

Introduction to RNA-Sequencing

By Aya Tarek & Eshraq Saeed

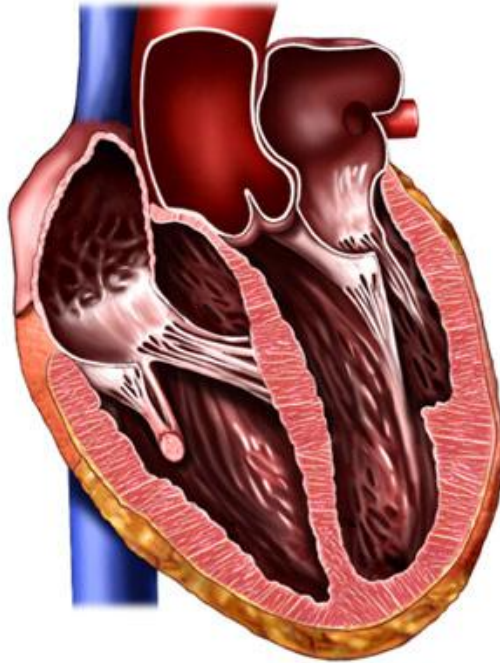
Agenda

- How to study a disease?
- Next generation sequencing
 - Microarrays
- RNA-Sequencing
 - Workflow
 - Guidelines
 - Challenges
 - Applications

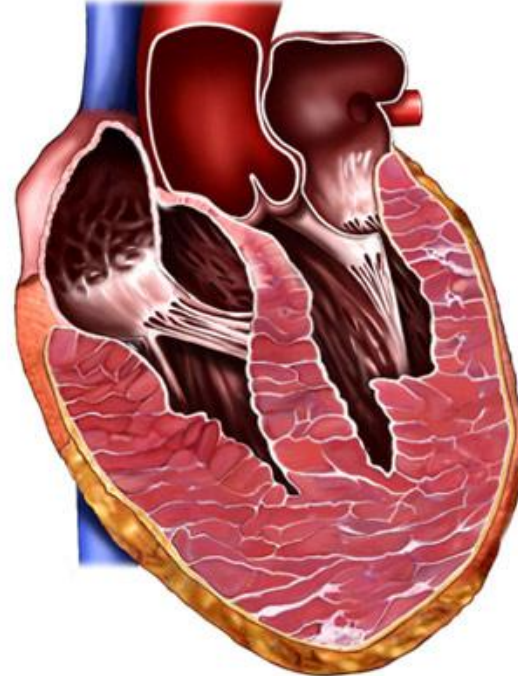


Cardiovascular Diseases

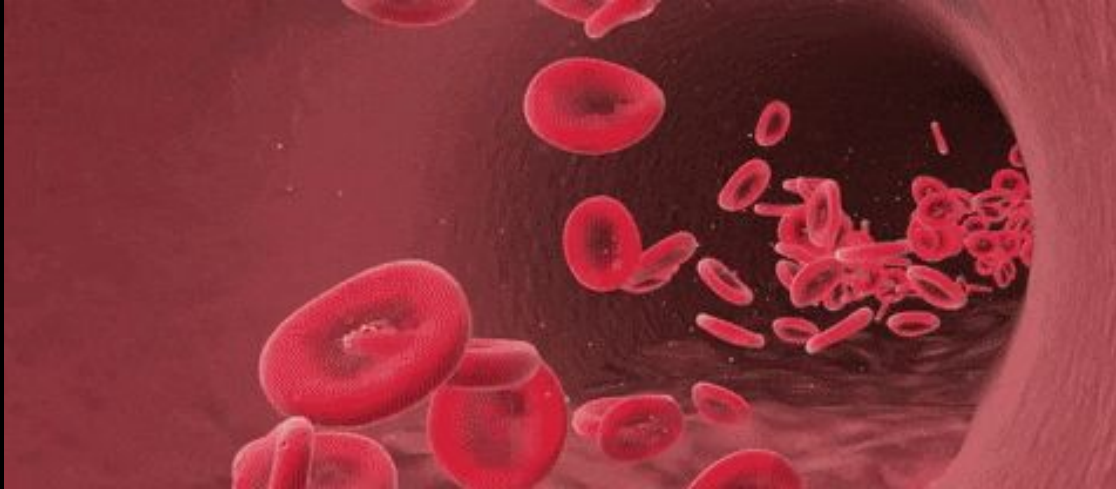
- Hypertrophic cardiomyopathy

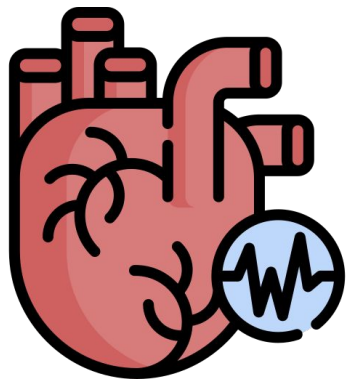


Normal heart
(cut section)

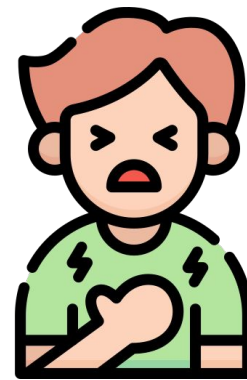
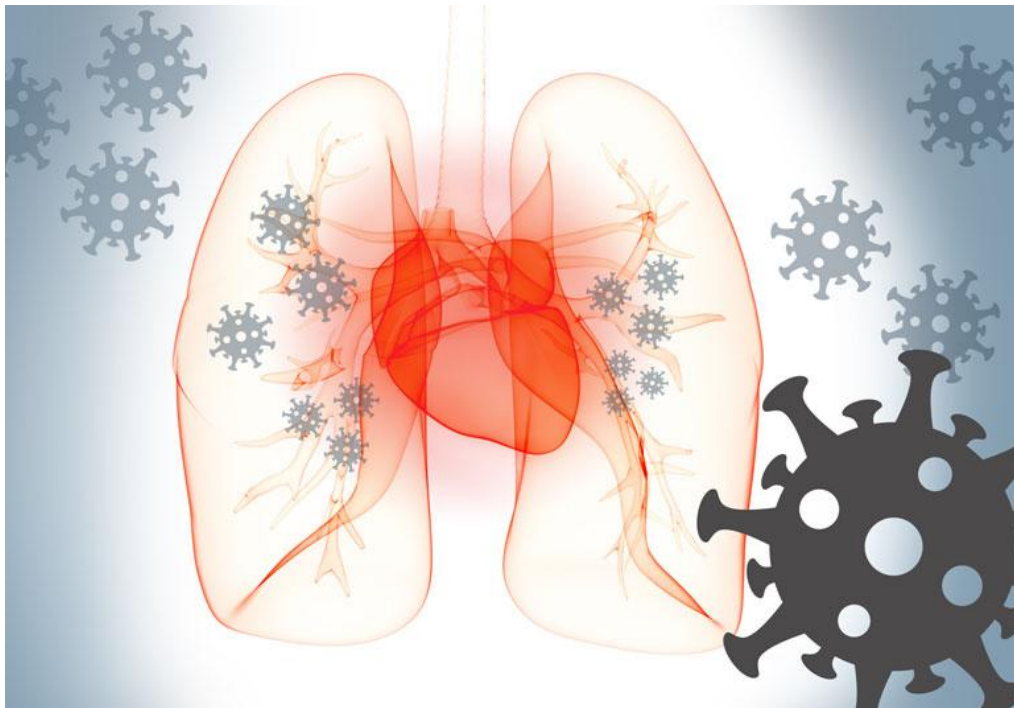


Hypertrophic
cardiomyopathy

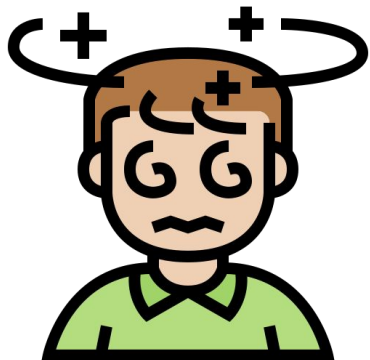




Arrhythmias



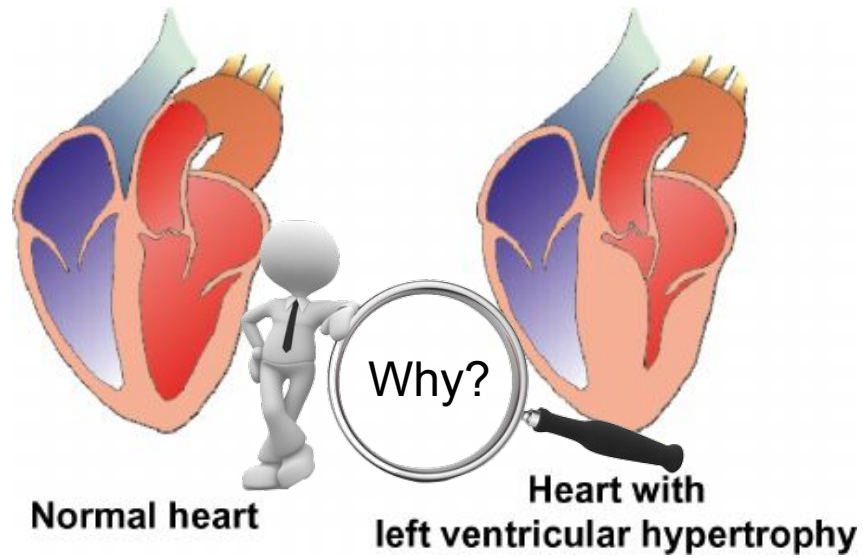
Chest pain

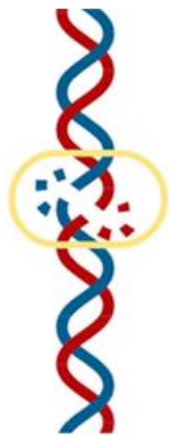


Dizziness

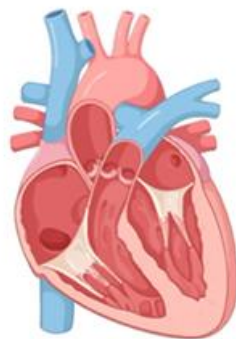


*Swelling in
legs*





=



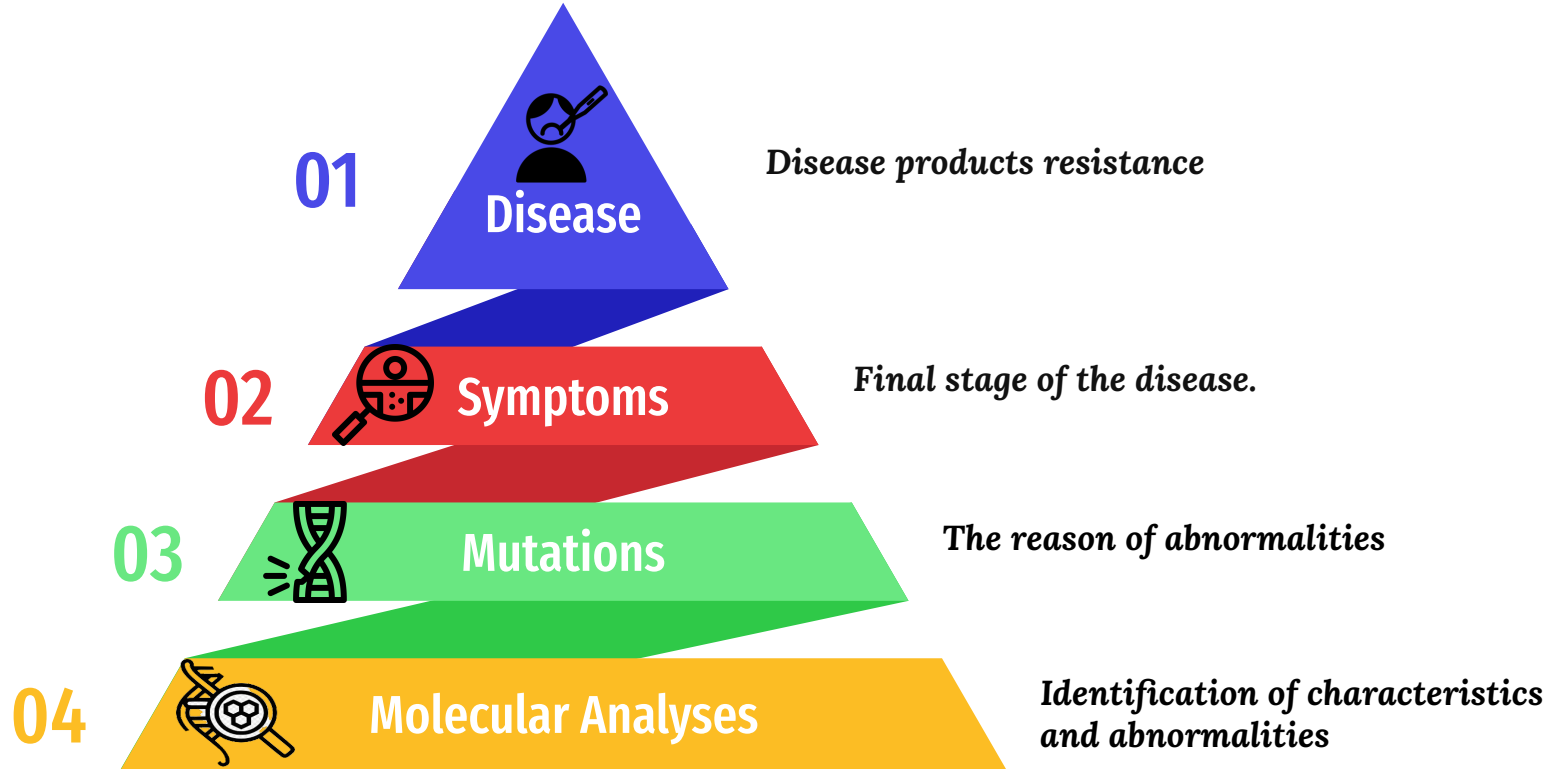
Healthy

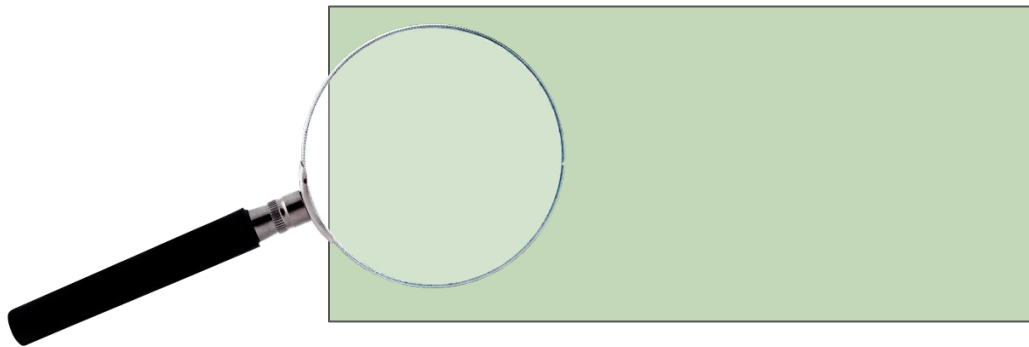


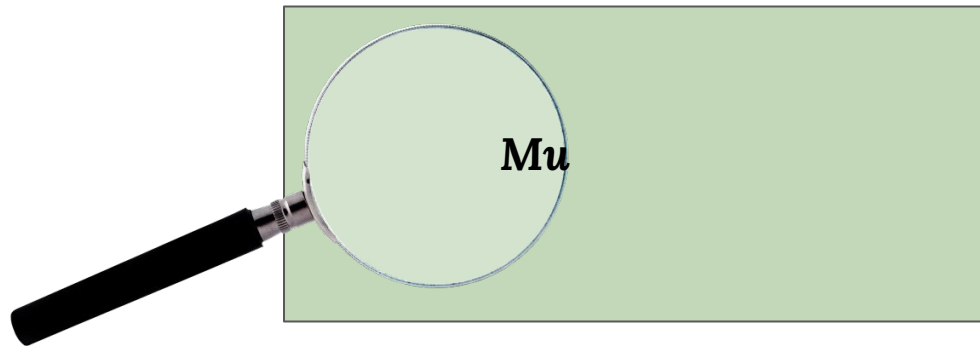
Hypertrophy

& many more...

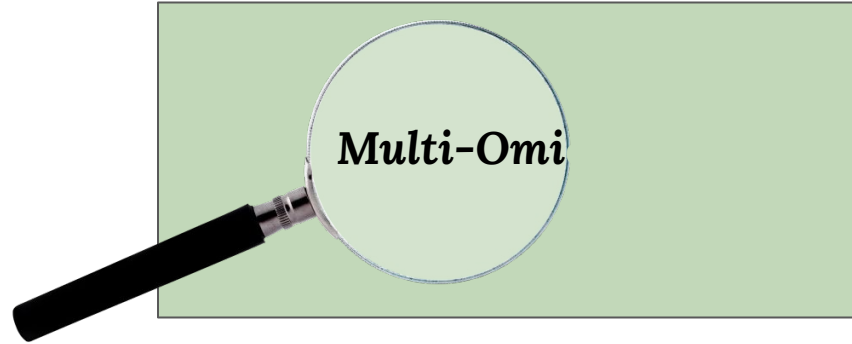
- *Cancer diseases*
- *Infectious diseases*
- *Rare diseases*

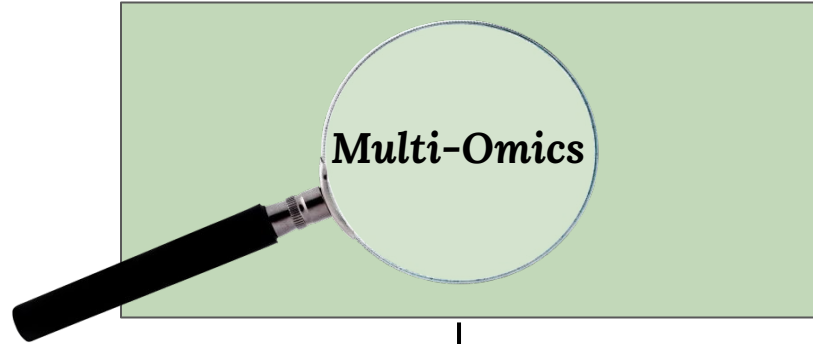






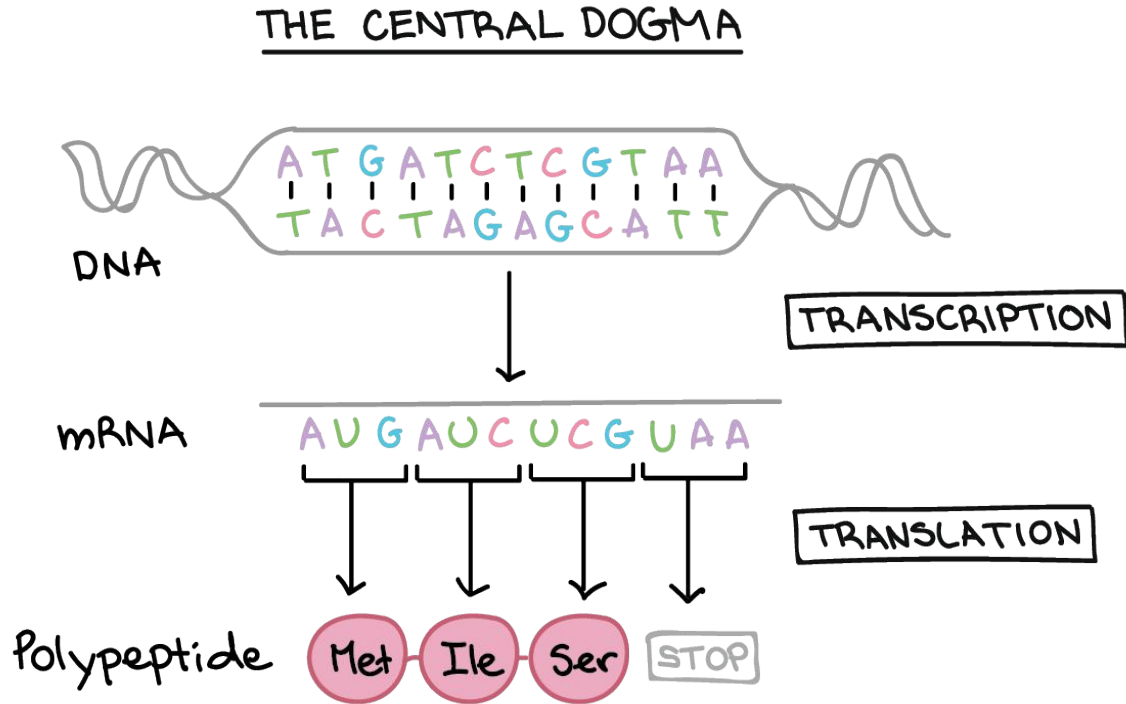




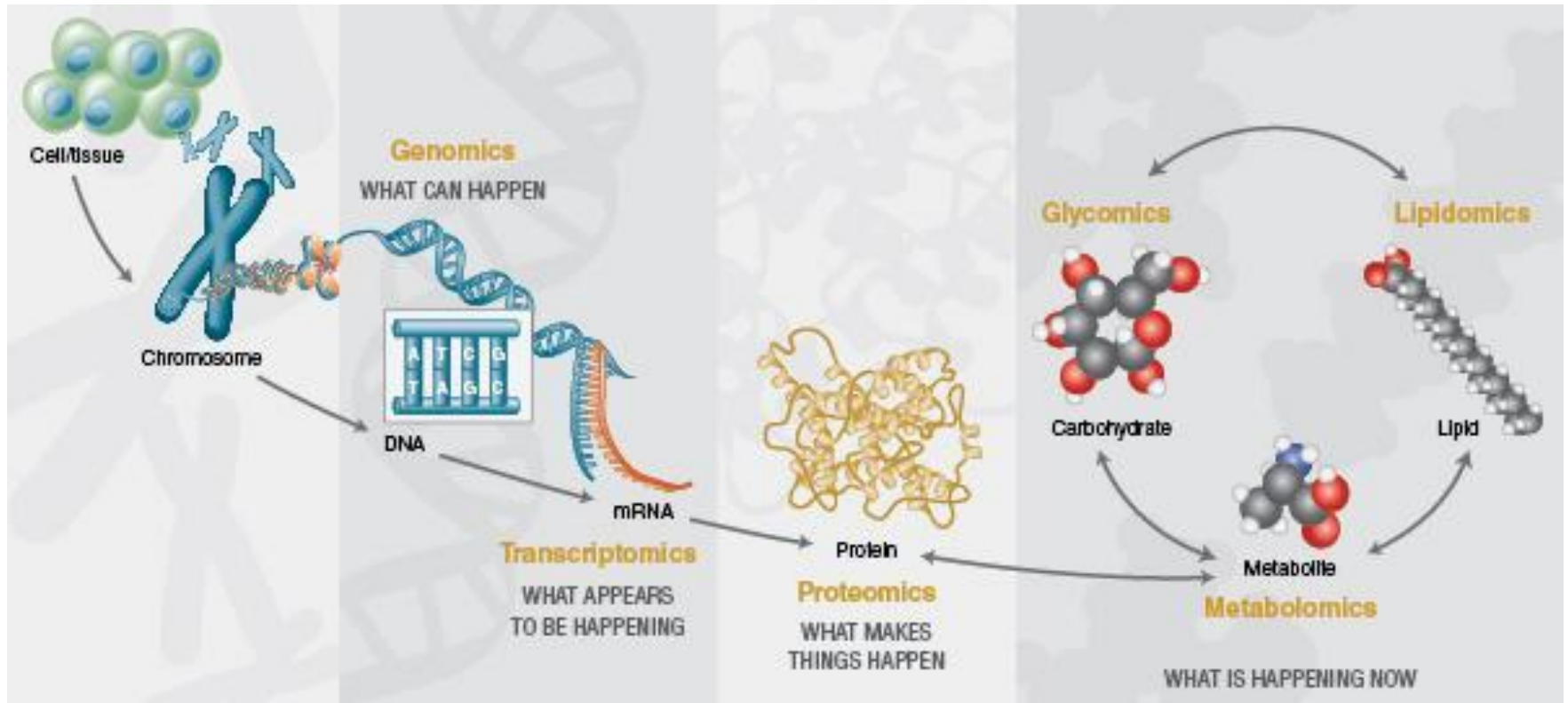


Omics can be defined as comprehensive study of set of molecules.

Central dogma

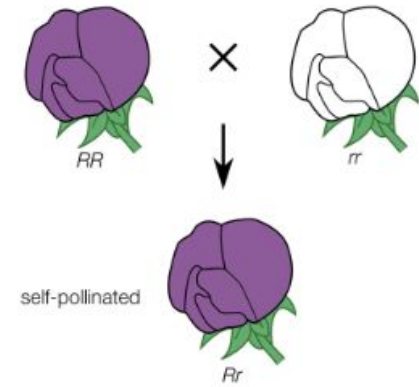
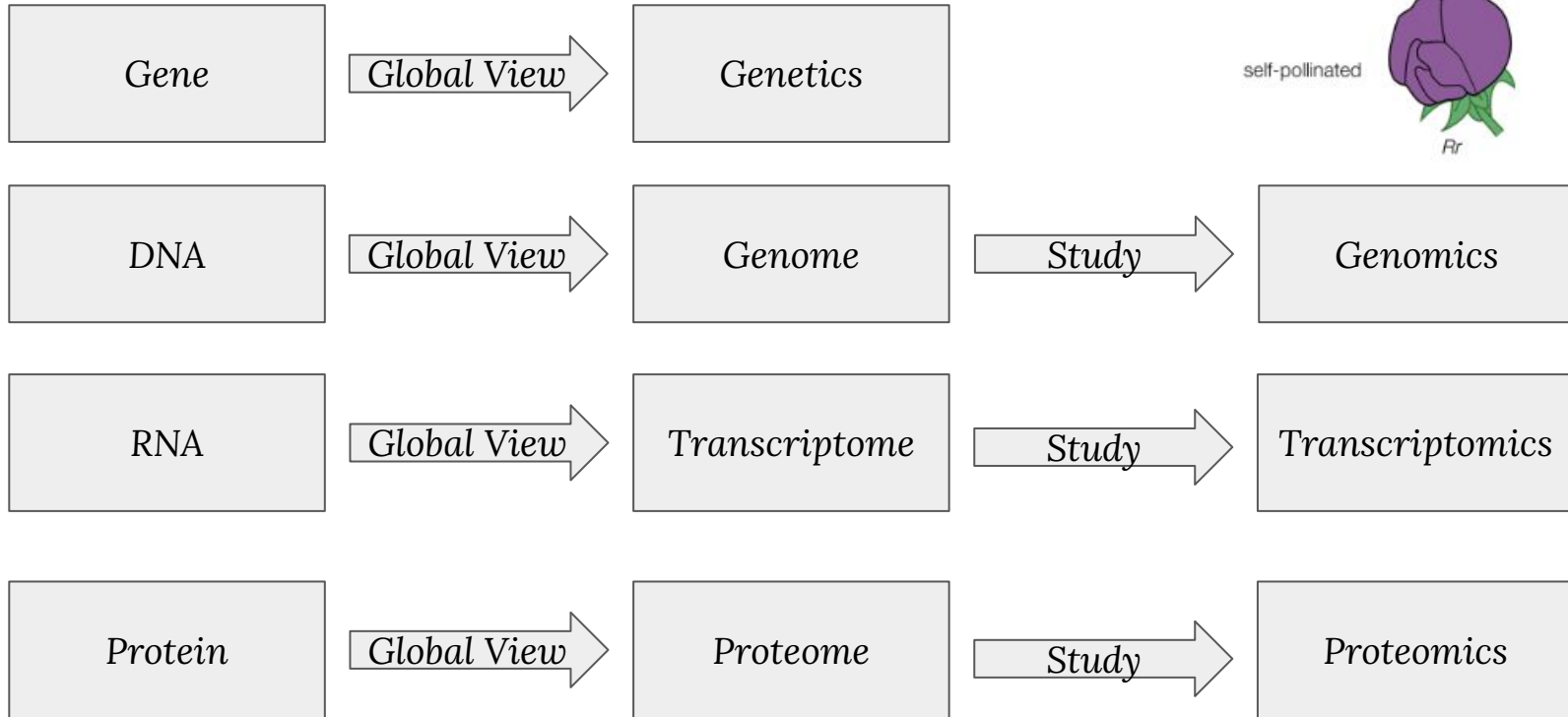


Multi-omics (Pan-omics) Cont.

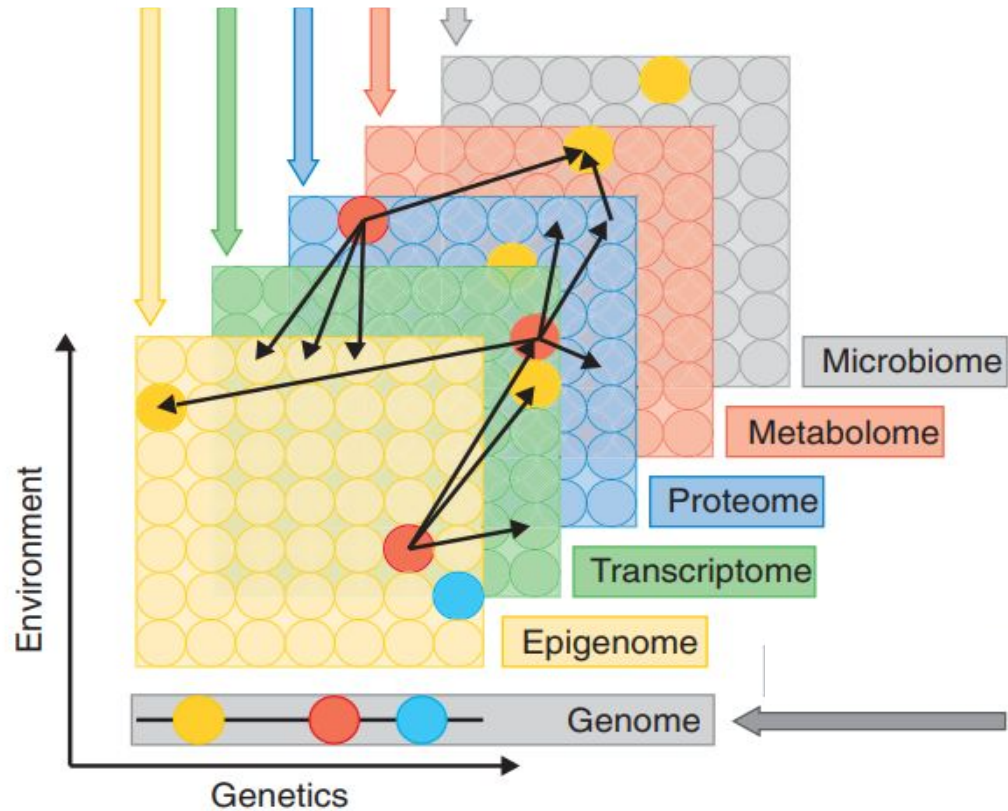


The relationship of major 'omics' technologies: ([Source](#))

Multi-omics (Pan-omics) Cont.



Omics?



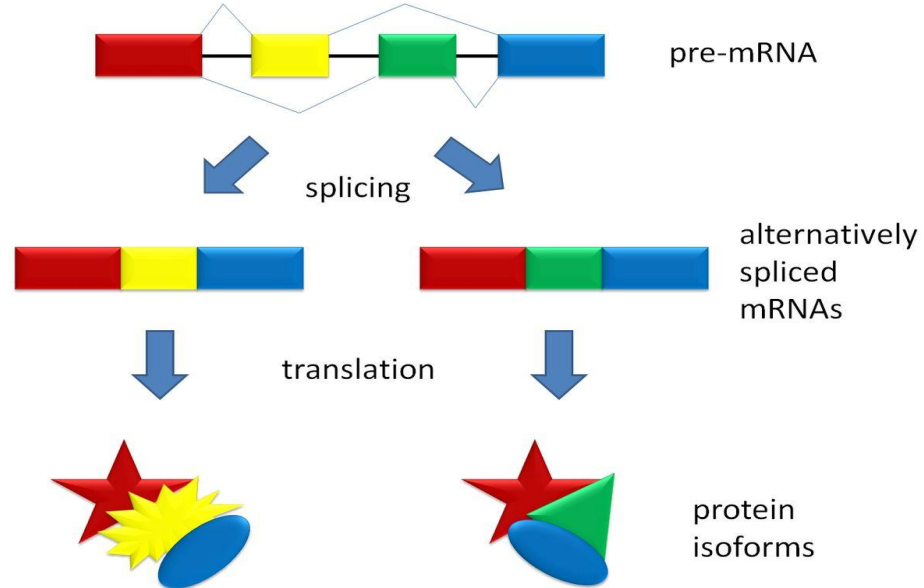
Transcriptomics

The study of the complete set of **mRNA** transcripts demonstrating the functional portion of the genome under

- Specific circumstances.
- In a specific cell.

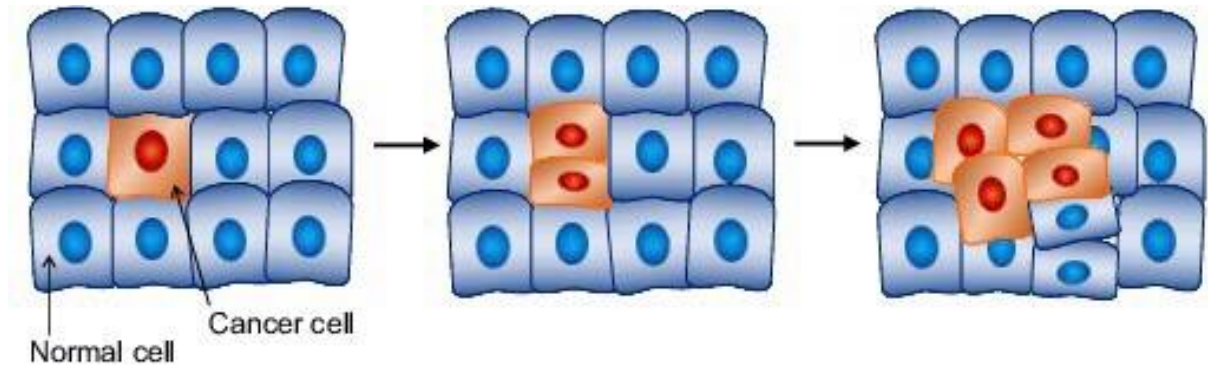
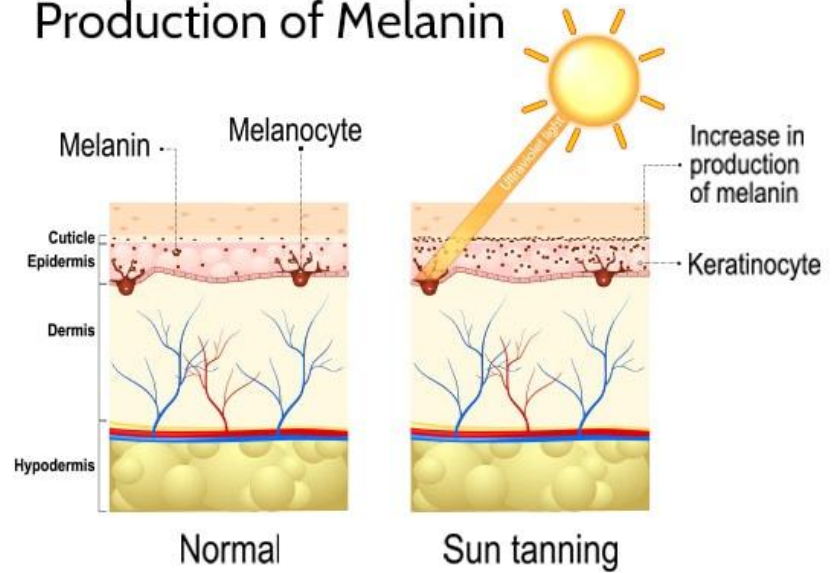
Using, **high-throughput methods** such as microarray analysis and RNA-Seq.

- Qualitative
 - *Alternative splicing and isoforms.*



- Quantitative analyses of RNA
 - Rate of expression

Production of Melanin



High-throughput sequencing

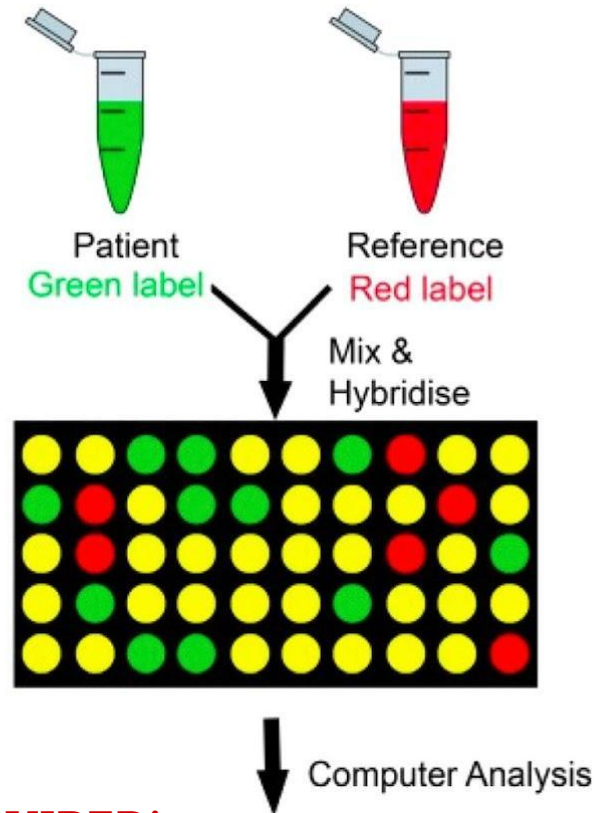
The use of automated equipment to rapidly test thousands to millions of samples.










- A. Large number of mRNA.
- B. Higher resolution.

- Microarrays
- Next generation sequencing
- RNA-sequencing

Microarray

- Hybridization

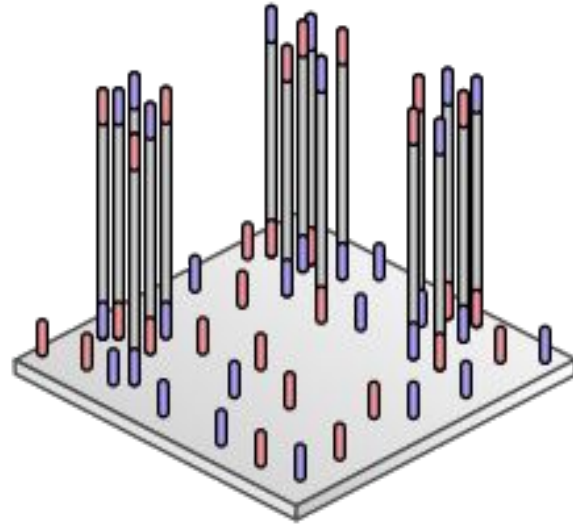


<u>Spot</u>	<u>Patient</u>	<u>Control</u>
	 2 copies	 2 copies
	 3 copies	 2 copies
	 1 copy	 2 copies

PRIOR KNOWLEDGE IS REQUIRED!

RNA-Seq

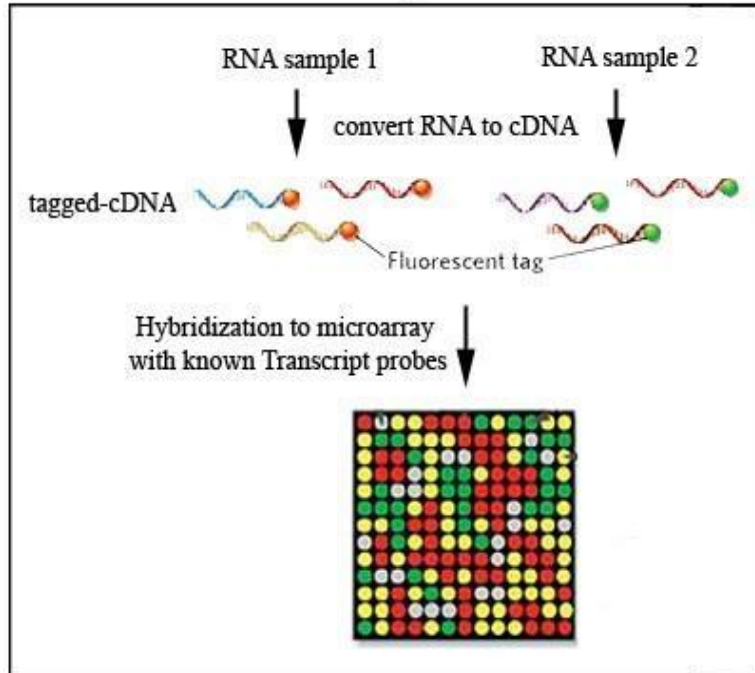
- *High-throughput sequencing*



Microarray

Expression level: Sequencing reads

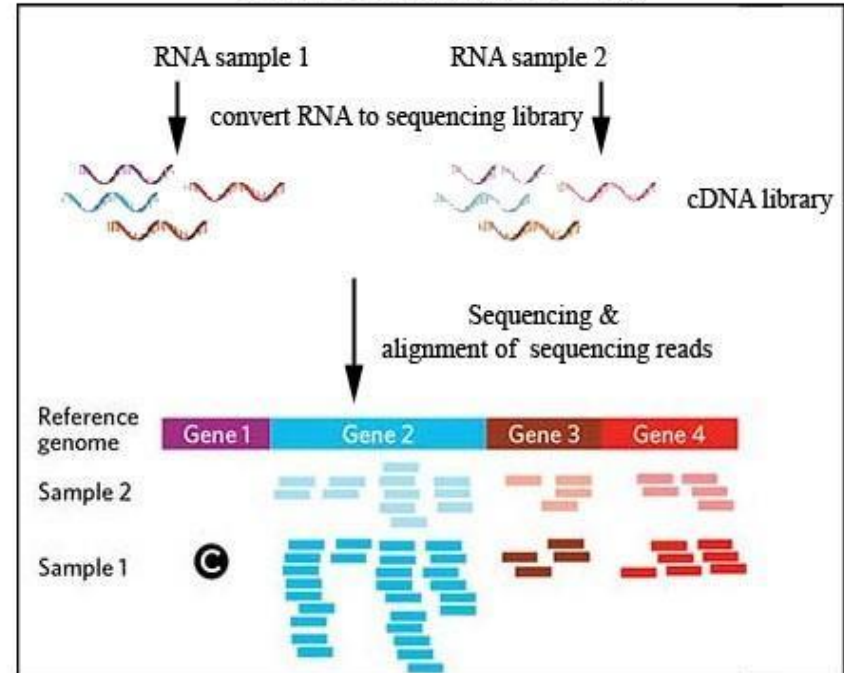
Microarray



RNA-Seq

Relative intensity

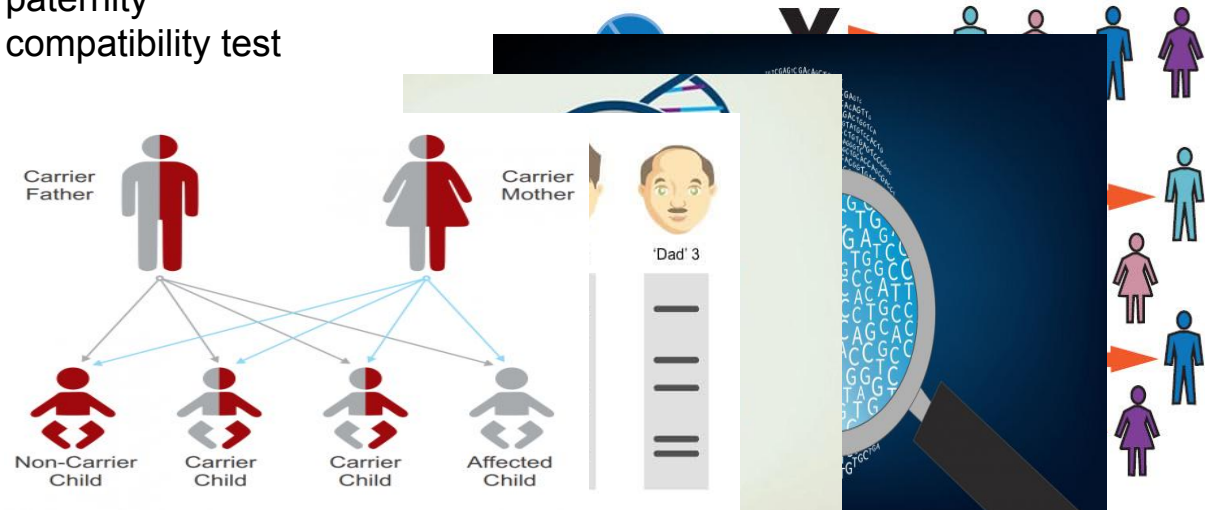
RNA Sequencing (RNA-Seq)



RNA-seq vs Microarray

RNA-Seq	Microarray
<i>Based on high-throughput sequencing</i>	<i>Based on hybridization</i>
<i>Allows for full sequencing of the whole transcriptome.</i>	<i>Gene fragments</i>
<i>Novel transcripts can be identified.</i>	<i>Only known transcripts to allow hybridization.</i>
<i>More specificity.</i>	<i>Less sensitivity.</i>
<i>Reveal alternative splicing</i>	<i>No alternative splicing information can be provided.</i>
<i>More expenses.</i>	<i>Lower cost</i>

- Genomics:
 - study of the entire genomes
 - Identify the genetic variants associated with disease, and response to treatment
 - Applications
 - Ancestry
 - Personalized and precision medicine
 - Genetic disease risk
 - Genetic paternity
 - Genetic compatibility test



- Proteomics
 - Interactions between proteins
 - the functions of a large fraction of proteins are mediated by post-translational modifications such as proteolysis, glycosylation, phosphorylation, nitrosylation, and ubiquitination. Such modifications play key roles in intracellular signaling, control of enzyme activity, protein turnover and transport, and maintaining overall cell structure
- Epigenomics:
 - modifications of DNA or DNA-associated proteins, such as DNA methylation or histone acetylation
 - non coding RNA
 - Metagenomics
- Metabolomics
 - metabolic function

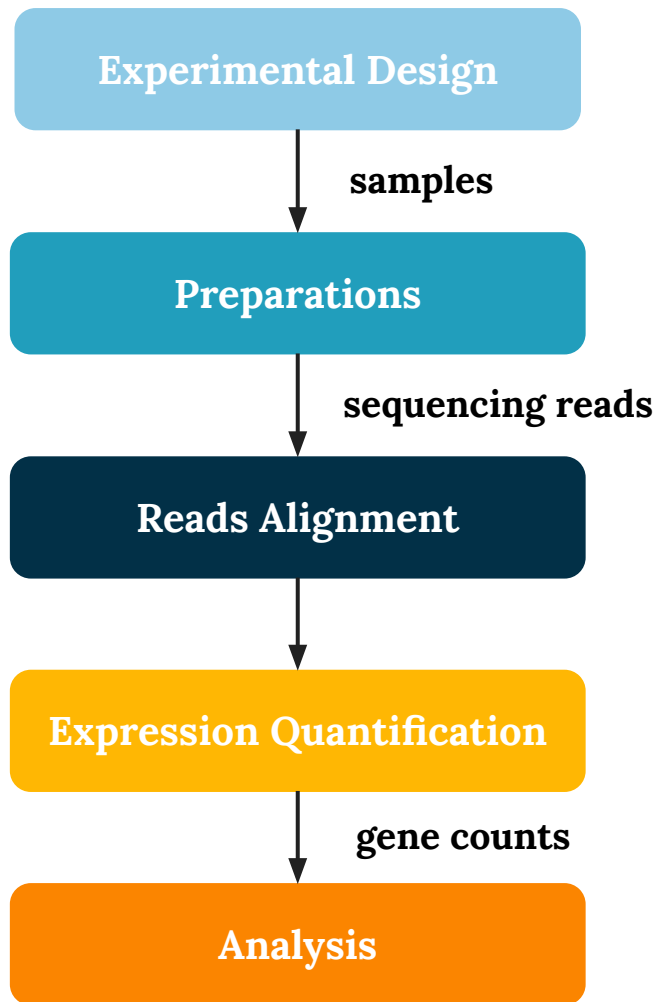
10 minutes break!

RNA-Sequencing

RNA-Seq Workflow

Used to determine RNA expression levels more accurately than microarrays. In principle, it is possible to determine the absolute quantity of every molecule in a cell population, and directly compare results between experiments.

NOW WHAT?!!!



Experiment vs Study

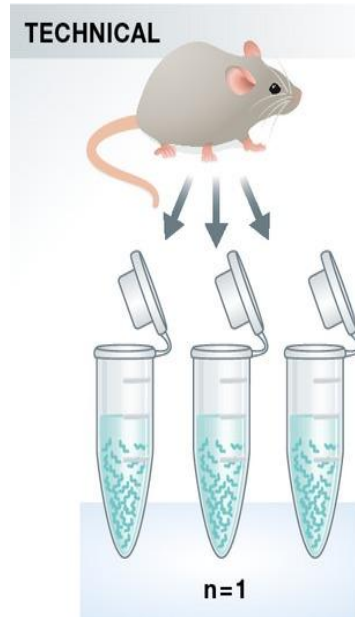
- *In an **experiment**, one uses highly controlled conditions to look at a (model) system, performs specific well-designed interventions at controlled times and intensities, and has an efficient assay to measure the effect of interest. You control the “experimental units” (such as cells, mice, and genotypes) and plan which experiments to perform and when.*
- *On the other hand, the **observations in a study** are made “in the wild”. It requires much bigger sample size than an experiment and is more complicated to analyze, usually requiring involvement of a specially trained expert at some point. In this series.*

I. Experimental Design

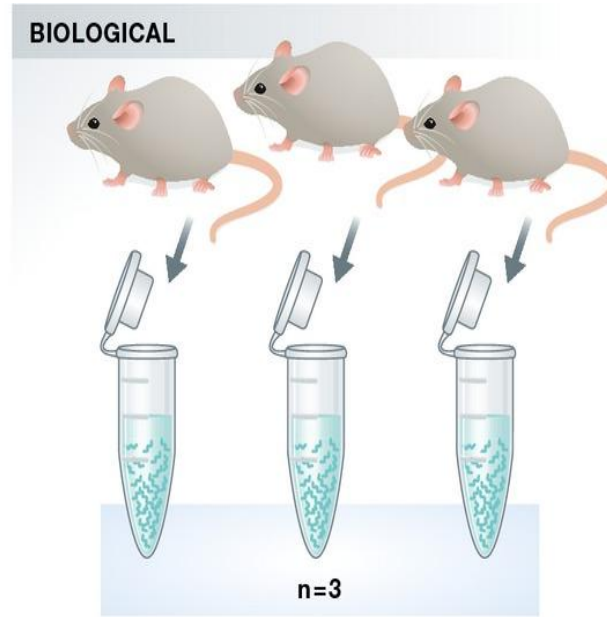
Important considerations that greatly affect the quality of a differential expression analysis:

- 1. Replicates**
- 2. Confounding**
- 3. Batch effects**

1. Replicates



**Technical
replicates**



**Biological
replicates**

Necessity?

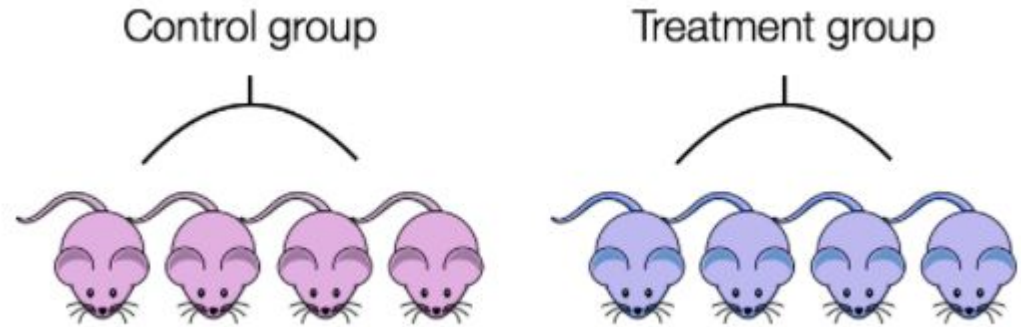
2. Confounding

HOW TO SOLVE?

Confounding is one of several threats to the internal validity of a research study. It is defined as a possible source of bias in studies in which an unmeasured third variable (the confounder) is related to the exposure of interest (although not causally) and causally related to the outcome of interest.

Eg: we know that gender has large effects on gene expression, and if all of our control mice were female and all of the treatment mice were male, then our treatment effect would be confounded by gender.

We could not differentiate the effect of treatment from the effect of gender.



3. Batch effect

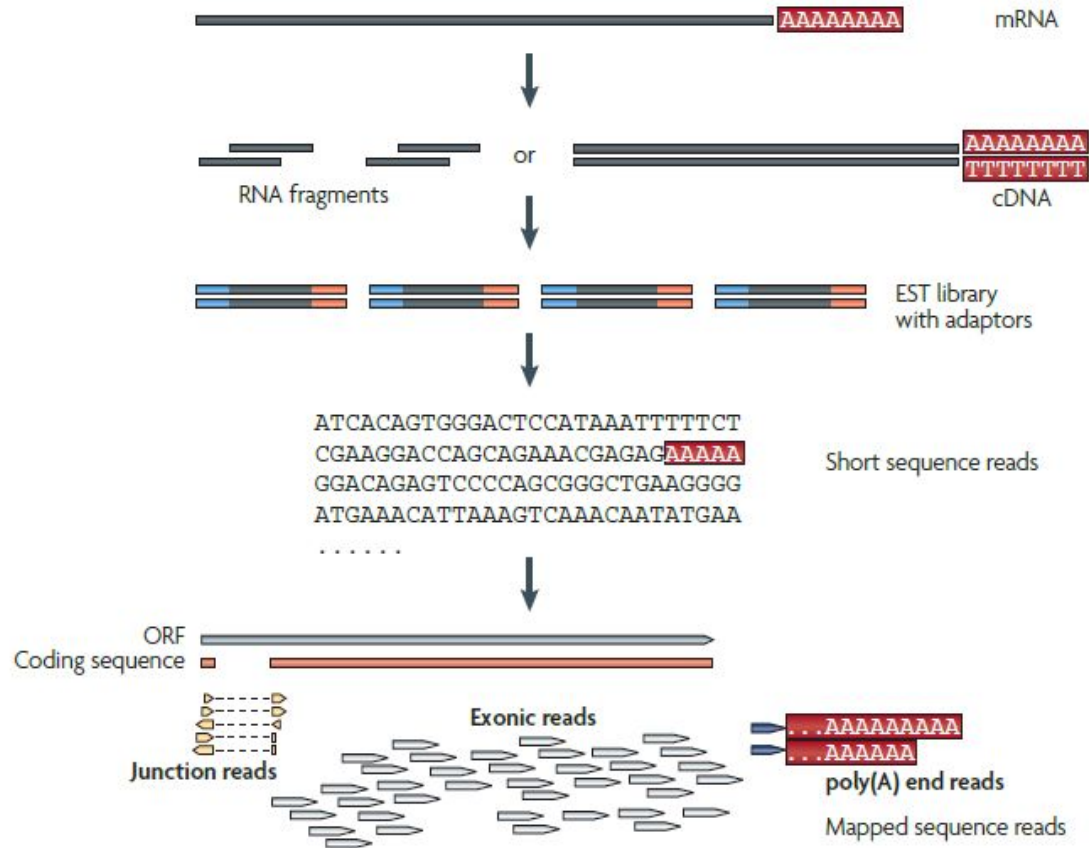
Questions criteria to ask whenever in a doubt:

- 1. Were all RNAs and library preparations performed on the same day?*
- 2. Did the same person perform the RNA isolation/library preparation for all samples?*
- 3. Did you use the same reagents for all samples?*
- 4. Did you perform the RNA isolation/library preparation in the same environment?*

II. Preparations

Mainly includes:

1. *RNA isolation*
2. *RNA or cDNA fragmentation*
3. *cDNA synthesis*
4. *Adapter ligation*
5. *Amplification*
6. *Bar-coding*
7. *Lane loading*



Wang Z et al. (2009) RNA-Seq: a revolutionary tool for transcriptomics Nature Reviews Genetics 10:57-63

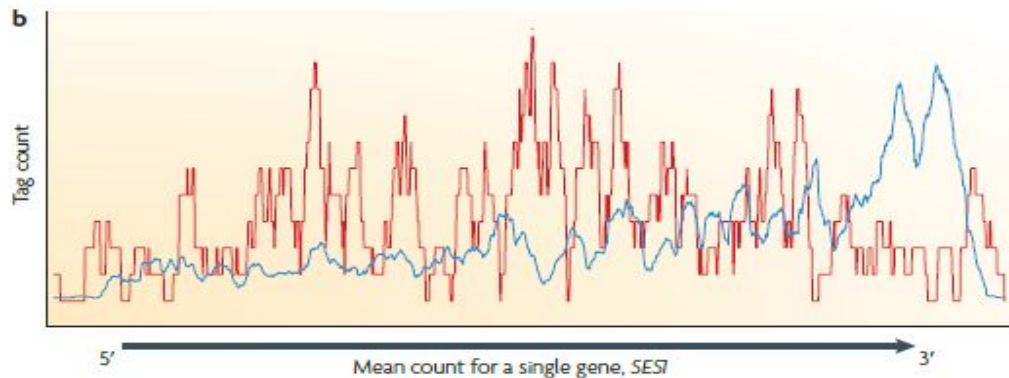
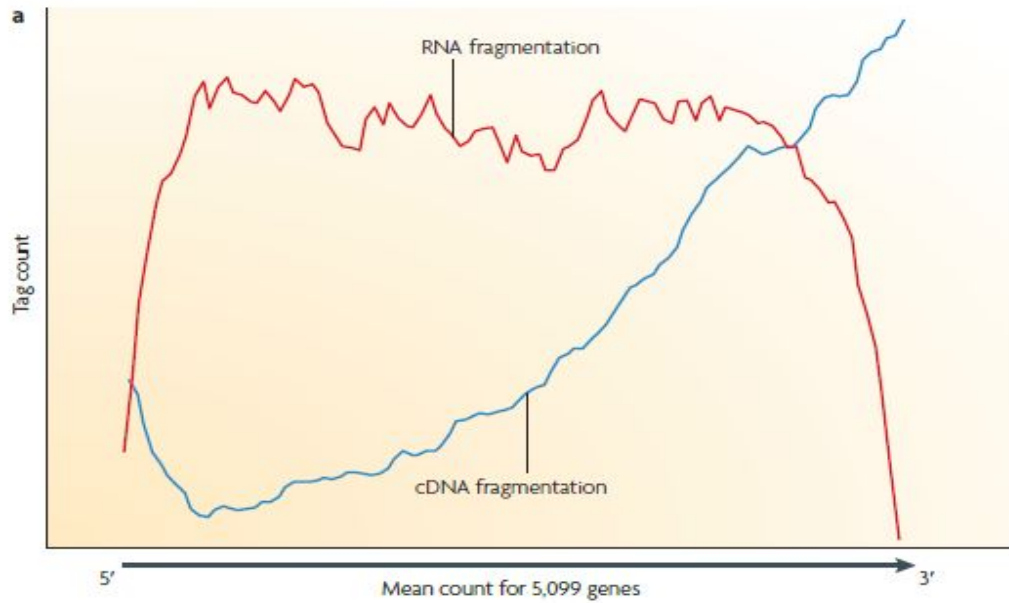
Challenges in RNA-Seq library preparation phase

1. RNAs:

- a. Several manipulation stages occur during the production of cDNA libraries, unlike small RNAs which can be directly sequenced after adaptor ligation, larger RNA molecules must be fragmented into smaller pieces (200–500 bp) to be compatible with most deep-sequencing technologies.

2. Fragmentation:

- a. **RNA fragmentation:** bias over the transcript body, but is depleted for transcript ends.
- b. **DNA fragmentation:** usually strongly biased towards the identification of sequences from the 3' ends of transcripts, and thereby provides valuable information about the precise identity of these ends



- a** | Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3' end of the transcript. RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5' and 3' ends. Note that the ratio between the maximum and minimum expression level (or the dynamic range) for microarrays is 44, for RNA-Seq it is 9,560. The tag count is the average sequencing coverage for 5,000 yeast ORFs. **b** | A specific yeast gene, *SES1* (seryl-tRNA synthetase), is shown.

Challenges in RNA-Seq library preparation phase (Cont.)

3. Short Reads:

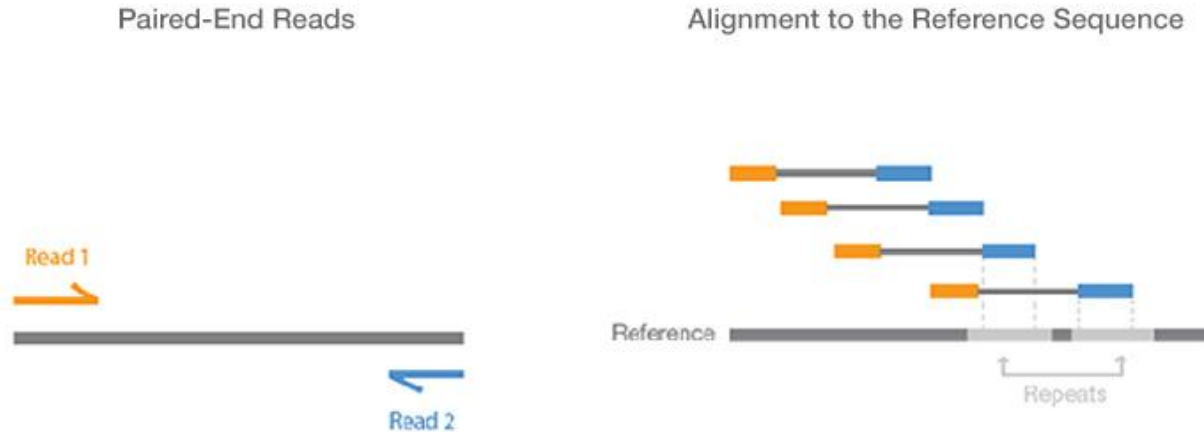
- a. *Shorts reads that are identical to each other can be obtained from cDNA libraries that have been amplified. These could be a genuine reflection of abundant RNA species, or they could be PCR artefacts.*
 - i. **Solution:** *to discriminate between these possibilities is to determine whether the same sequences are observed in different biological replicates.*

Challenges in RNA-Seq library preparation phase (Cont.)

4. Strand-specific libraries:

- a. Another key consideration concerning library construction is **whether or not to prepare strand-specific libraries**. These libraries have the **advantage** of yielding information about the orientation of transcripts, which is valuable for transcriptome annotation, especially for regions with overlapping transcription from opposite directions. However, they are currently laborious to produce because they require many steps or direct RNA-RNA ligation, which is inefficient.

Single-end Vs Paired-end?



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

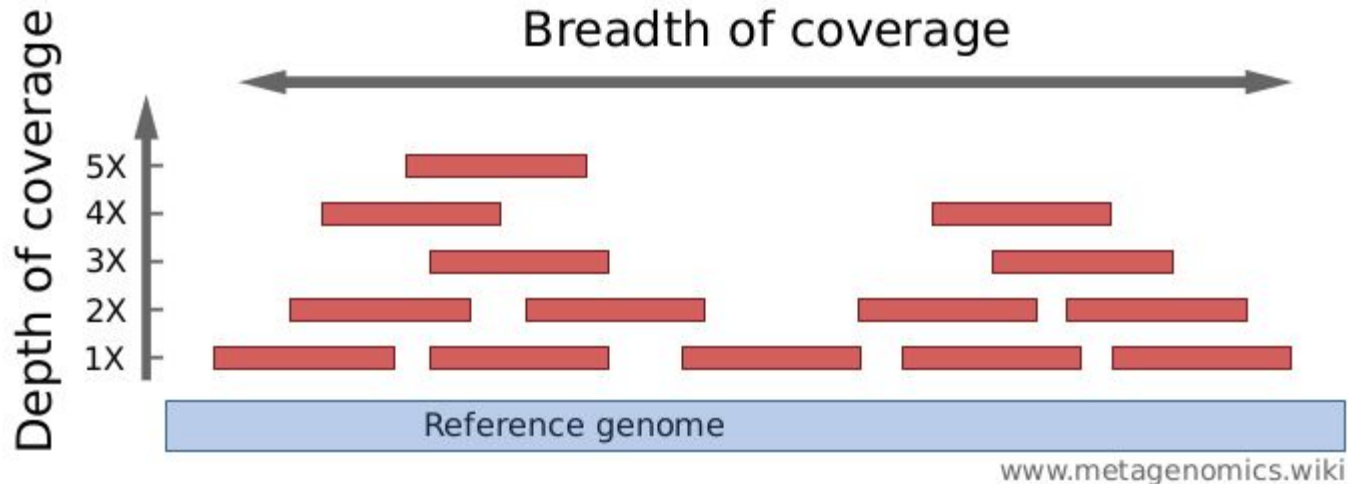
Single-end Vs Paired-end

When to use?

- *Single-end sequencing: when studying changes in gene expression*
- *Paired-end sequencing:*
 - *Whole genome sequencing*
 - *Alternative splicing*
 - *De novo transcriptome studies*

Important guidelines

- If mRNA extracted is uncorrupted, then pass it for cDNA conversion by poly(A) selection, but if degraded, then repair the mRNA strand and then proceed.
- **Sequencing depth:**



Important guidelines Cont. (Sequencing depth)

1. *The average depth of sequencing coverage can be defined theoretically as LN/G , where L is the read length, N is the number of reads and G is the haploid genome length.*
2. *Depth of coverage is affected by the accuracy of genome alignment algorithms and by the uniqueness or the 'mappability' of sequencing reads within a target genome.*
3. *Bigger the genome or transcriptome, means we need greater number of reads to cover the entire length of the sequence.*

Coverage Vs Cost

- Another important issue is sequence coverage, or the percentage of transcripts surveyed, which has implications for cost. Greater coverage requires more sequencing depth. To detect a rare transcript or variant, considerable depth is needed.
- In simple transcriptomes, such as yeast for which there is no evidence of alternative splicing, **30 million 35-nucleotide reads** from poly(A) mRNA libraries are **sufficient** to observe transcription from **most (>90%) genes** for cells grown under a single condition (that is, in nutrient-rich medium)
- This depth is probably more than sufficient for most purposes, as the number of expressed genes detected by RNA-Seq reaches **80% coverage at 4 million uniquely mapped reads**, after which doubling the depth merely increases the coverage by 10%

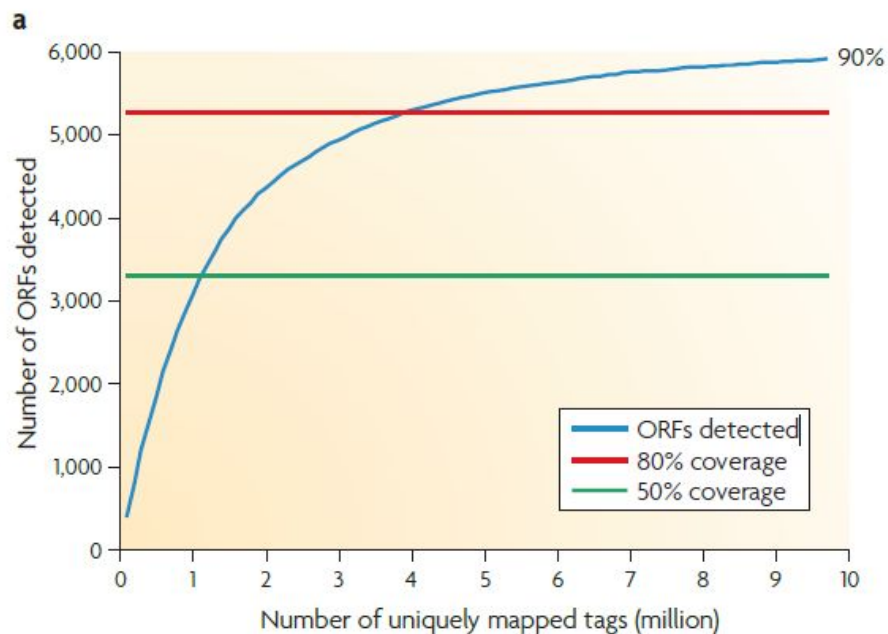
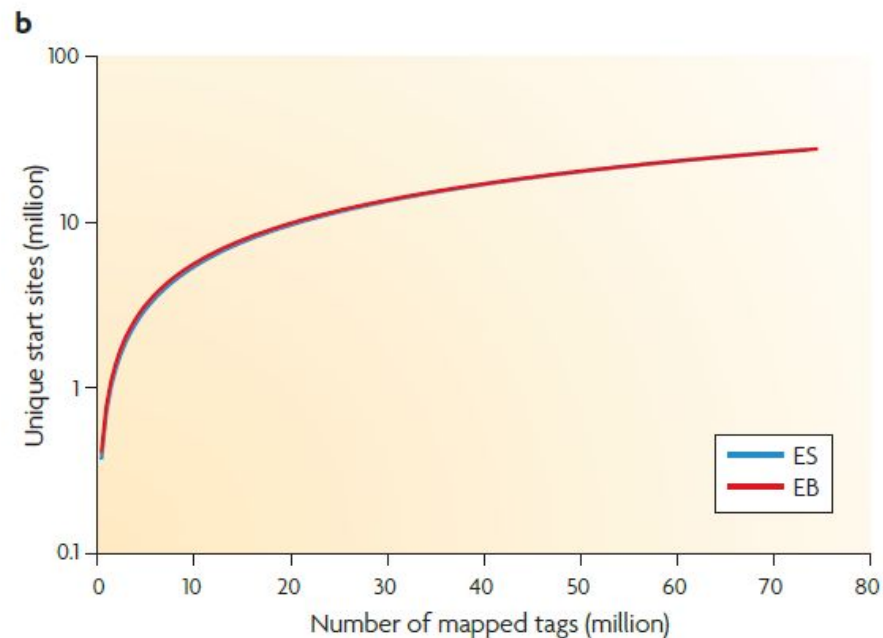
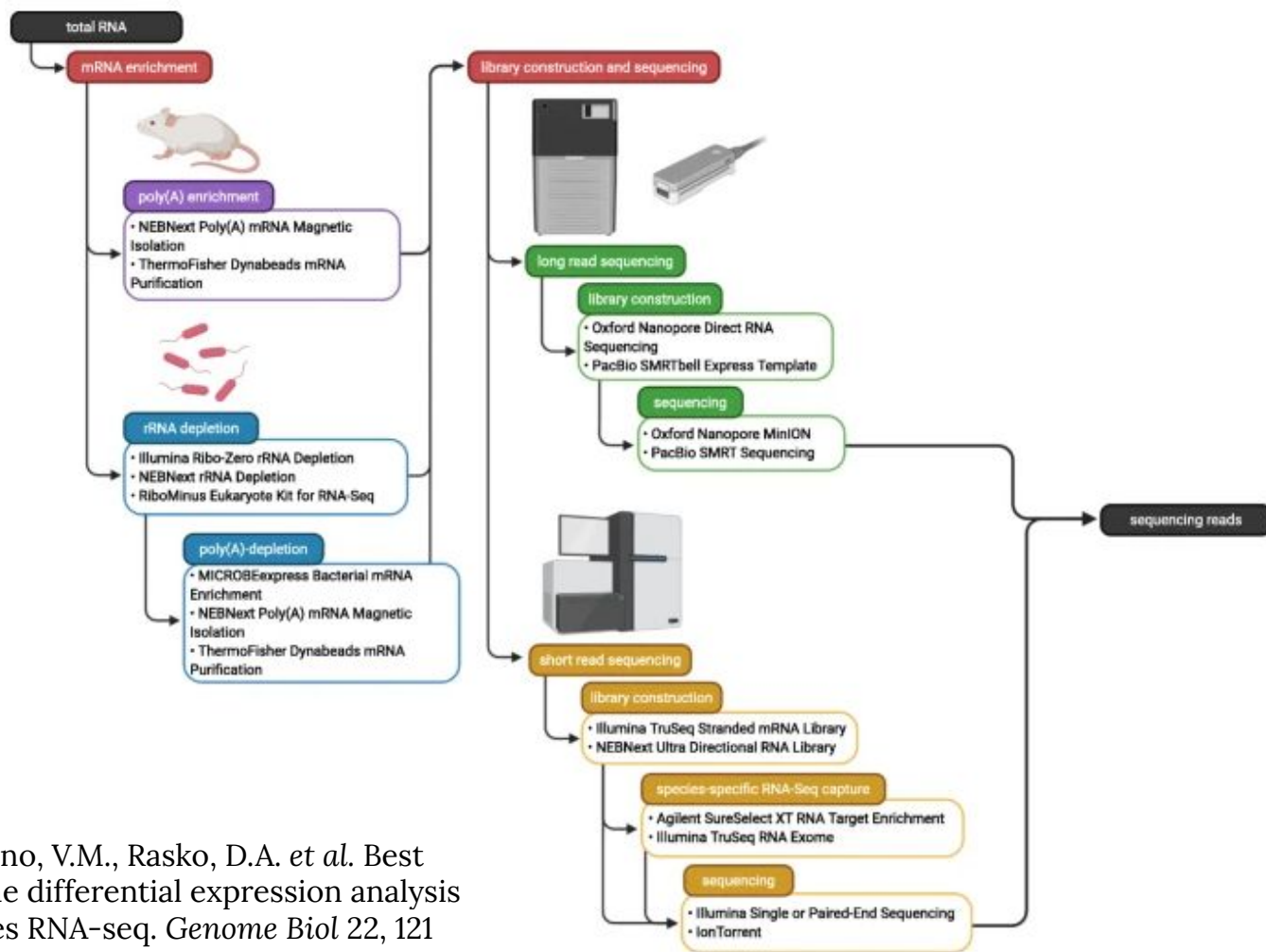


Figure 5 | Coverage versus depth. a | 80% of yeast genes were detected at 4 million uniquely mapped RNA-Seq reads, and coverage reaches a plateau afterwards despite the increasing sequencing depth. Expressed genes are defined as having at least four independent reads from a 50-bp window at the 3' end. Data is taken from REF. 18.



b | The number of unique start sites detected starts to reach a plateau when the depth of sequencing reaches 80 million in two mouse transcriptomes. ES, embryonic stem cells; EB, embryonic body. Figure is modified, with permission, from REF. 22 © (2008) Macmillan Publishers Ltd. All rights reserved.

- Analyzing many different conditions can further increase the coverage; in *S. pombe* 122 million reads from six different growth conditions detected transcription from >99% of annotated genes.
- **In general**, the larger the genome, the more complex the transcriptome, the more sequencing depth is required for adequate coverage.



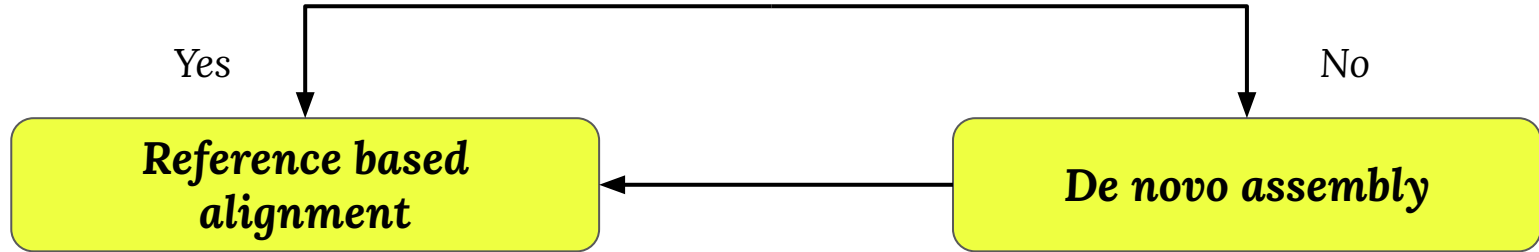
Chung, M., Bruno, V.M., Rasko, D.A. *et al.* Best practices on the differential expression analysis of multi-species RNA-seq. *Genome Biol* 22, 121 (2021).

Initial processing of sequencing reads

1. **Demultiplex** by index or barcode
2. **Remove** adapter sequences
3. **Trim** reads by quality
4. **Discard** reads by quality/ambiguity
5. **Filter** reads by k-mer coverage

III. Reads Alignment

Is there an available reference for all genome in the analysis?



- Percentage of mapped reads 70-90 %.
- The uniformity of read coverage on exons and the mapped strand
- Results:
 - SAM files

- PE strand-specific sequencing and long reads are preferred.
- Results:
 - Contigs

Sequence reads challenges

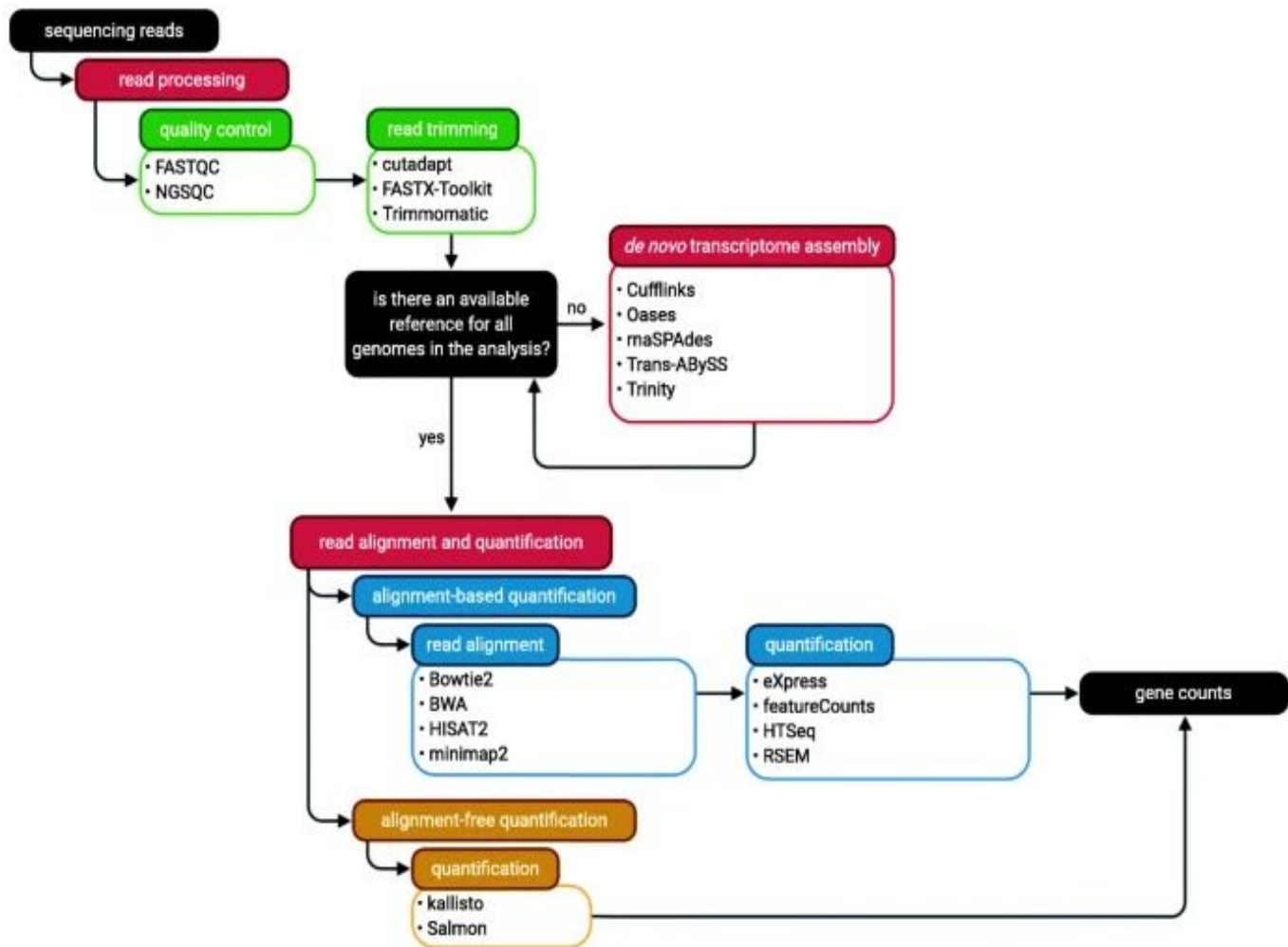
- Short transcriptomic reads also contain reads that span exon junctions or that contain poly(A) ends – these cannot be analysed in the same way.
- For genomes in which splicing is rare (for example, *S. cerevisiae*) special attention only needs to be given to poly(A) tails and to a small number of exon–exon junctions.
- **Transcriptome maps:**
 - a. For complex transcriptomes it is more difficult to map reads that span splice junctions, owing to the presence of extensive alternative splicing and trans-splicing.
 - i. **Solution?** Compile a junction library that contains all the known and predicted junction sequences and map reads to this library

Sequence reads challenges

- For large transcriptomes, alignment is also complicated by the fact that a significant portion of sequence reads match multiple locations in the genome.
 - a. **Solution?** Assign these multi-matched reads by proportionally assigning them based on the number of reads mapped to their neighbouring unique sequences.
- Sequencing errors and polymorphisms can present mapping problems for all genomes, not just for repetitive DNA. Generally, single base differences are not problematic, because most mapping algorithms accommodate one or two base differences.

V. Expression Quantification

- It is the approach of quantifying gene expression by RNA-seq is to count the number of reads that map (i.e. align) to each gene (read count)
- **In the case of alignment-dependent quantification:**
 - Achieved using the resulted indices files for alignment (i.e, aligned reads) and a gene transfer format (GTF) file.
- **In the case of alignment-independent quantification:**
 - Reads are quantifies based on a pseudo-alignment or quasi-mapping of read k -mers.
 - As inputs, k -mer-based tools require an index generated from a nucleotide Fasta file containing the transcript sequences of the target organism along with paired-end Fastq files.
 - Developed tools can operated this process in an efficient time standards.



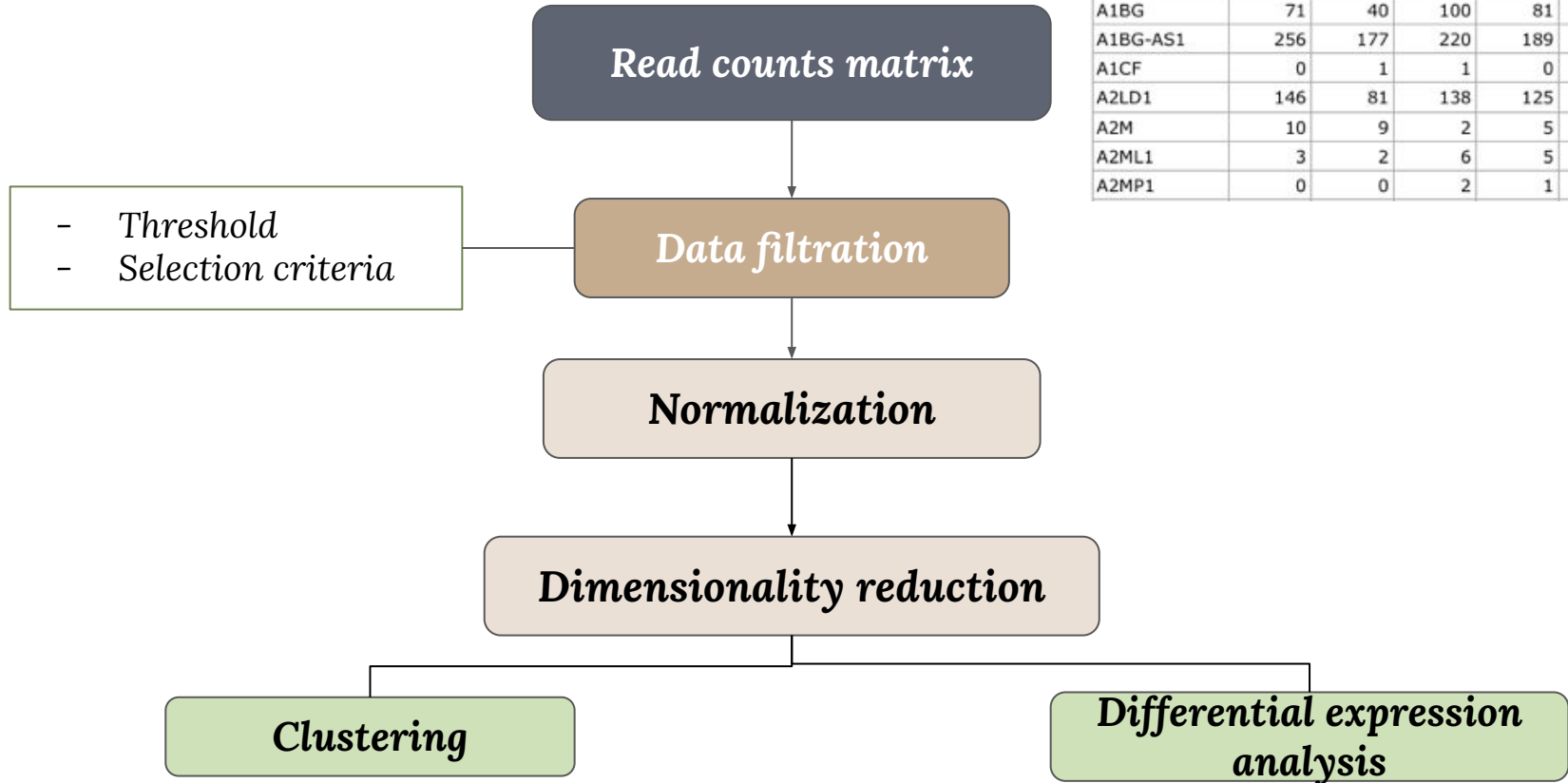
VI. Analysis

Having the gene counts matrix, **let the most interesting phase begin!**

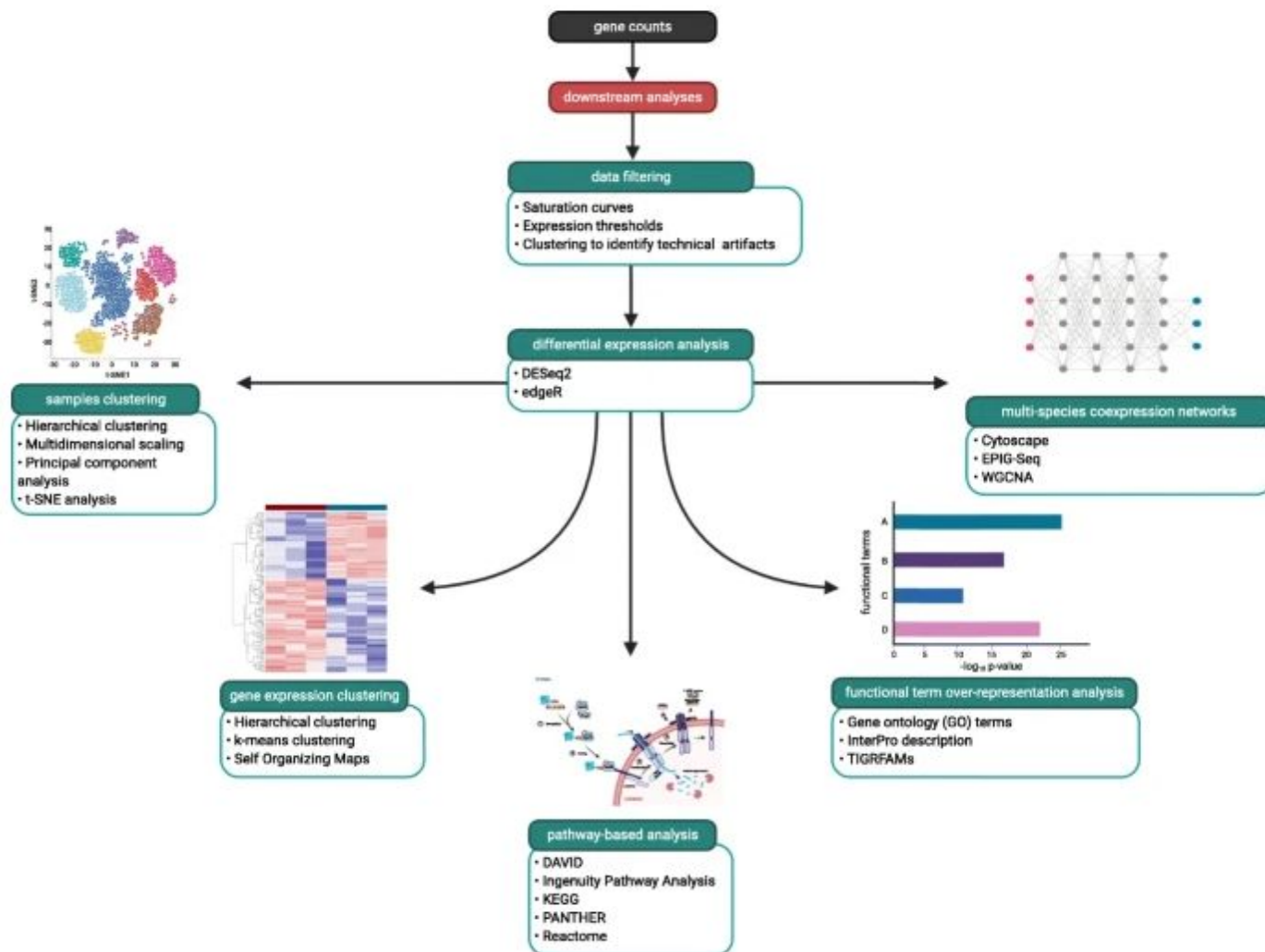
In the data analysis phase, we aim to gain insights about our biological data and answer the biological question we raised. Accordingly, we can perform our analysis by multiple steps depending on our goal or the question we want to solve.

A fundamental research aim in many RNA-seq studies is to identify differentially expressed genes between distinct sample groups.

Analysis pipeline



GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1
1/2-SBSRNA4	57	41	64	55	38	45
A1BG	71	40	100	81	41	77
A1BG-AS1	256	177	220	189	107	213
A1CF	0	1	1	0	0	0
A2LD1	146	81	138	125	52	91
A2M	10	9	2	5	2	9
A2ML1	3	2	6	5	2	2
A2MP1	0	0	2	1	3	0



Any Questions?

Thank You!