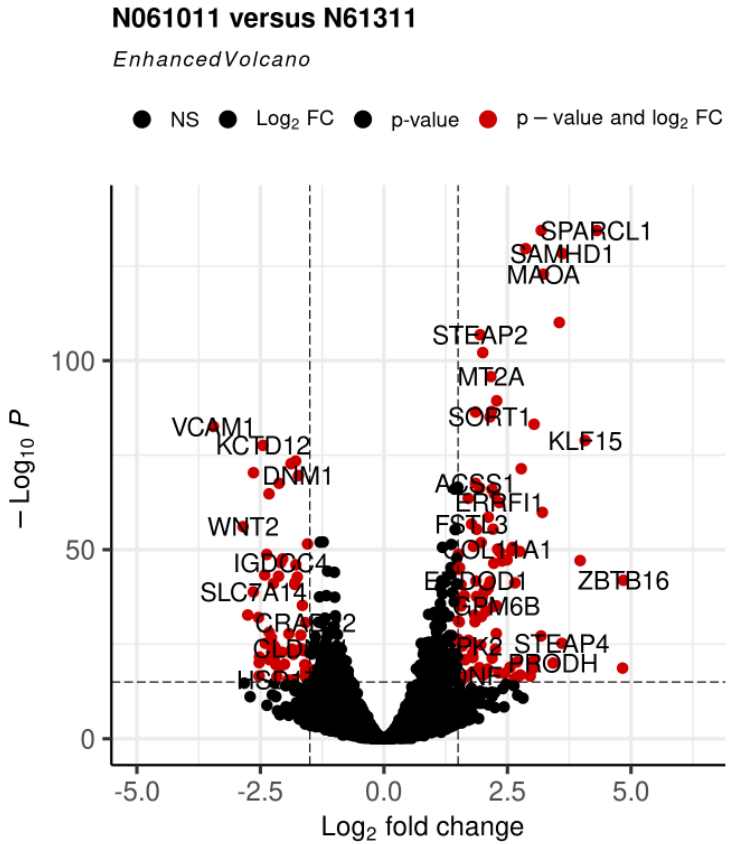

sampletype
sampletype
control
MOV10_knockdown
MOV10_overexpression
2
1
0
-2
Mov10_oe_3
Mov10_oe_1
Mov10_oe_2
Irrel_kd_3
Irrel_kd_1
Irrel_kd_2
Mov10_kd_2
Mov10_kd_3

How people typically present DGE results...

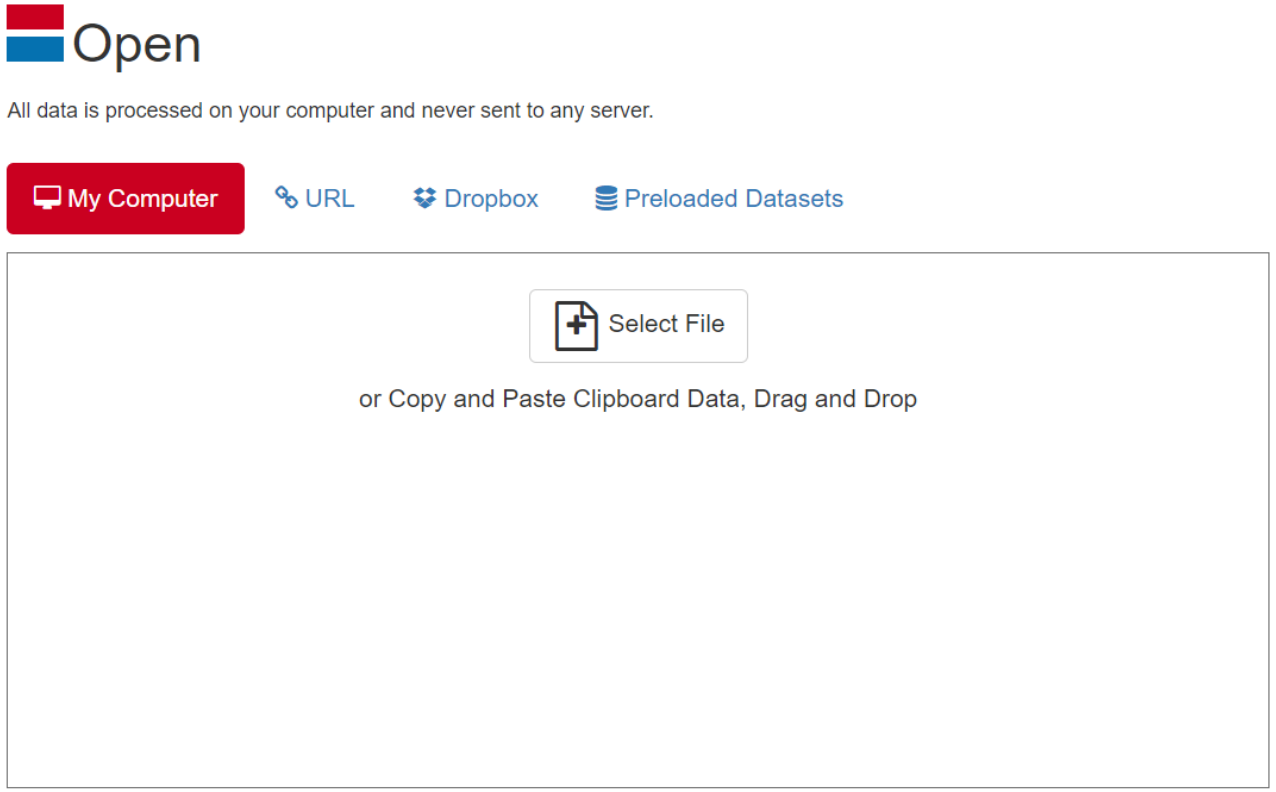
Heatmap

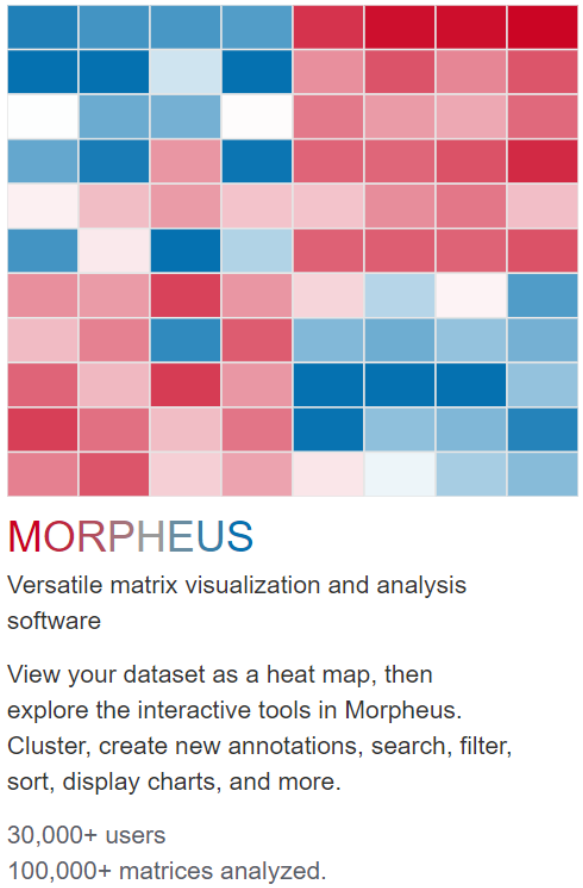


Volcano plot



If you don’t know how to write codes, you can use an online tool like Morpheus to make a heatmap





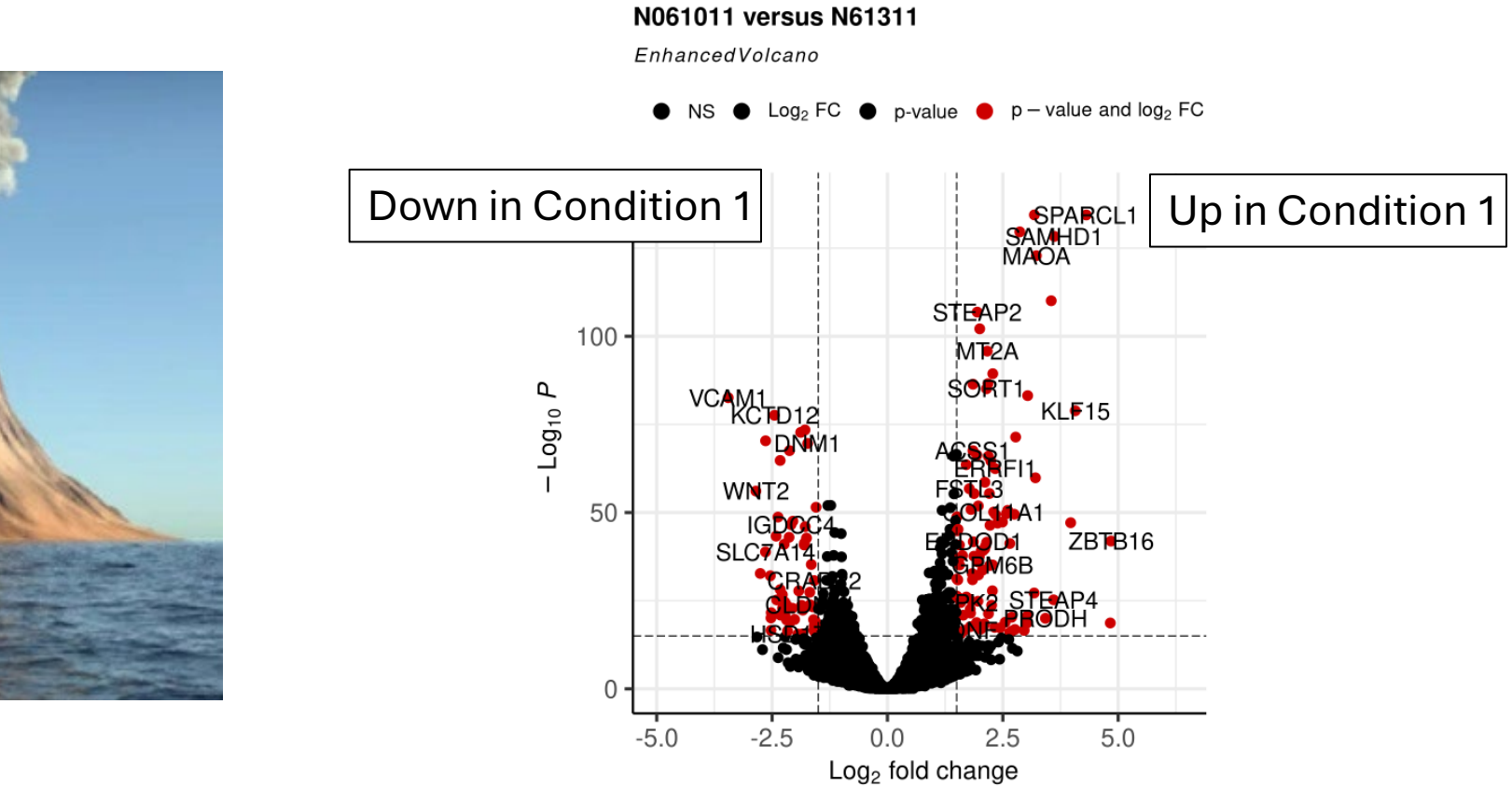
https://software.broadinstitute.org/morpheus/

Volcano plot:

- Differentially-expressed genes shown in specific portions in the plot



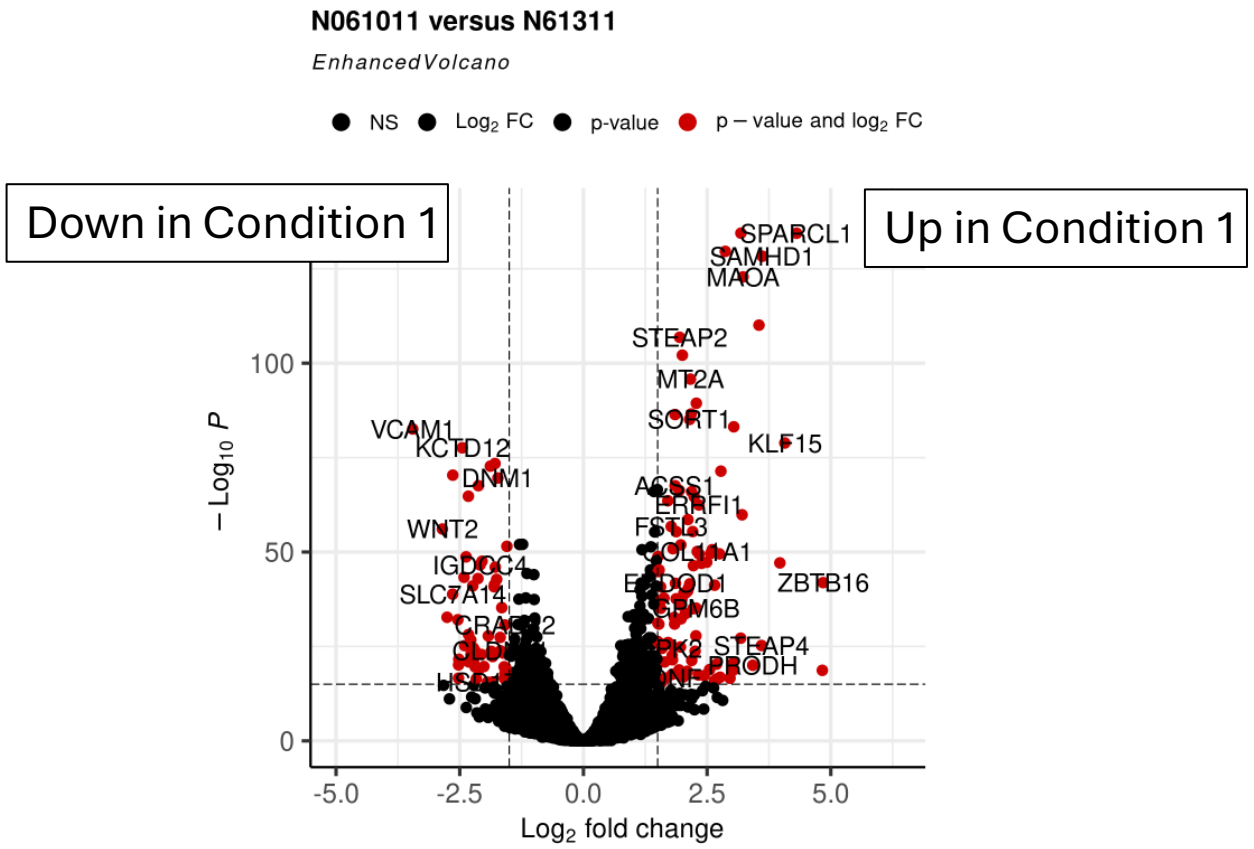
Condition 1 vs. Condition 2



Volcano plot:

- Differentially-expressed genes shown in specific portions in the plot

Condition 1 vs. Condition 2

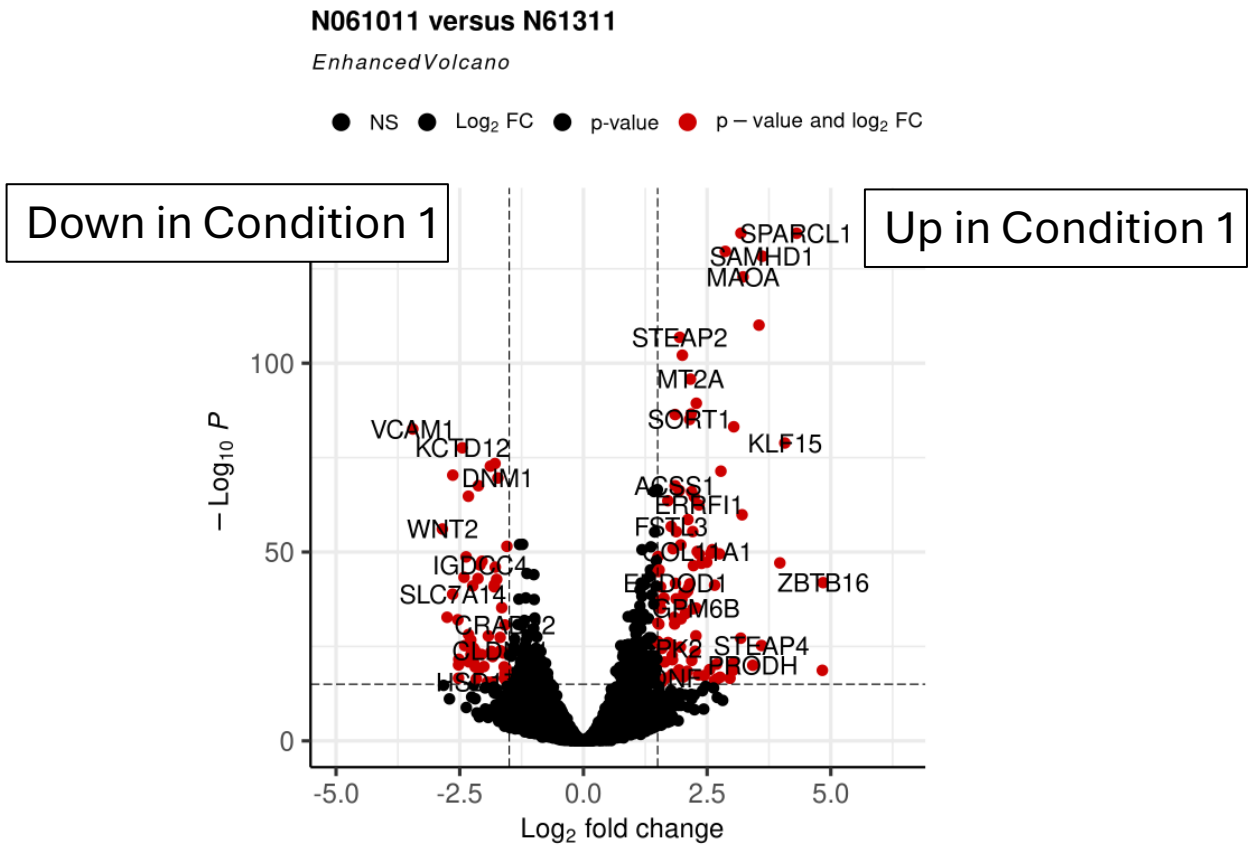


- Each data point represents a gene
- X-axis: the degree of differential expression level (Log2 fold change difference here)
- For example, if the expression of gene N in Condition 1 is 4 times more than that of Condition 2, the Log2 fold change will be $\log_2(4) = 2$
- If the expression is higher in Condition 2, we assign the fold difference in negative value

Volcano plot:

- Differentially-expressed genes shown in specific portions in the plot

Condition 1 vs. Condition 2



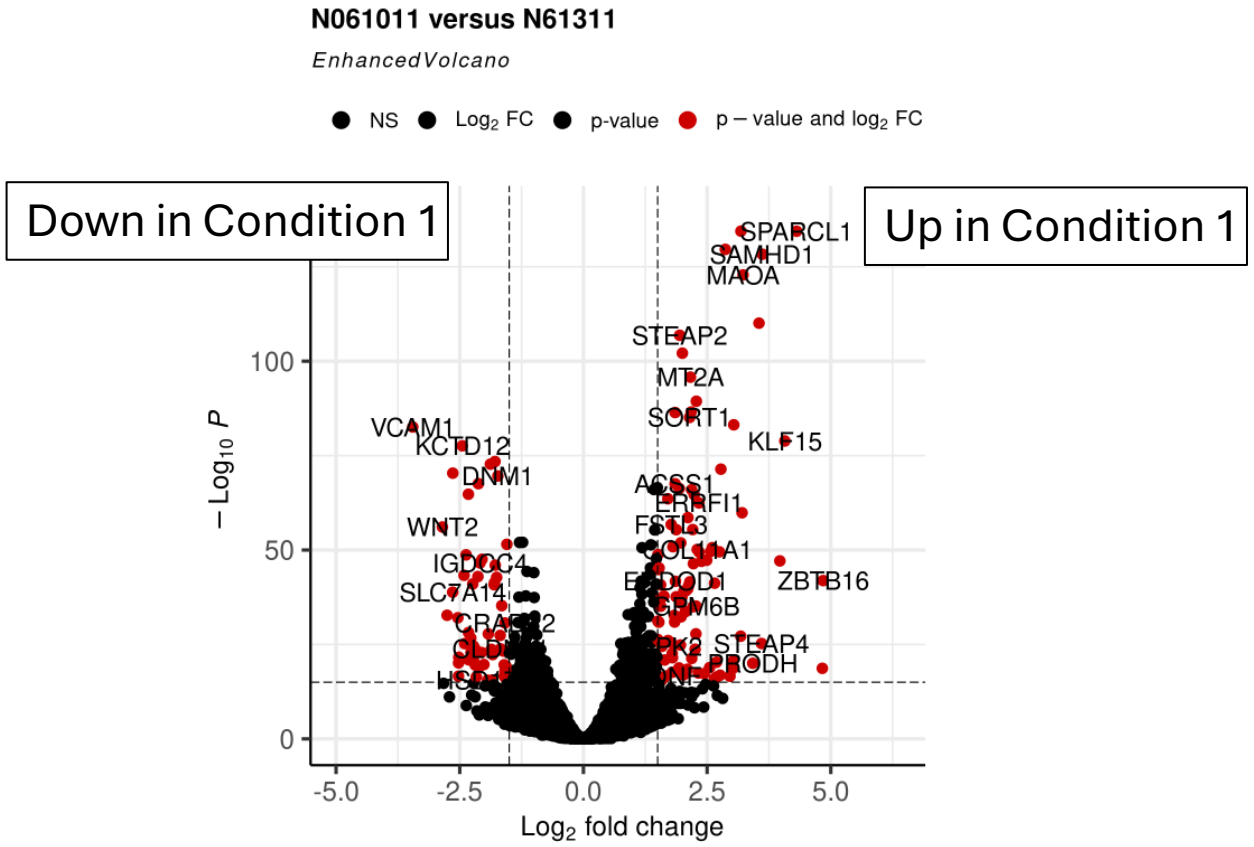
Quiz:

If the expression of gene M in Condition 2 is 8 times more than that in Condition 1, what would be the position on the “Log2 fold change” x-axis here?

Volcano plot:

- Differentially-expressed genes shown in specific portions in the plot

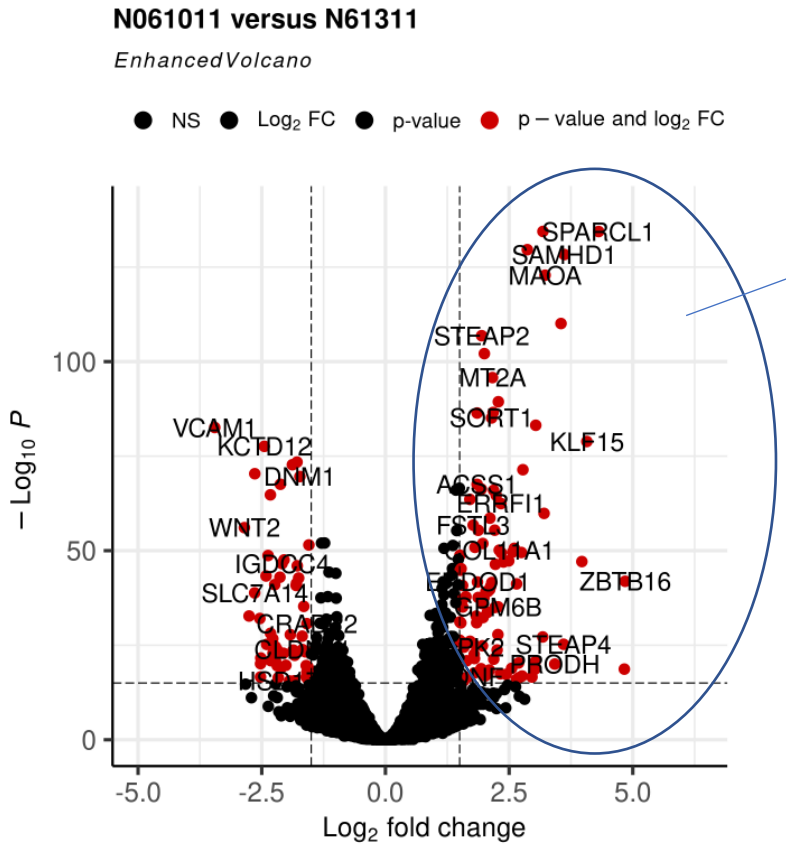
Condition 1 vs. Condition 2



- Each data point represents a gene
- Y-axis: the degree of statistical significance level (-Log10 P; more accurately, adjusted *p* values)
- Why negative? The lower the *p* values, the more statistically significant the difference will be…

Sometimes when you see DE genes, you can immediately tell what it means. However, sometimes you don't. What would you do?

Volcano plot



I have no idea about these genes

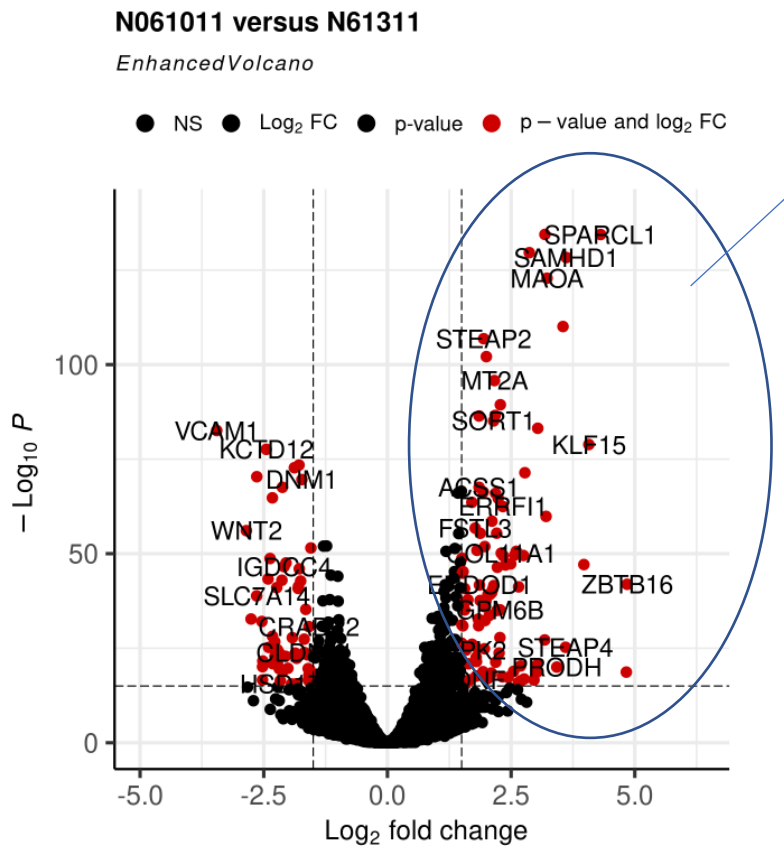
Do these genes belong to known functional pathways?

Two major approaches:

1. Gene Ontology (GO)/Over-Representation Analysis (ORA)
2. Gene Set Enrichment Analysis (GSEA)

GO/ORA Analysis

Volcano plot



Select DE genes

The analysis would investigate the fraction of those DE genes that are in a specific pathway.

If, say, 0.3% of all genes are in a pathway of interest, if we pick a list of genes randomly from all genes, we should still find ~0.3% of the genes in the list to be in that pathway.

But if 20% of the genes in the list you provided are within this pathway, it's over-represented, meaning that the genes in the list could represent enrichment of that pathway.

GO/ORA Analysis

All genes
(~30,000 genes)



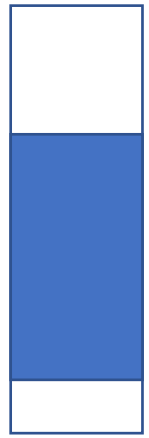
3% are known to be in pathway X

Randomly-selected list



~ 3% in pathway X

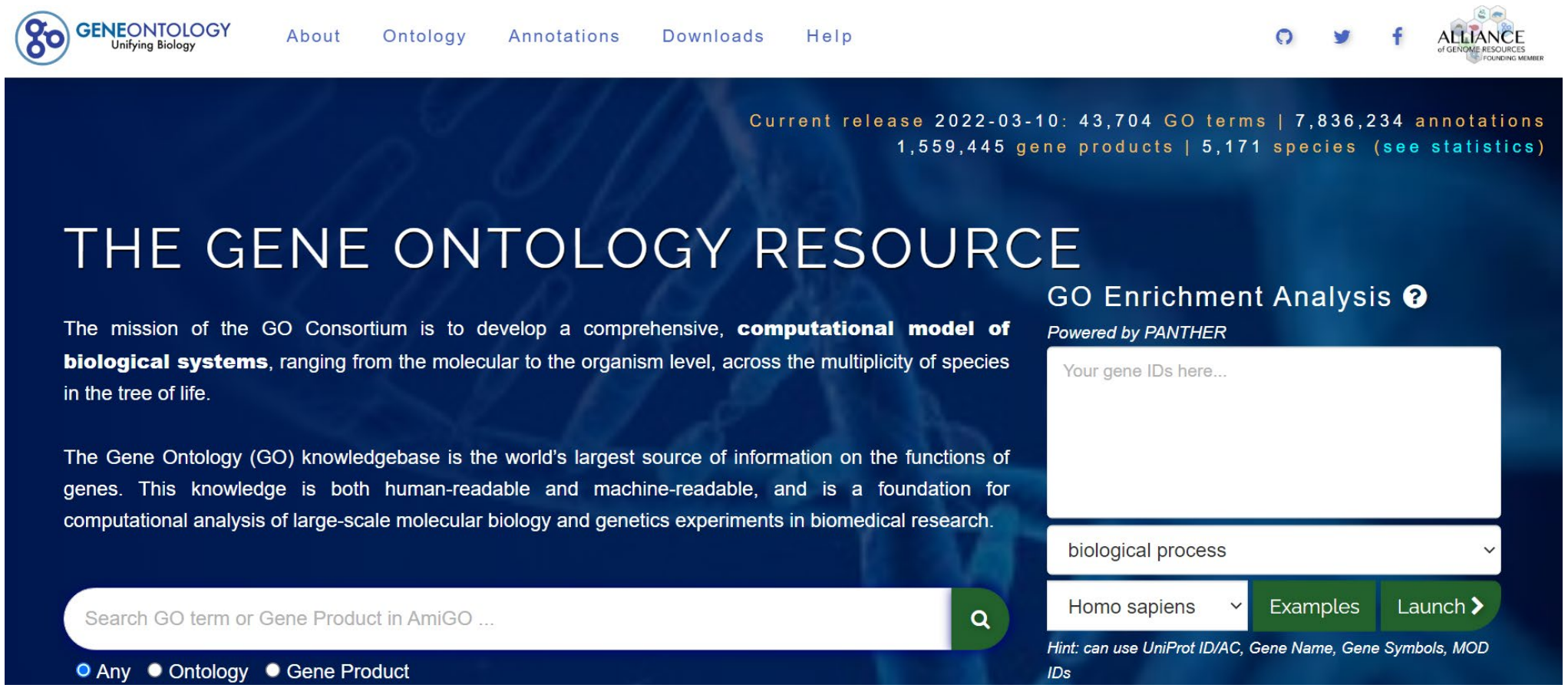
A list you get from DGE



60% in pathway X

This suggests based on the DE genes you have, pathway X should be **over-represented** so pathway X should be up-regulated in that condition

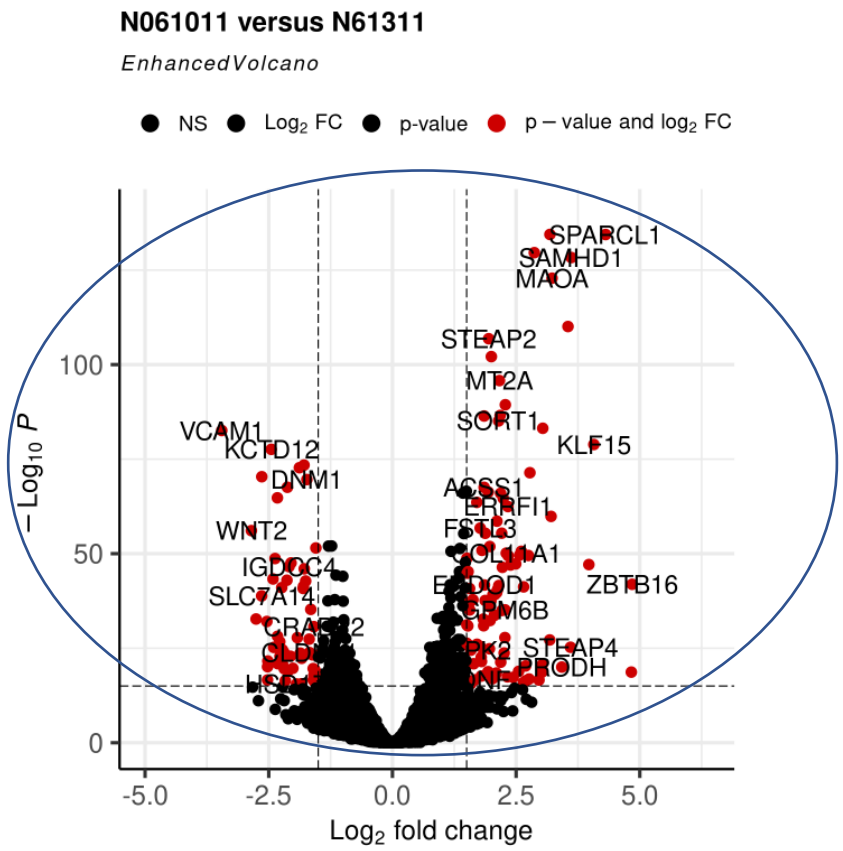
You can try running GO/ORA yourself…



http://geneontology.org/

Gene Set Enrichment Analysis (GSEA)

Volcano plot



Use all genes that are RANKED!

You will rank them and weigh them by some statistic (fold change, or log(p) for example). The software will walk through from top to bottom, investigating a pathway. If it runs into a gene in the pathway it'll boost your enrichment score, if it doesn't it'll subtract some points. The number of points it gives depends on the rank/weight, such that the higher the rank and fold change, the bigger the boost in enrichment score.

Gene Set Enrichment Analysis (GSEA)

Up in Condition 1

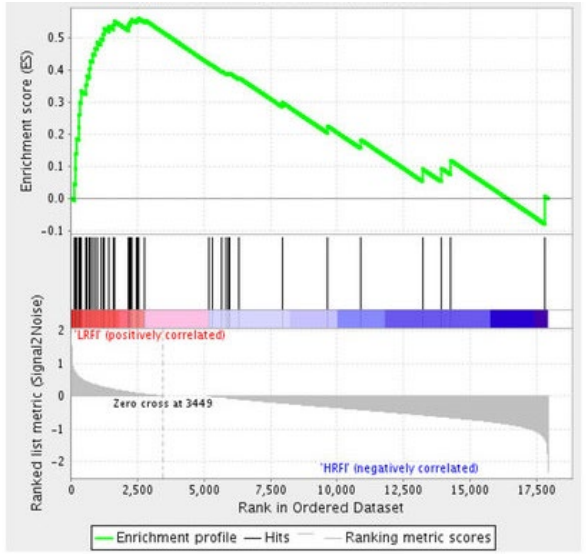
Up in Condition 2

A ranked list
(from DGE)

A
B
C
D
E
.
.
.
X
Y
Z

Pathway alpha
Gene set

B
C
D
F
H



- Walk through the ranked list; added points if the gene is in the gene set
- **Lots of points added at the beginning**
- Points deducted toward the end
- Pathway alpha likely enriched in Condition 1 (Upward peak toward the left)

Gene Set Enrichment Analysis (GSEA)

A ranked list
(from DGE)

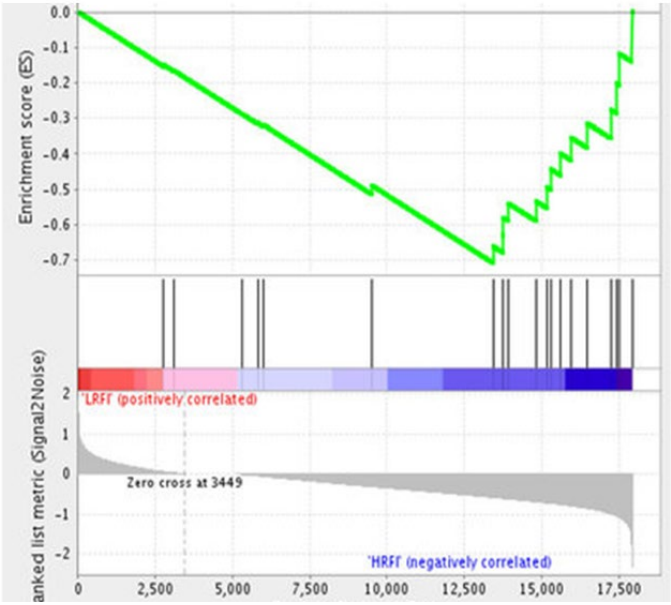
Up in Condition 1

A
B
C
D
E
.
.
.
X
Y
Z

Up in Condition 2

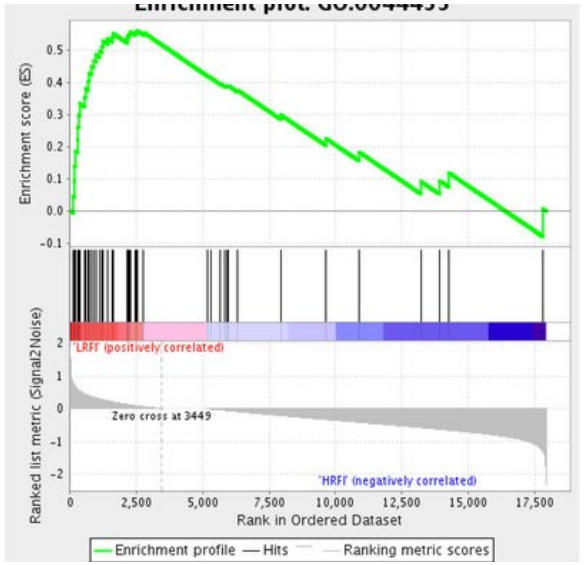
Pathway beta
Gene set

Q
V
W
Y
Z

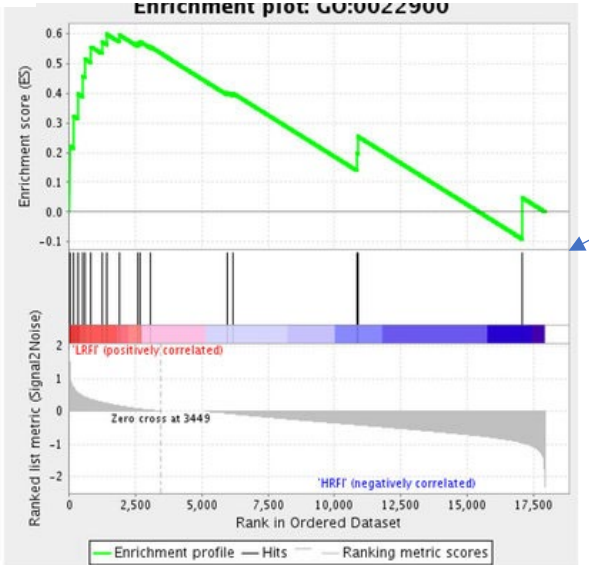


- Walk through the ranked list; added points if the gene is in the gene set
- Points deducted at the beginning
- **Lots of points added toward the end**
- Pathway beta enriched in Condition 2 (Downward peak toward the right)

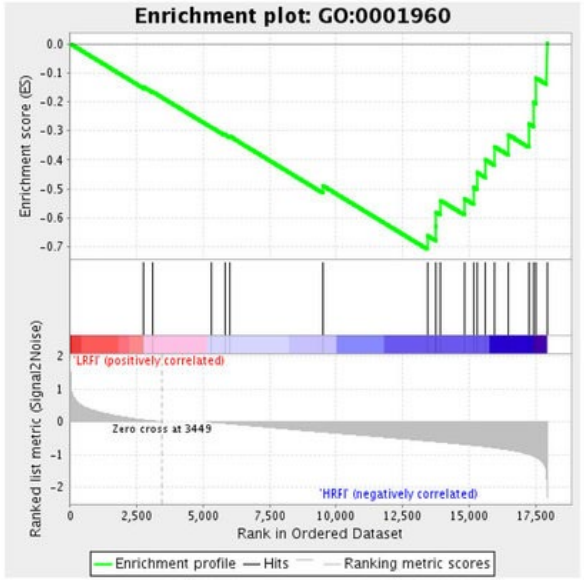
MITOCHONDRIAL_MEMBRANE_PART

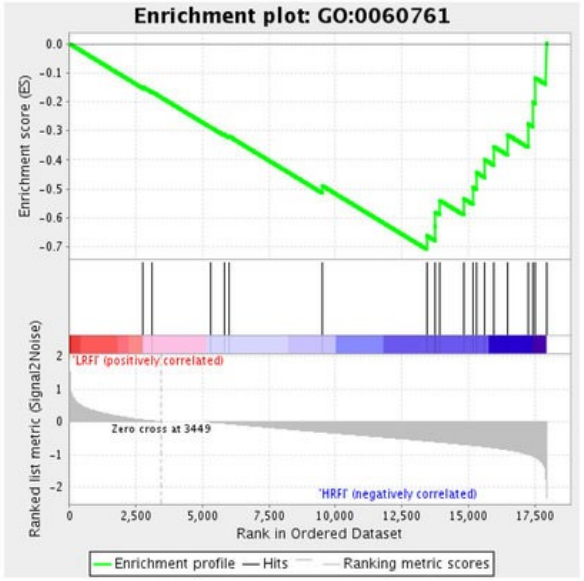


ELECTRON_TRANSPORT_CHAIN



NEGATIVE_REGULATION_OF_RESPONSE_TO_CYTOKINE_STIMULUS





Genes in the gene set shown here;
Showing where they are in the ranked list

GSEA Enrichment Plots

Yang et al. *BMC Genomics*. 2020

The leading edge subset is the set of gene members in the gene set that appear in the ranked list prior to the peak score → genes in the gene set that contribute to the enrichment the most!

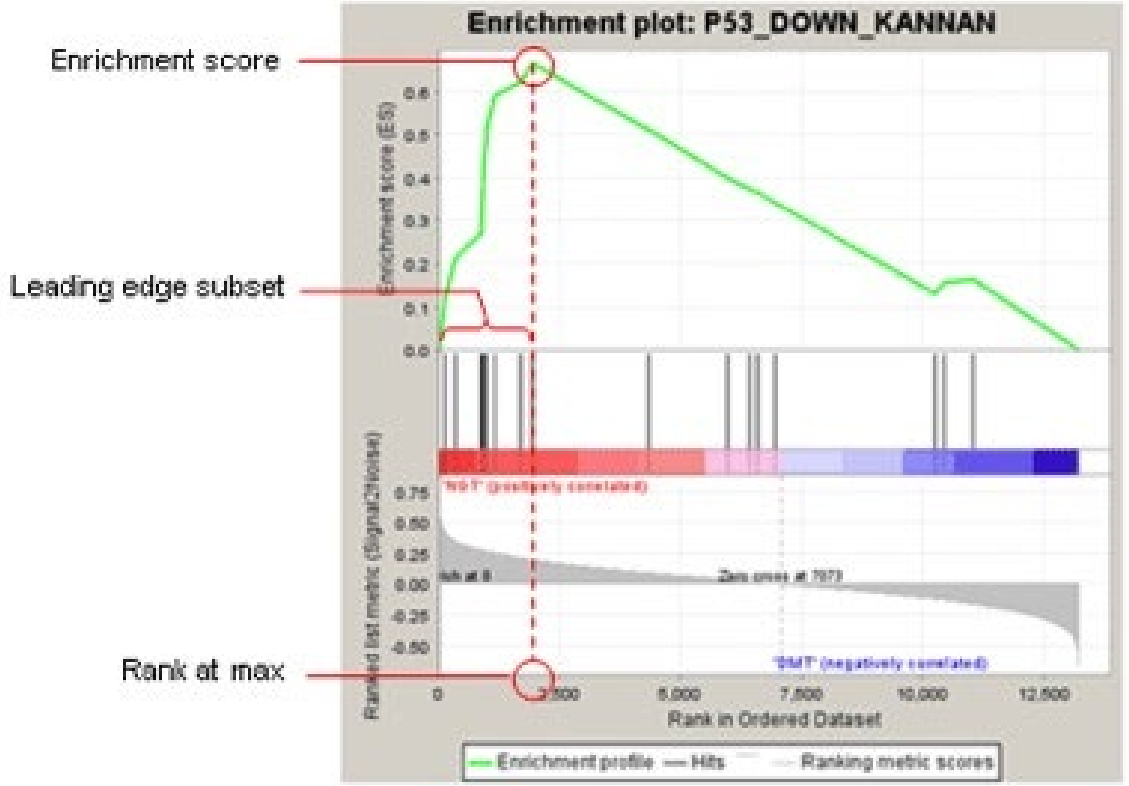
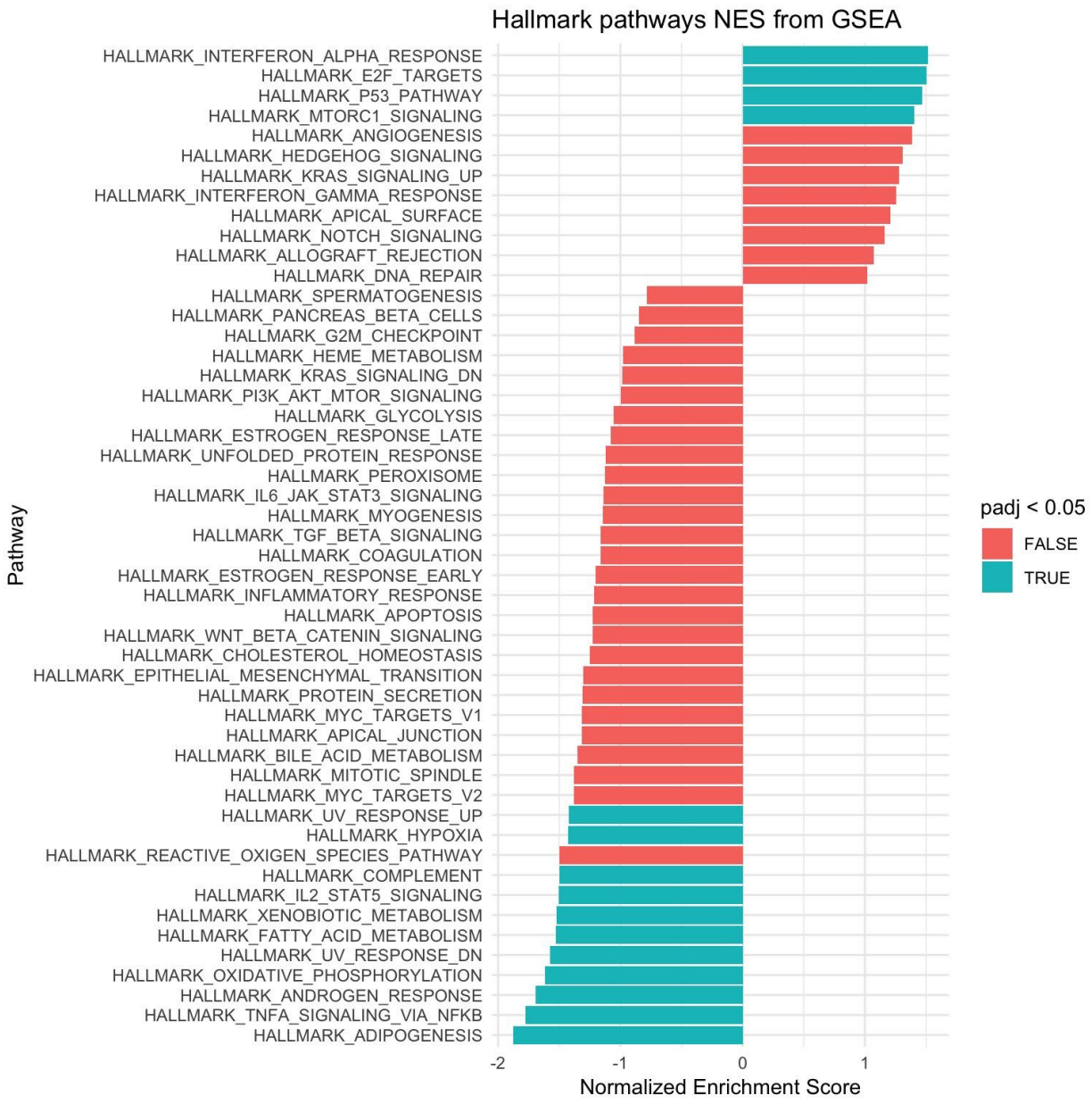


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

$$NES = \frac{\text{actual ES}}{mean(\text{ESs against all permutations of the dataset})}$$



GSEA-identified pathways

https://stephenturner.github.io/deseq-to-fgsea/

What if you'd like to sequence bacterial/viral mRNA?

Do you think things are going to be easier? Or more difficult?

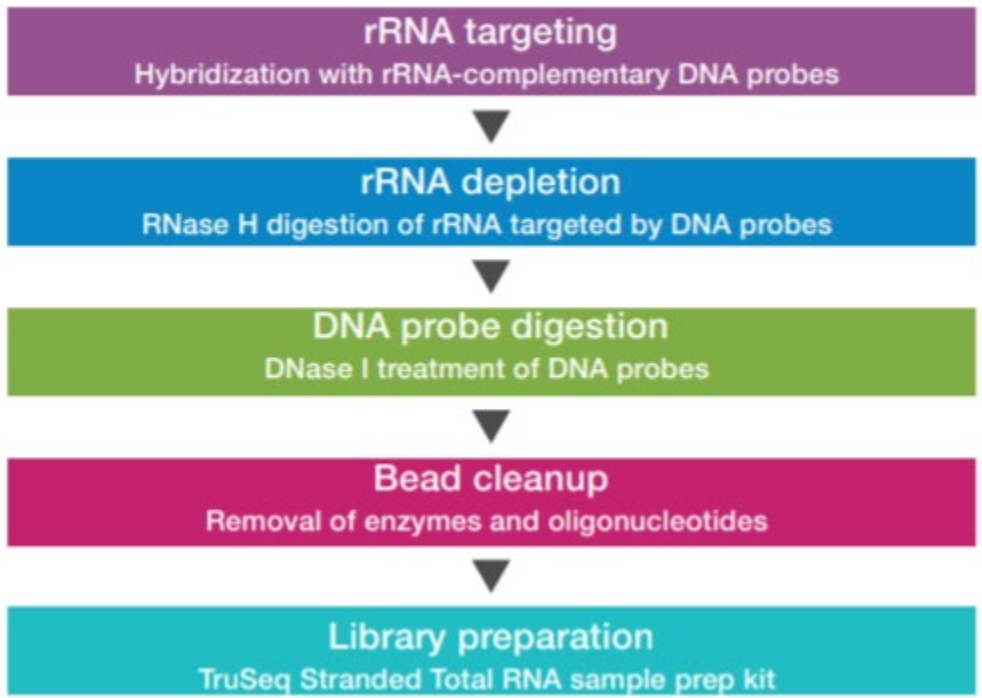
Challenges in sequencing bacterial/viral mRNA

- In prokaryotes, poly(A) tails are present only in 2-60% of the molecules of a given mRNA species, making poly(A) tails, typical targets for RT primers, not suitable. For some viruses, they do not use polyadenylation at all.
- >95% of cellular RNA is composed of ribosomal RNA

illumina®

Microbial RNA sequencing enabled with the Illumina Ribo-Zero™ Plus rRNA Depletion Kit

Focus on analyzing high-value portions of the transcriptome and save on sequencing costs.

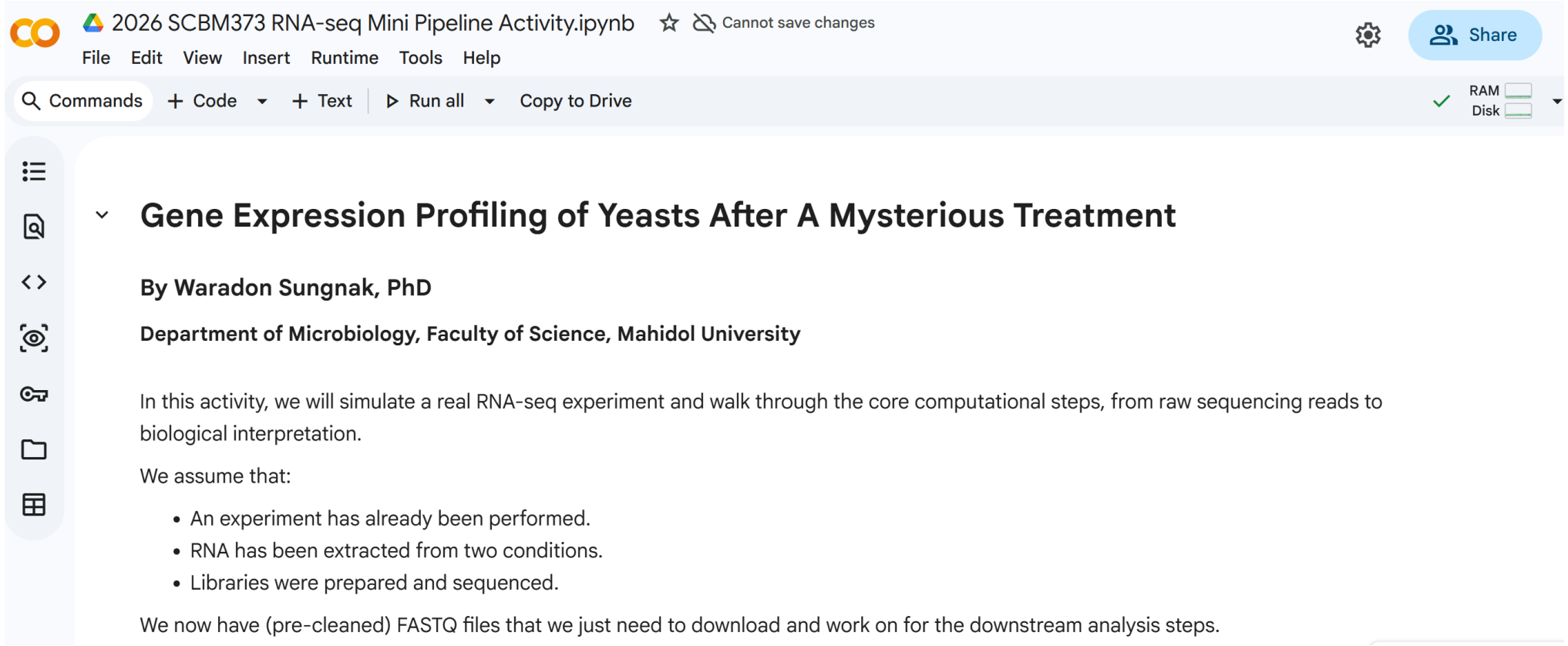


Take-home messages

- Cellular phenotype can be reflected by gene expression i.e. mRNA expression
- Transcriptomics – comprehensive info on gene expression
- **Transcriptomics ~ RNA-seq**
- From RNA-seq, you get **reads** from a sequencer (most likely Illumina) you would then need to **align/map** them to **reference** sequences
- **Differential gene expression (DGE)** tells you how different gene expression is in 2 different sample groups
- Functional **gene ontology (GO)** or **gene set enrichment analysis (GSEA)** tells you what DE genes might represent pathway-wise
- Special considerations needed for bacterial cells or viral RNA

Class activity:

Gene Expression Profiling of Yeasts After A Mysterious Treatment



https://colab.research.google.com/drive/14CqHGrdMpOP91UC4uJmDAJyPfHdXm_K9?usp=sharing