

FACULTAD DE  
CIENCIAS  
BIOLÓGICAS  
Universidad de Concepción



CELL MORPHODYNAMICS  
LATINAMERICA



Universidad de Concepción

# “PRINCIPLES OF TRANSCRIPTOMICS IN DEVELOPMENT”

COURSE “OPTICS, FORCES & DEVELOPMENT”

12TH MARCH 2024

**ESTEFANÍA TARIFEÑO-SALDIVIA PHD.**

FACULTAD DE CS. BIOLÓGICAS  
UNIVERSIDAD DE CONCEPCIÓN

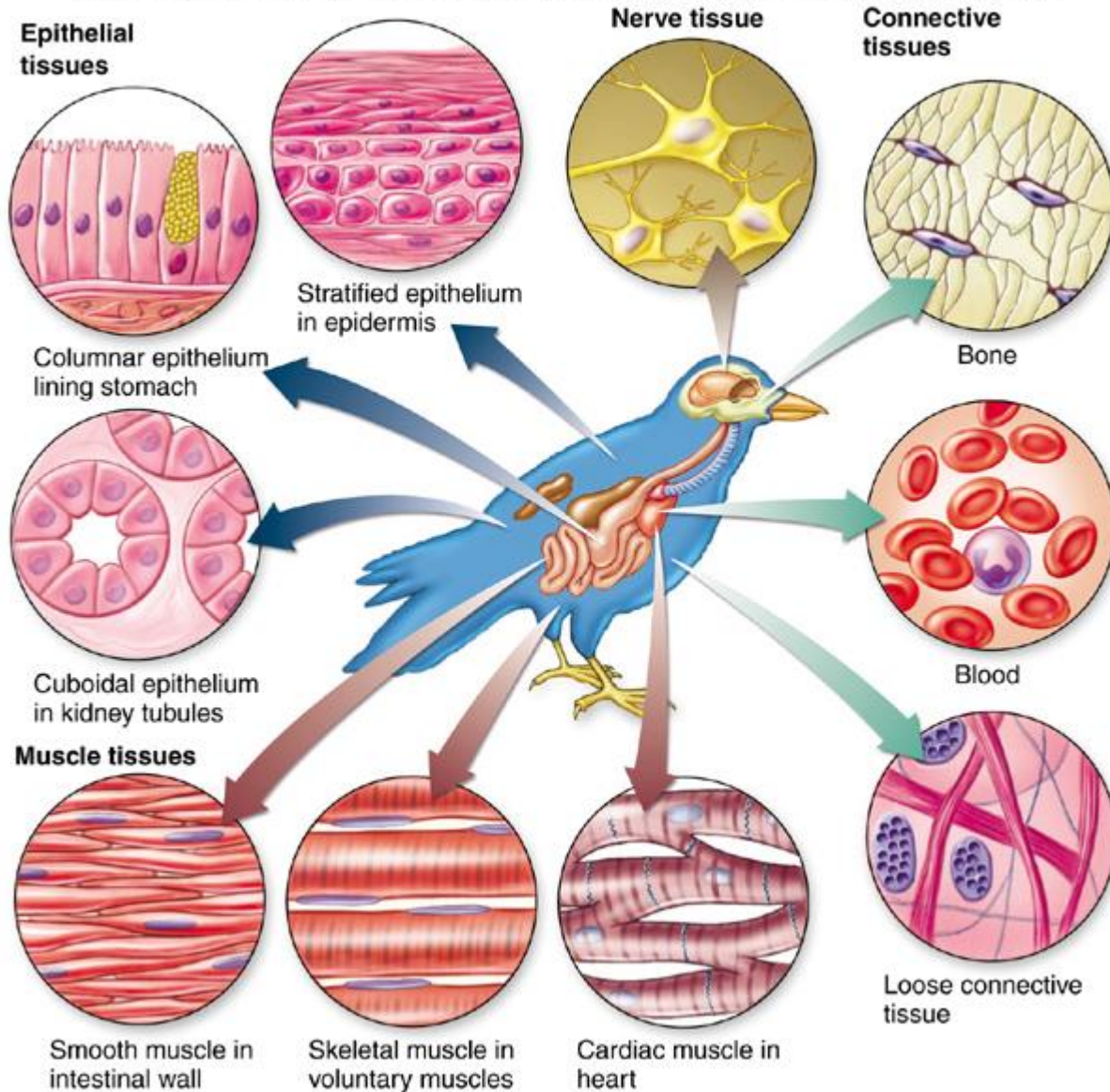
# Class Outline



- **Transcriptomics**
- **Sequencing Technologies**
- **Repositories**
- **Data Analysis**
- **Results visualization**
- **Developmental studies**

# TRANSCRIPTOMICS -- What distinguishes one cell from another?

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.



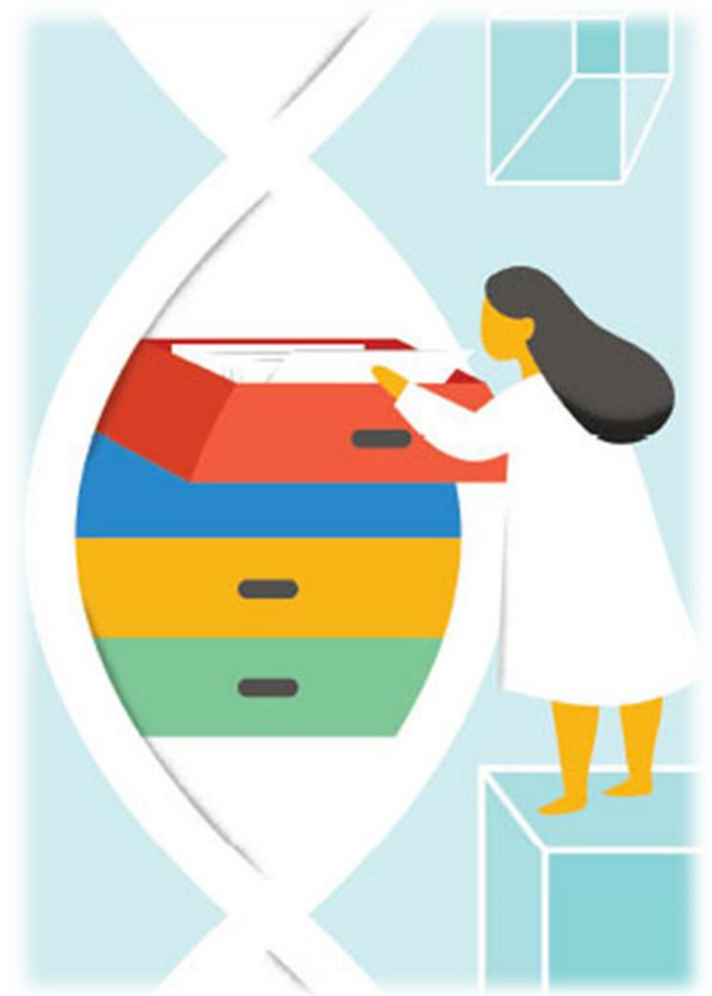
Approximately 20,000 coding and 20,000 non-coding genes are expressed in varying combinations and intensities, contributing to the definition of the **cell-specific transcriptomic fingerprint**.

The **transcriptome** is defined as the complete set of **transcripts**, encompassing both coding and non-coding RNAs, expressed in a cell or tissue at a specific time or condition.

# Transcriptomics

The cell transcriptome is **dynamic**, undergoing changes triggered by external stimuli.

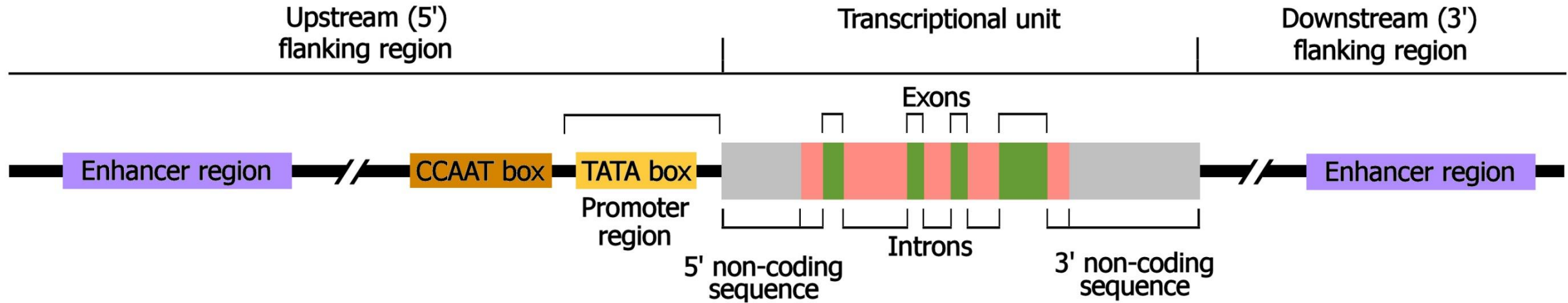
Transcriptomics is the comprehensive study of the transcriptome, to understand how stimuli modulate transcripts expression.





# What are the differences between gene and transcript?

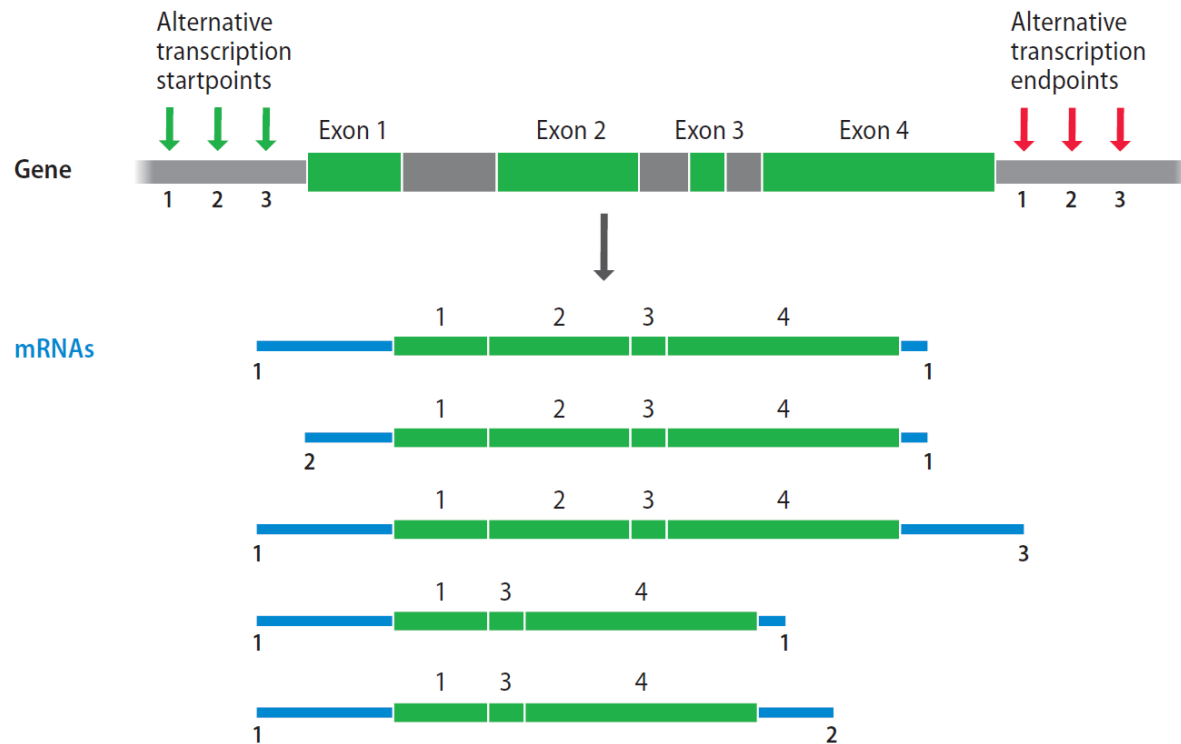
What elements does the basic structure of a gene contain?



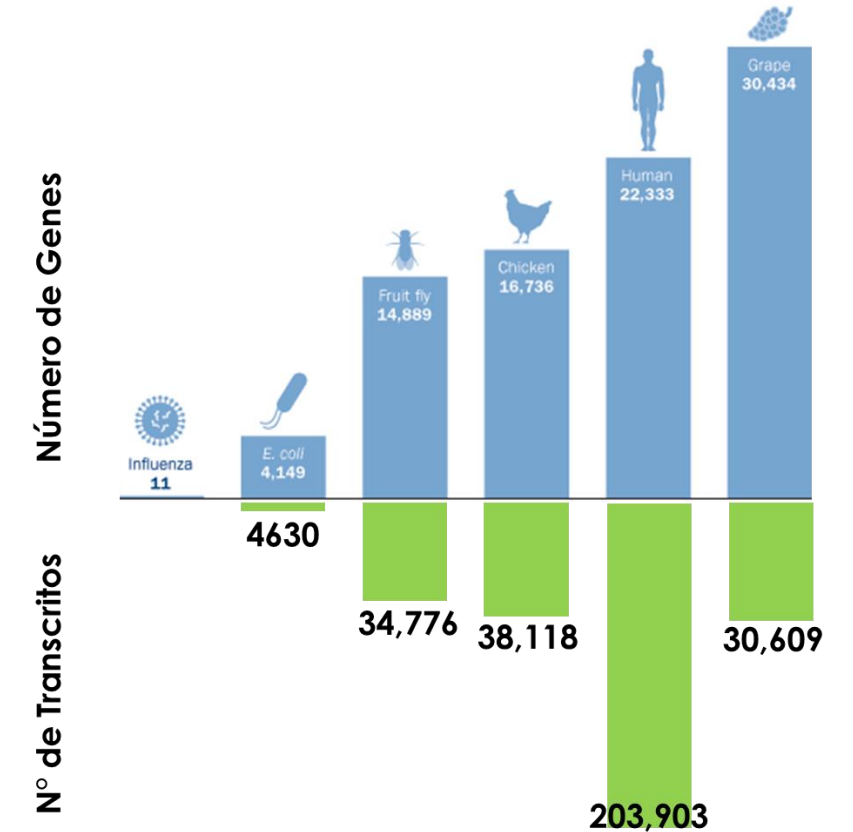
Is there any difference in the structure between coding and non-coding genes?

# What are transcripts?

Are molecules of RNAs synthesized from a DNA template



Related to the complexity of the organism



# So, How can we define a gene?

Any interval of DNA transcribed into a functional RNA molecule

## Accounts for:

- ✓ Non-coding RNAs
- ✓ Coding Genes
- ✓ Splicing Variants

## Exclude:

- **Pseudogenes:** non-functional copies of genes often resulting from gene duplication events, mutations, or evolutionary processes.

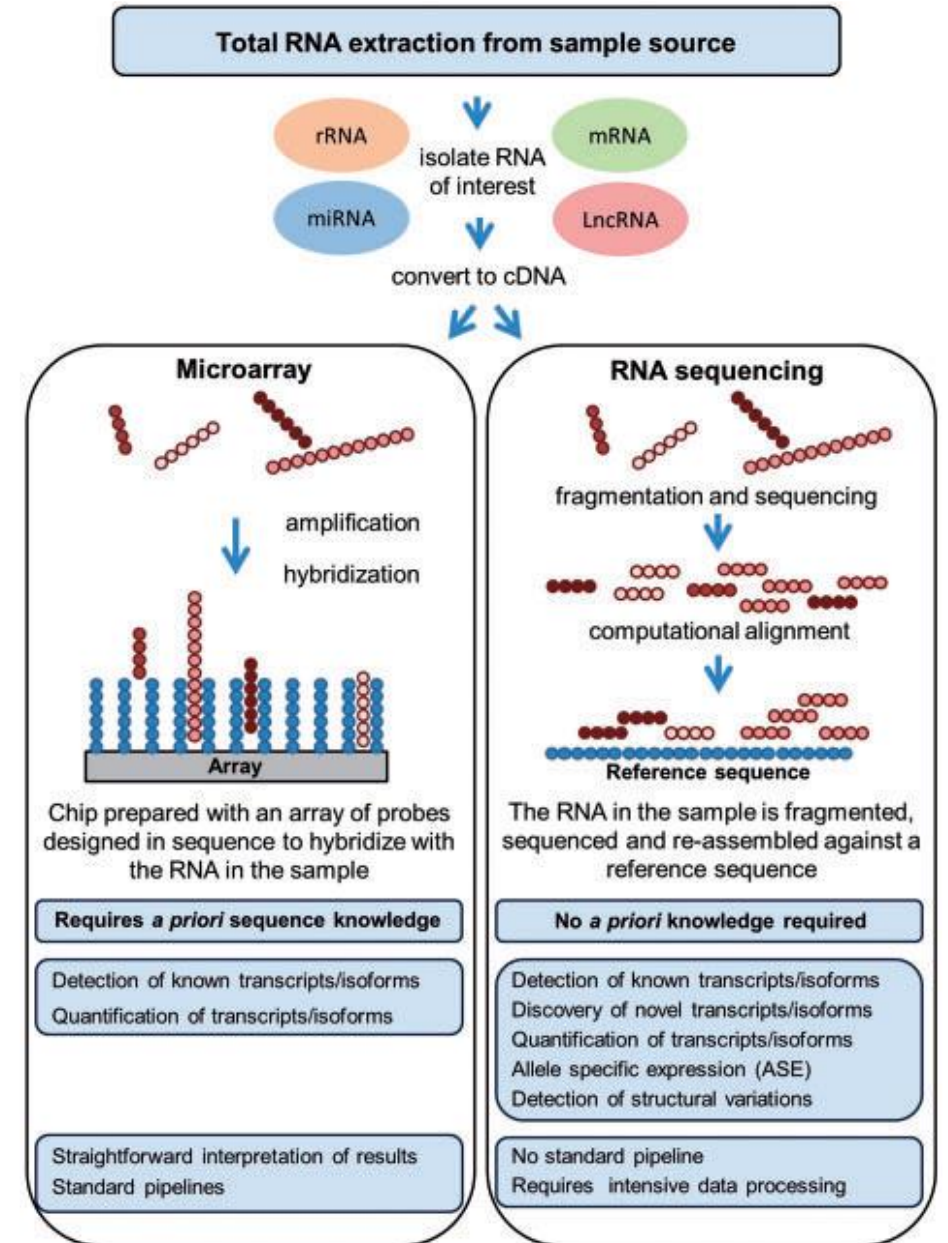
# Transcriptomics is performed by

## 1) RNA-Microarrays

- Based on probe hybridization of a predefined set of genes.
- Required previous knowledge of the sequence to be hybridized
- It is quantitative

## 2) RNA-Sequencing

- Allows the discovery of new transcripts/genes/alternative splicings
- It is quantitative
- The analysis is more complex than microarray





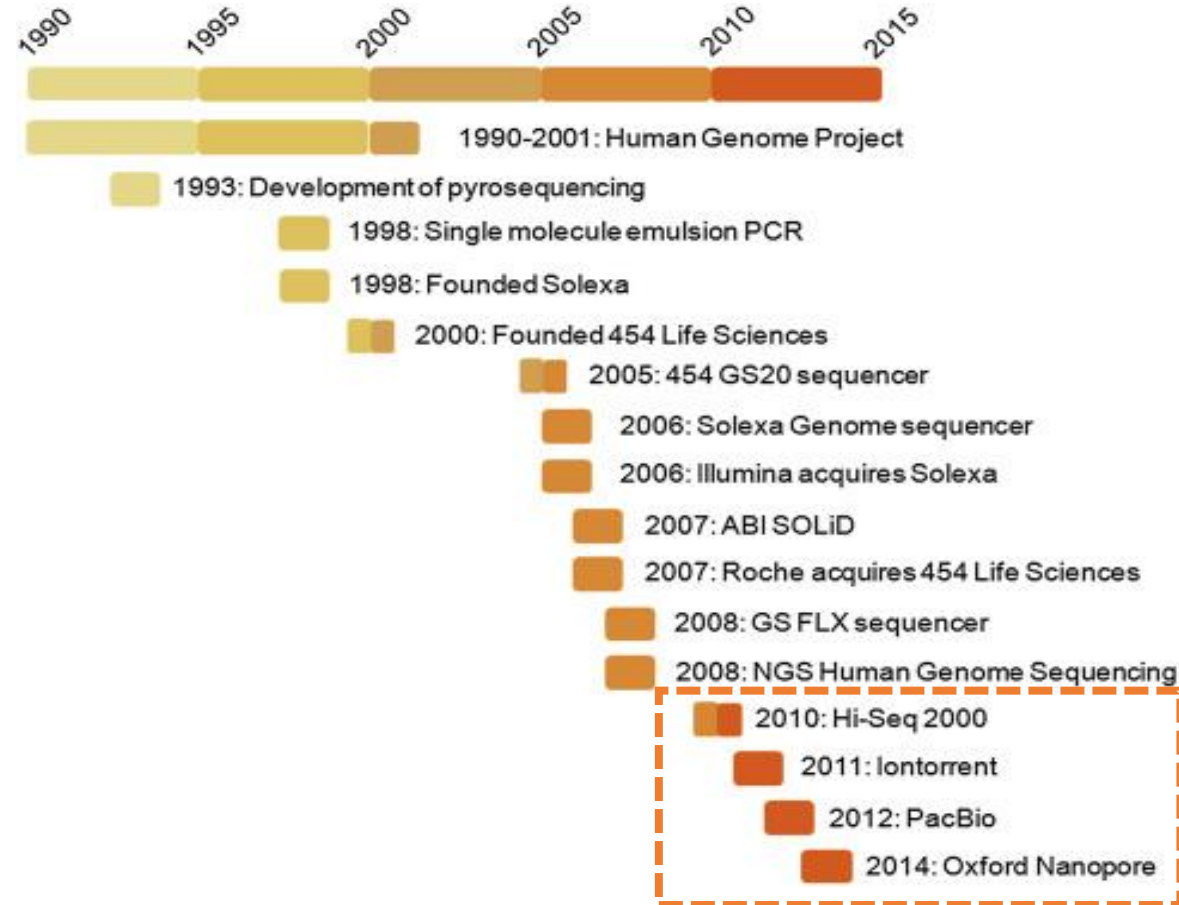
# Transcriptomic analysis output

- Estimate the presence/absence and quantify transcripts.
- Evaluate alternative splicing to determine or predict protein isoforms.
- Quantitatively estimate the influence of genotype on gene expression.

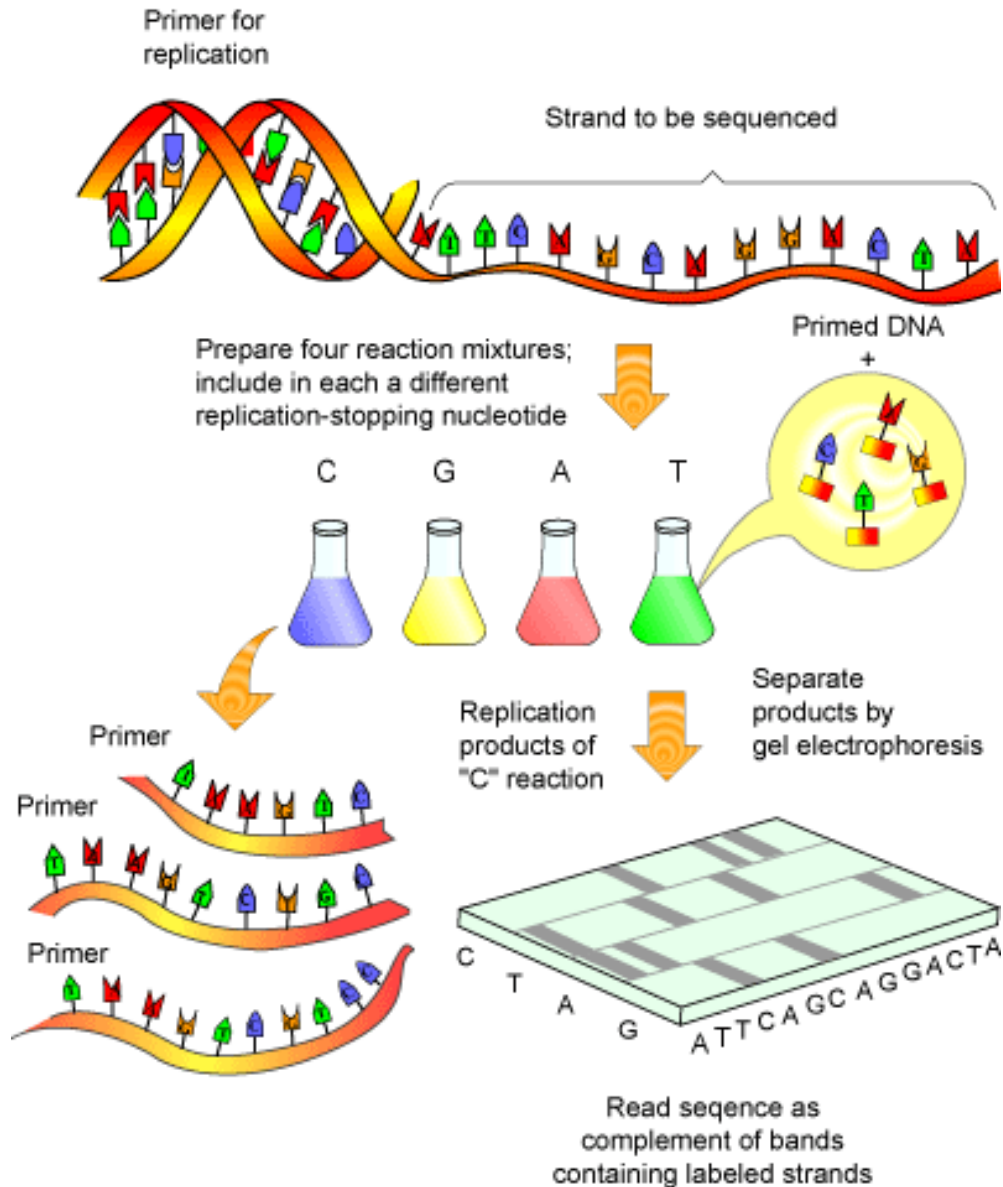


# SEQUENCING TECHNOLOGIES -- Let's talk about nucleotide sequencing

3 sequencing generations

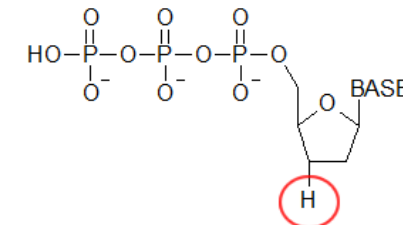
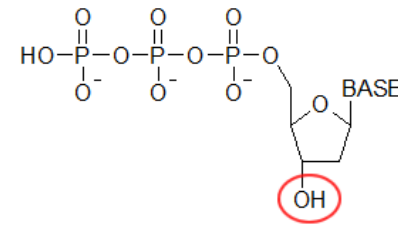


# Sanger sequencing or chain termination method



Based on PCR amplification → It is known as the synthesis method

Uses dideoxynucleotide triphosphate (ddNTPs) lacking 3'-OH required for phosphodiester bond



In the beginning, nucleotides were radio-labeled (<sup>35</sup>S dCTP)

To identify the last nucleotide incorporated, 4 reactions were needed, and fragments were separated by an acrylamide gel with different nucleotide resolutions depending on the concentration.

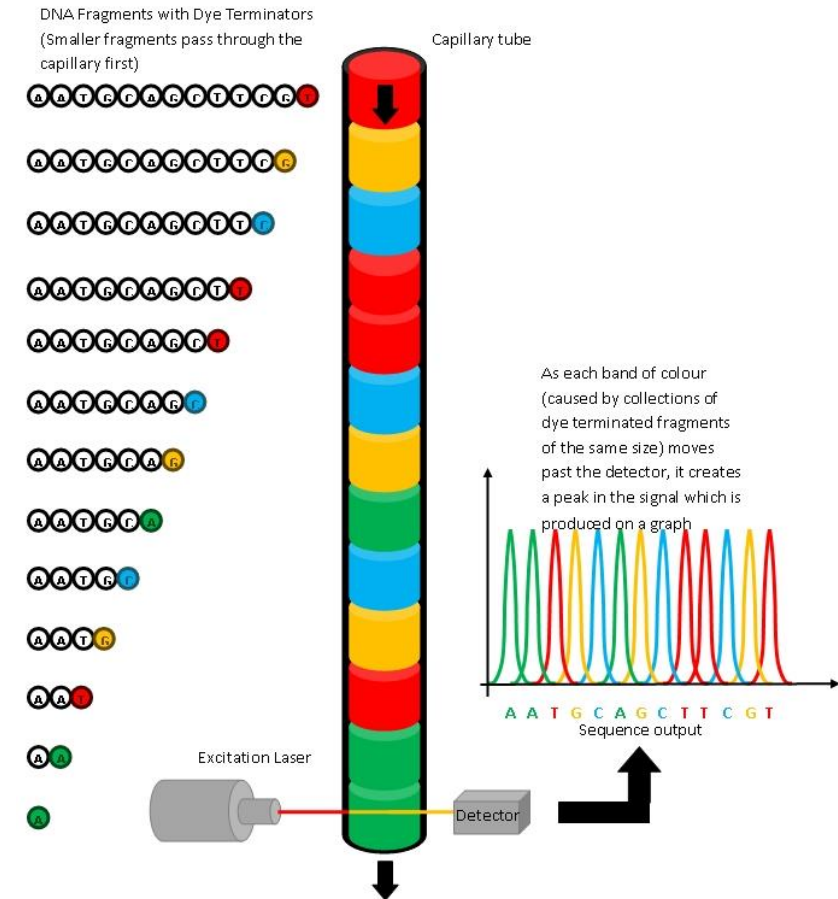
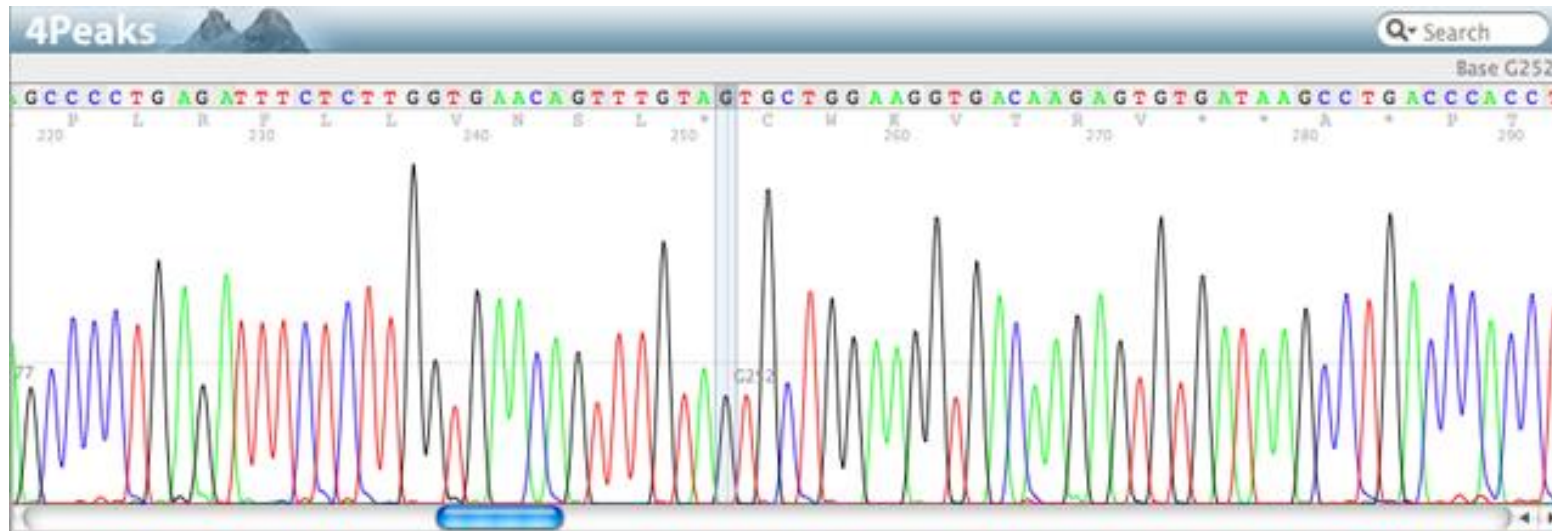


# Automatized sanger sequencing with fluorescent terminators (current)

A different fluorescence to each terminator allows for running just one reaction.

Fragments separation perform on capillary electrophoresis coupled to an electronic detector

The detector is connected to a computer and the signal is read on a chromatogram (electropherogram)



# Shotgun sanger sequencing

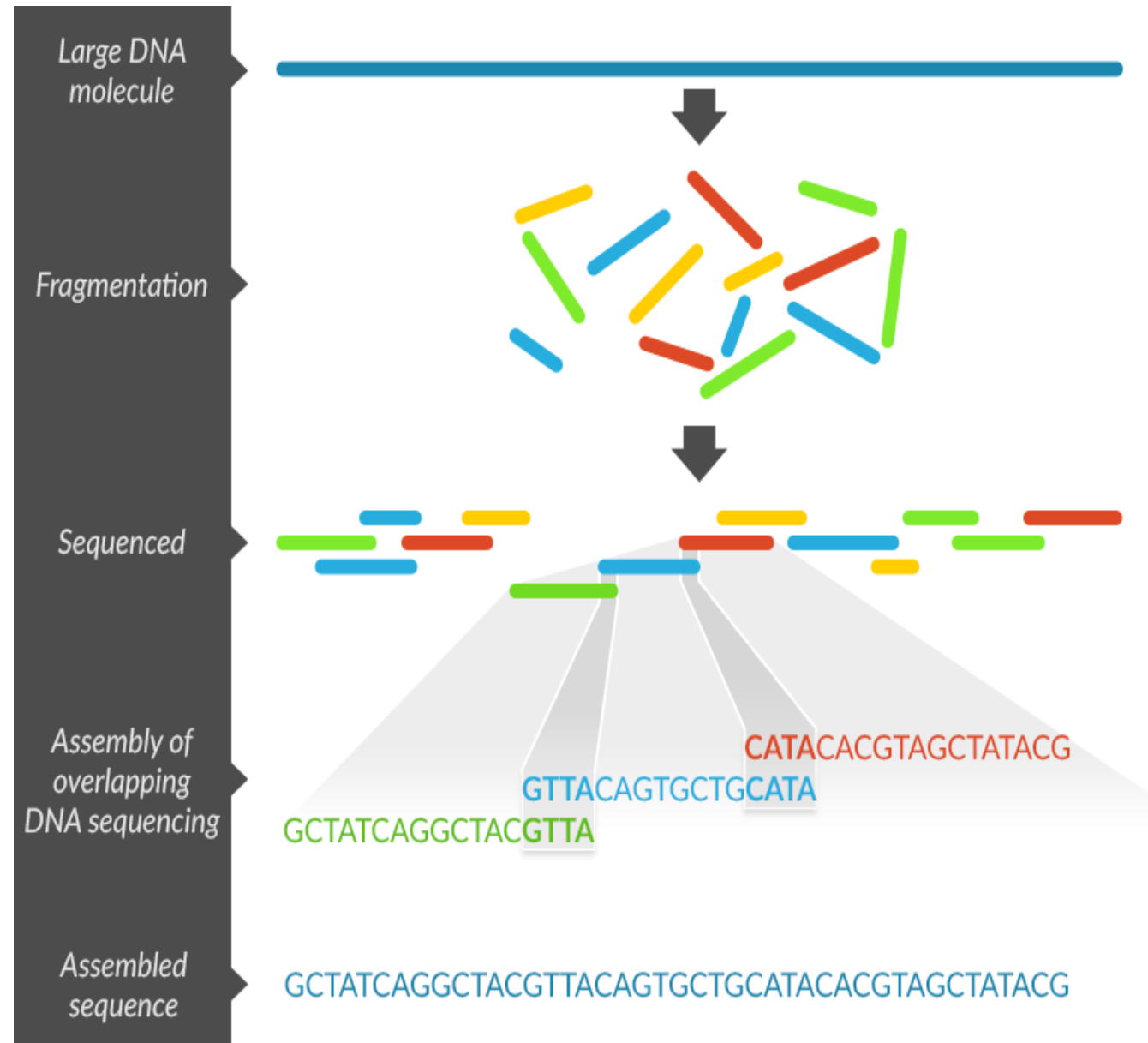
Sequencing Range: 100-1000bp

It was implemented to sequence long DNA fragments

The principle is based on random fragmentation of the long DNA piece.

Those **fragments** are sanger sequenced and posteriorly assembled based on overlapping sections to reconstruct the original DNA molecule.

For **RNA-seq**, transcripts will be randomly fragmented on small pieces (300bp)





# Key Sequencing Concepts



Two concepts/features of sequencing that must be defined before we continue

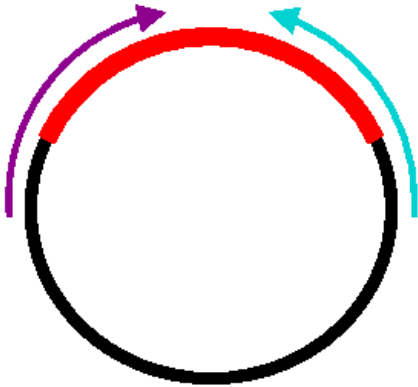


- > What do we sequence?
- > Sequencing Depth and Coverage

# What do we sequence in RNA-seq experiments?

A fragment of DNA with two adapters sequences attached, one at each end, which are used for the sequencing reaction

On shotgun sequencing, the known regions around the insert will play as adapters



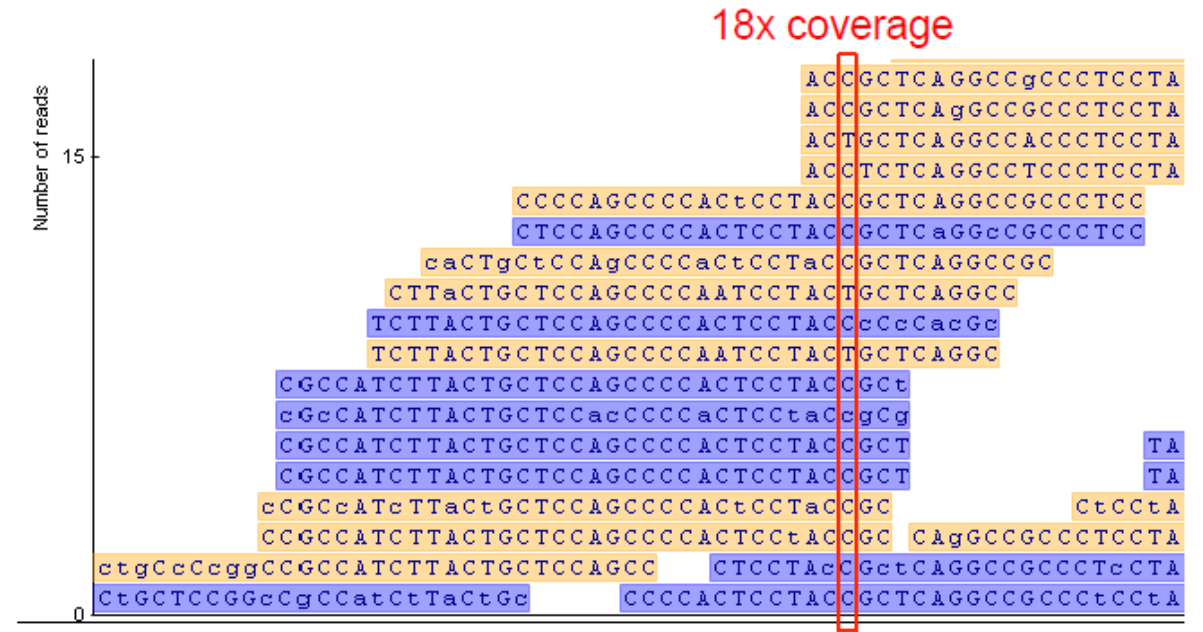
On next-generation sequencing, this sequences must be compatible with the sequencing technology



# Sequencing Depth and Read Coverage

**Sequencing depth** is the number of reads (so, DNA fragments) sequenced. This number must be defined before sequencing according to the transcriptome/genome size. Organisms with big genomes will require more sequenced reads than organisms with small genomes.

**Read Coverage** is the number of sequences covering a specific region of the genome or transcriptome.



partly overlapping sequencing reads result from the multiple templates being sequenced across the flow cell

# Read Coverage for RNAseq

**Read Coverage** for RNA-seq is often calculated in terms of sequence depth by sample and will depend on your objective.

## Recommendations for eukaryotes organisms (by Illumina):

- For **quick snapshots** of highly expressed genes, **25 millions reads** by samples is enough
- For a **global view of gene expression** and some information on alternative splicing, typically **60 million reads** by sample will work (most of the published works are using this sequence depths).
- **In-depth transcriptome exploration** will need a minimum of **200 million reads** is need
- **Targeted sequencing** **3 millions**

## Read length recommendations:

- mRNA profiling → SE 75bp
- Transcriptome assembly → PE 75bp or 100bp
- Small RNA → SE 50bp

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

## Evaluating the Impact of Sequencing Depth on Transcriptome Profiling in Human Adipose

Yichuan Liu, Jane F. Ferguson, Chenyi Xue, Ian M. Silverman, Brian Gregory, Muredach P. Reilly, Mingyao Li

Published: June 24, 2013 • <https://doi.org/10.1371/journal.pone.0066883>

Article	Authors	Metrics	Comments	Media Coverage
Abstract				
Introduction				
Materials and Methods				
Results				
Discussion				
Supporting Information				
Author Contributions				
References				
Reader Comments				
Figures				

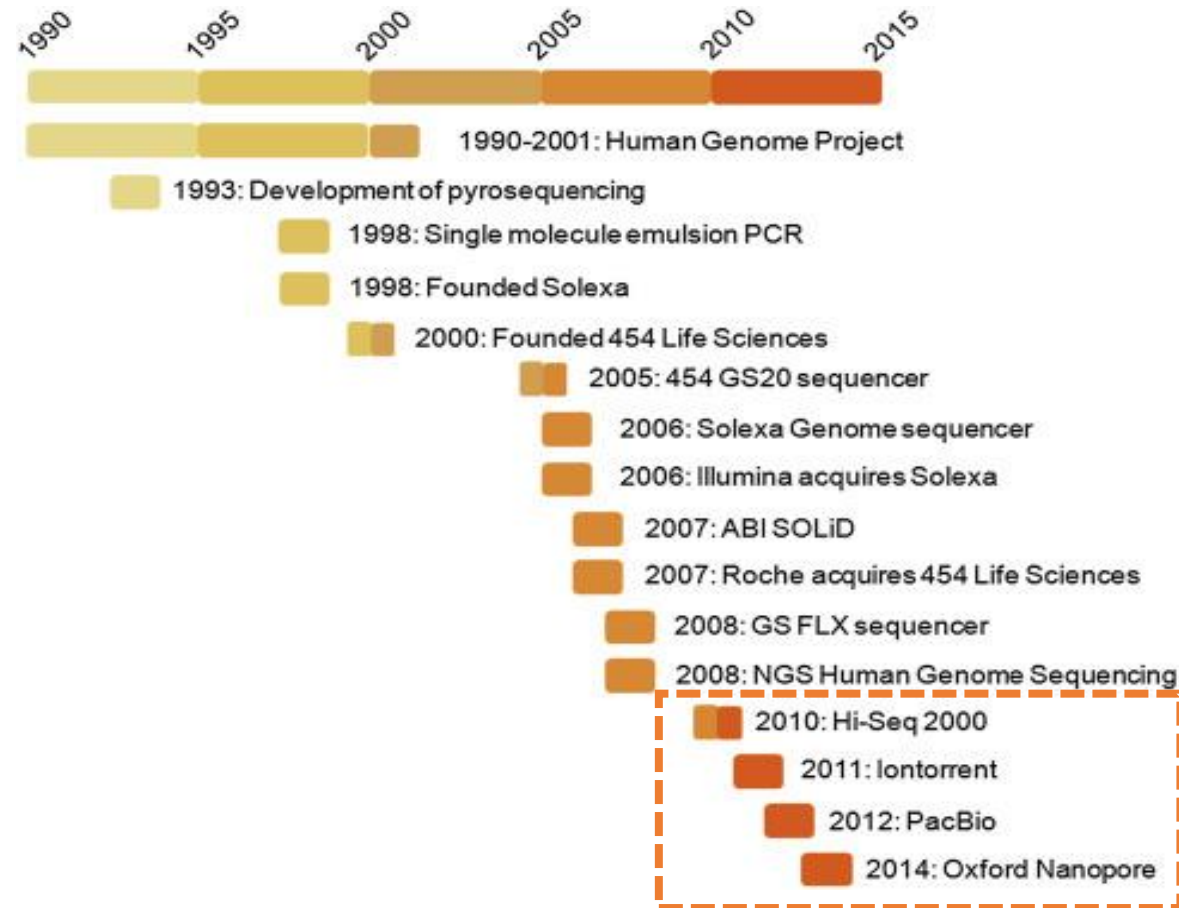
Abstract

Recent advances in RNA sequencing (RNA-Seq) have enabled the discovery of novel transcriptomic variations that are not possible with traditional microarray-based methods. Tissue and cell specific transcriptome changes during pathophysiological stress in disease cases versus controls and in response to therapies are of particular interest to investigators studying cardiometabolic diseases. Thus, knowledge on the relationships between sequencing depth and detection of transcriptomic variation is needed for designing RNA-Seq experiments and for interpreting results of analyses. Using deeply sequenced Illumina HiSeq 2000 101 bp paired-end RNA-Seq data derived from adipose of a healthy individual before and after systemic administration of endotoxin (LPS), we investigated the sequencing depths needed for studies of gene expression and alternative splicing (AS). In order to detect expressed genes and AS events, we found that ~100 to 150 million (M) filtered reads were needed. However, the requirement on sequencing depth for the detection of LPS modulated differential expression (DE) and differential alternative splicing (DAS) was much higher. To detect 80% of events, ~300 M filtered reads were needed for DE analysis whereas at least 400 M filtered reads were necessary for detecting DAS. Although the majority of expressed genes and AS events can be detected with modest sequencing depths (~100 M filtered reads), the estimated gene expression levels and exon/intron inclusion levels were less accurate. We report the first study that evaluates the relationship between RNA-Seq depth and the ability to detect DE and DAS in human adipose. Our results suggest that a much higher sequencing depth is needed to reliably identify DAS events than for DE genes.

Figures

# SEQUENCING TECHNOLOGIES -- Let's talk about nucleotide sequencing

3 sequencing generations






# Next Generation sequencing – What is Second Generation?

- It is **massive sequencing** which is characterized by high depth (millions of fragments sequenced at once)
- Compared with Sanger, it is  $>100x$  cheaper and faster
- During the development of this technology appear several platforms, however, **Illumina** is the gold standard now.

Instrument	Method	Read Length	Yield	Quality	Value
Illumina	synthesis + fluorescence	250	++++	+++++	++++
SOLiD	ligation + fluorescence	75	++++	+++	+++
Roche 454	non-term NTP + luminescence	600	+	++++	++

# Illumina Sequencers



Sequencing System	iSeq™	MiniSeq™	MiSeq®	NextSeq®	HiSeq®	HiSeq® X	NovaSeq®
					4000	Five/Ten	6000
<b>Output per run</b>	1.2 Gb	7.5 Gb	15 Gb	120 Gb	1.5 Tb	1.8 Tb	1 Tb - 6 Tb <sup>1</sup>
<b>Instrument price</b>	\$19.9K	\$49.5K	\$99K	\$275K	\$900K	\$6M <sup>2</sup> /\$10M <sup>2</sup>	\$985K
<b>Installed base<sup>3</sup></b>	NA	~600	~6,000	~2,400	~2,300 <sup>4</sup>		~285

1. Output per run for the S1, S2 and S4 flow cells equal 1 Tb, 2 Tb and 6 Tb, respectively assuming two flow cells per run
2. Based on purchase of 5 and 10 units for HiSeq X Five and HiSeq X Ten, respectively
3. Based on end of fiscal year 2017
4. Combined HiSeq family

# Video of how Illumina sequencing works

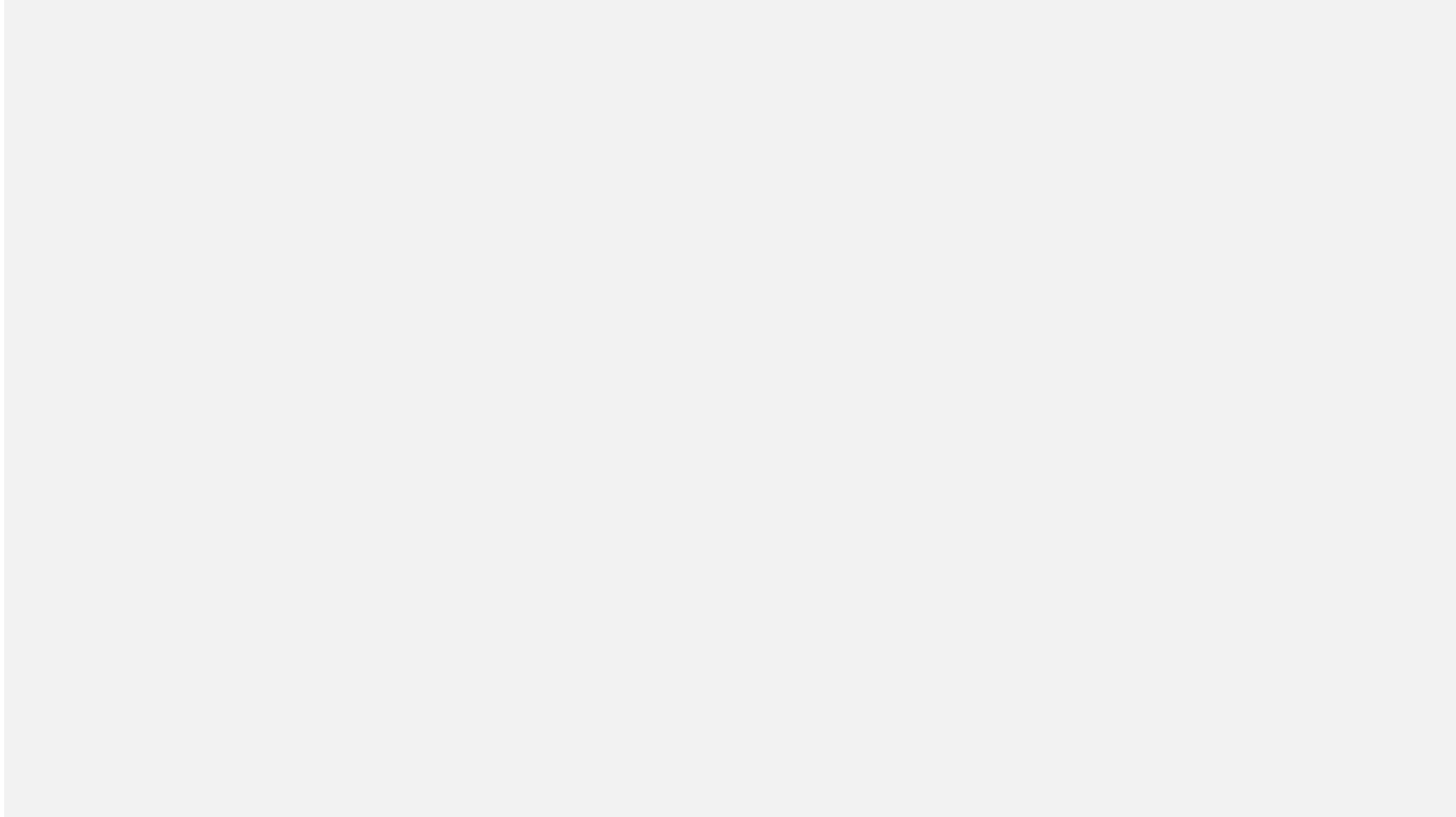
# Third Generation sequencing

It was developed based on the **need for larger reads**, which help resolve complex genomic structures such as repetitive elements, copy number alterations, alternative splicing, and structural variations.

	PacBio <sup>1</sup>		Oxford Nanopore <sup>2</sup>	
Instrument Specifications	RS II (P6-C4)	Sequel	MinION	PromethION
Average read length	10 – 15 kb	10 – 15 kb	Variable (up to 900 kb) <sup>3,4</sup>	*
Error rate	10 – 15 %	10 – 15 %	5 – 15 % <sup>4,5</sup>	*
Output	500 Mb – 1 Gb	5 Gb – 10 Gb	~5 Gb <sup>4</sup>	*
# of reads	~50k	~500k	Variable (up to 1M) <sup>6,7</sup>	*
Instrument price/Access fee <sup>a</sup>	\$700k	\$350k	\$1000 <sup>8</sup>	\$135k bundle <sup>9</sup>
Run price	~\$400	~\$850	\$500-\$900 <sup>7</sup>	*

# Third-Generation sequencing— PacBio Platform

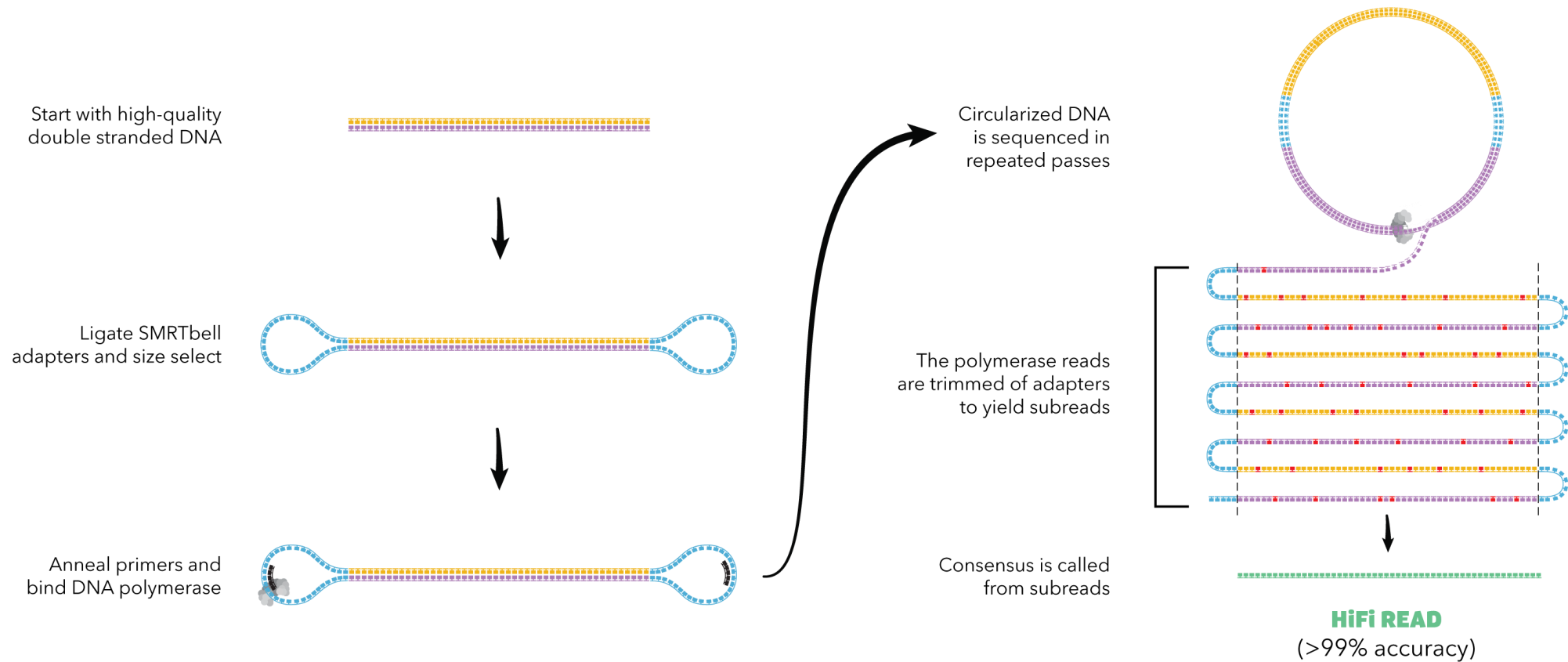
The third-generation sequencing is also known as single-molecule real-time (SMRT) sequencing. In the case of PacBio technology, it is based on the use of zero-mode waveguides (ZMWs).





# PacBio utilize circular consensus sequences to improve accuracy

Circular consensus sequencing involves **multiple passes** of the sequencing template. The system records multiple reads of the same circular DNA molecule, generating a consensus sequence by aligning these reads. This process helps correct errors that may occur in individual reads, improving overall accuracy.





We got a big amount of data...so, what now??

# Why do we care about repositories??



## ✓ Open Source and Open Science Advocacy:

- Encourage transparency in scientific research by sharing raw data with the scientific community.

## ✓ Journal Publication Requirement:

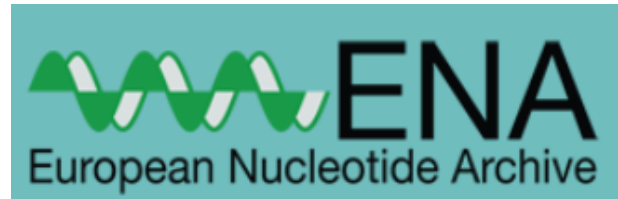
- Reputable scientific journals often require the release of data before accepting and publishing a research paper.
- Fulfilling this requirement demonstrates a commitment to the highest standards of scientific integrity.

## ✓ Global Scientific Collaboration:

- Acknowledge the role of shared biological data as a valuable resource for the global scientific community.
- By contributing to a collective pool of information, researchers worldwide can access and utilize this data for diverse scientific endeavors.

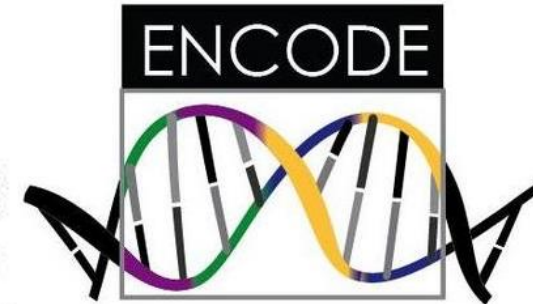
# Some well-know data repositories

- ENA / European Nucleotide Archive
- SRA /Sequence Read Archive
- GEO /Gene Expression Omnibus
- BioSD / BioSamples Database
- TIARA /Total Integrated Archive of short-Read and Array



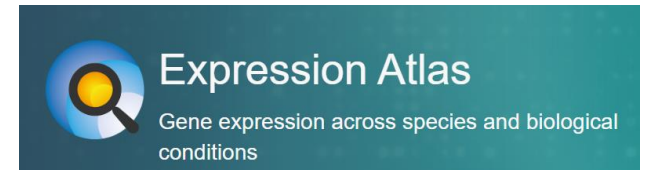
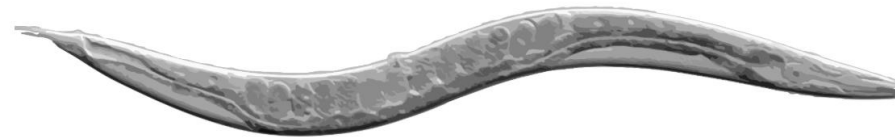
RNA  
Seq  
Atlas

National Human Genome Research Institute

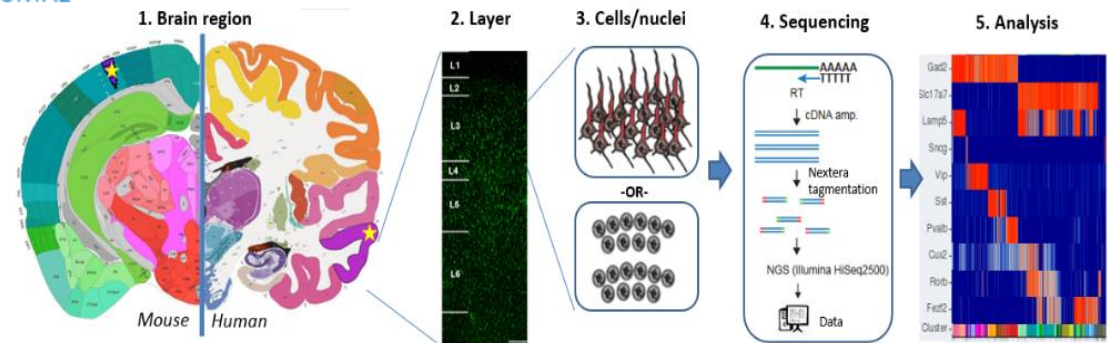


A Cell Atlas of Worm

The *C. elegans* transcriptome at single cell resolution



ALLEN BRAIN ATLAS  
DATA PORTAL

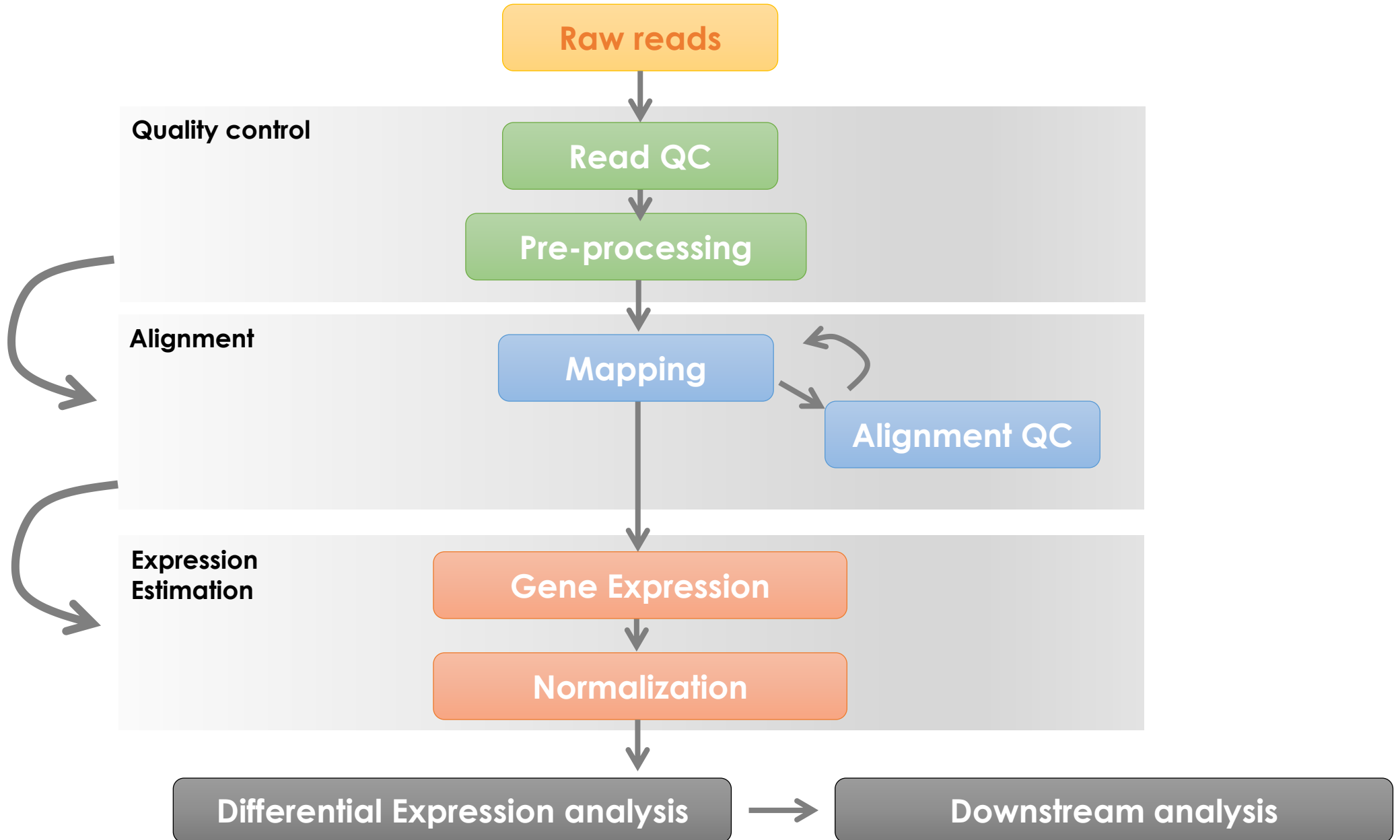


Check complete list here: <https://www.nature.com/sdata/policies/repositories>

# Data Analysis General Pipeline

# Data Analysis General Pipeline

3 main steps



# Primary format for NGS data

## FASTQ (FASTA + Quality)

- Format that associates sequences with quality value by nucleotide base

## BAM

- Format for aligned and not aligned sequences. It is binary (compressed)

## SAM

- Format for aligned and not aligned sequences. It is text (extended and human readable)

## BED

- Format describing one feature (CDS, Exon, Intron, UTR, etc..) per line

## GFF/GTF (gene annotation)

- Format describing one feature (CDS, Exon, Intron, UTR, etc..) per line, often contains more information than bed format



# Formato FASTA: Componentes

Start  
symbol

Sequence ID  
(no spaces)

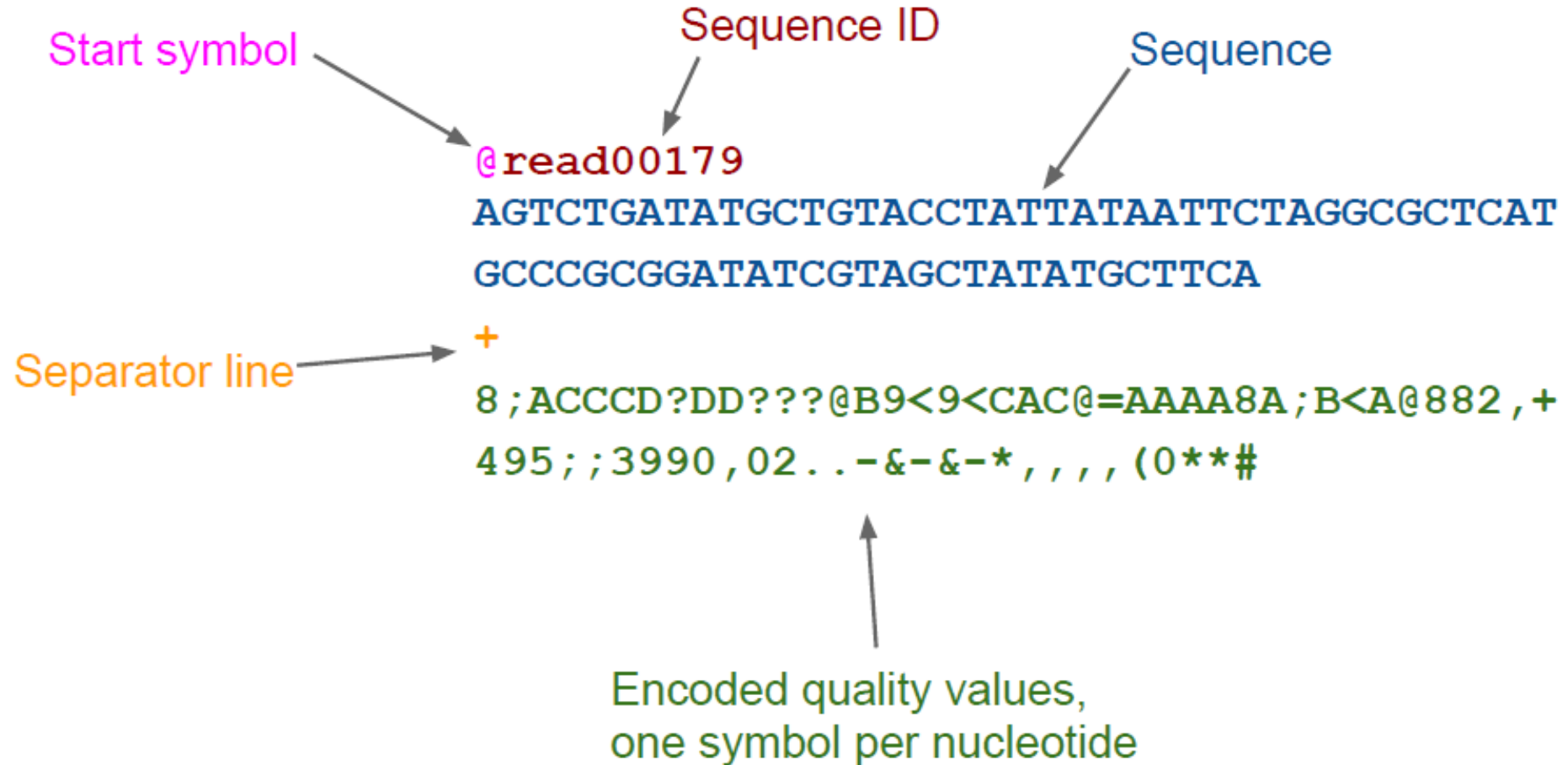
Sequence description  
(spaces allowed)

```
>dnaA chromosomal replication initiator protein DnaA  
MSLSLWQQCLARLQDELPATEFSMWIRPLQAELSDNTLALYAPNRFVLDWVRDKYL  
EALRDLLALQEKLVTIDNIQKTVAEYYKIKVADLLSKRRSRSVARPRQMAMALAKE  
LLHAVGNGIMARKPNAKVVYMHSERFVQDMVKALQNNAIIEEFKRYYRSVDALLIDD  
FSLPEIGDAFGGRDHTTVLHACRKIEQLREESHDIKEDFSNLIRTLSS
```

The sequence  
(usually 60 letters per line)

# Formato FASTQ

Is an extension of the **FASTA format carrying quality values** associated with each base



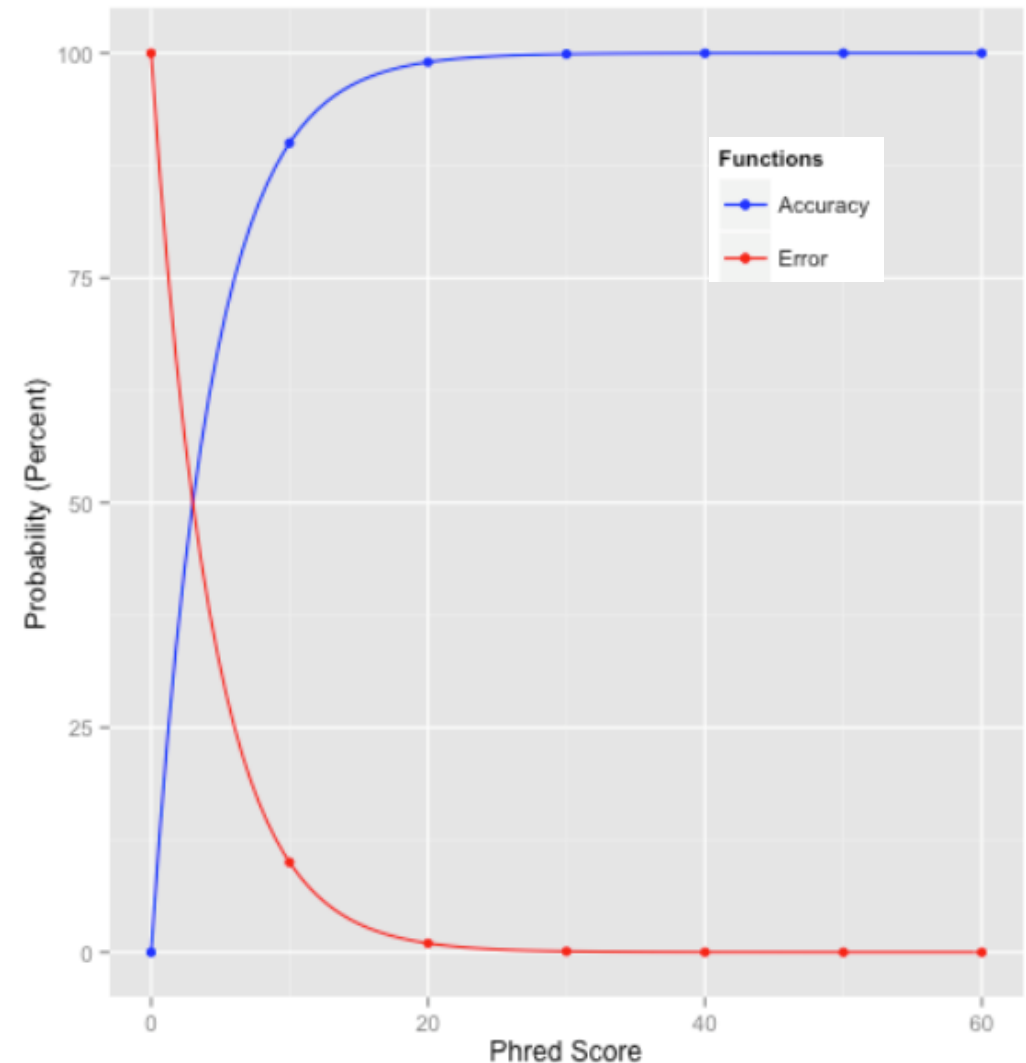
# Quality Value are expressed by phred score

Sequencing quality scores measure the probability that a base is called incorrectly. With sequencing by synthesis (SBS) technology, each base in a read is assigned with a quality score by a phred algorithm.

The sequencing quality score of a given base,  $Q$ , is defined by the following equation:

$$Q = -10\log_{10}(e)$$

where  $e$  is the estimated probability of the base call being wrong.



# How quality value are generated

Illumina quality scores are calculated for each base call in a **two-step** process:

1. Quality predictor values are **observable properties of clusters** from which base calls are extracted.
  - a) **intensity profiles**
  - b) **signal-to-noise ratios**
2. A **quality model**, also known as a **quality table or Q-table**, lists combinations of **quality predictor values** and relates them to corresponding quality scores.

Table 1: Q-Scores and Error Probabilities

Quality Score	Error Probability
Q40	0.0001 (1 in 10,000)
Q30	0.001 (1 in 1,000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)

# Quality value code

The quality values are coded on letters and symbols

The coding system is the [ASCII](#) (American Standard Code for Information Interchange)

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMN O PQRSTU VWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|         |         |         |         |
33          59      64      73          104          126
0.....26...31.....40
          -5....0.....9.....40
           0.....9.....40
            3.....9.....41
0.2.....26...31.....41
```

S - Sanger Phred+33, raw reads typically (0, 40)  
X - Solexa Solexa+64, raw reads typically (-5, 40)  
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)  
J - Illumina 1.5+ Phred+64, raw reads typically (3, 41)  
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)  
(Note: See discussion above).  
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

!"#\$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ				
Q0	Q10	Q20	Q30	Q40
<i>bad</i>	<i>maybe</i>	<i>ok</i>	<i>good</i>	<i>excellent</i>



# How data quality impacts on my results?

## Quality recommendations

### Experimental design

Minimize variability in your samples  
Have at least 3 biological replicates

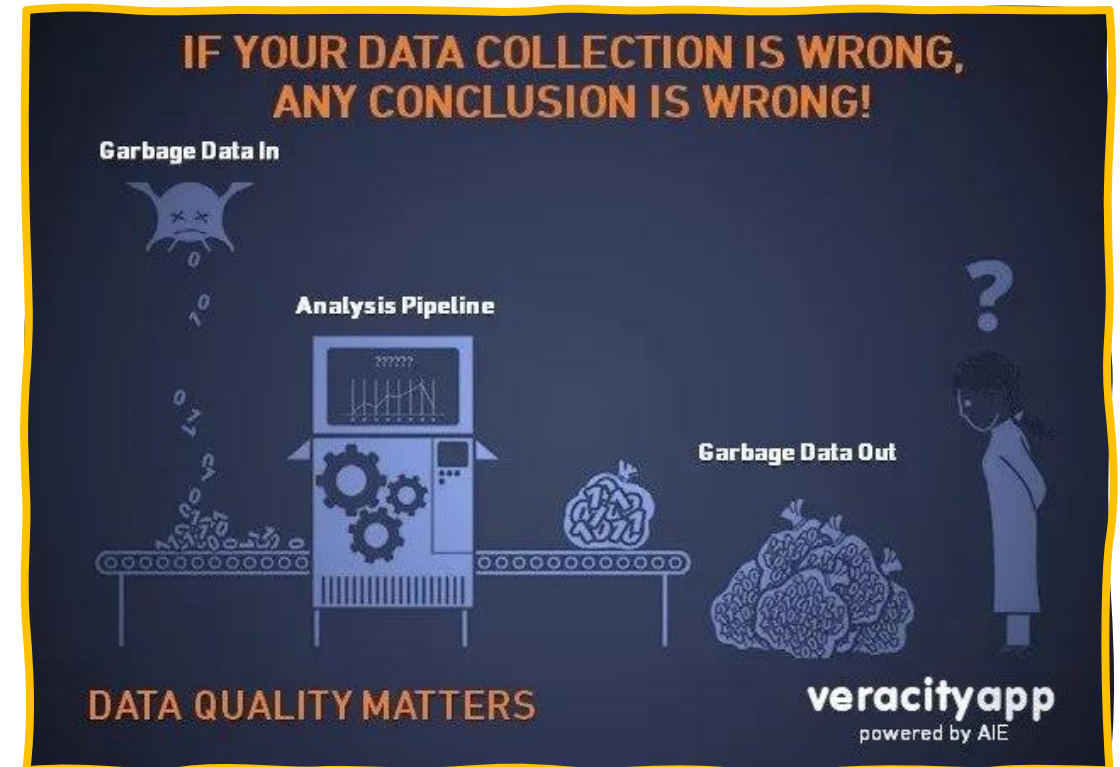
### Samples

Avoid degradation, use an RNA Stabilization Reagent  
Integrity of RNA must be  $RIN \geq 8$   
Samples must be DNA-free (DNase treatment)  
Select appropriate library prep method for RNA available

### Sequencing

Consider sequencing depth based on exp objective  
Run read QC controls, reproducibility test and Mapping QC.

## The GIGO (GARBAGE IN – GARBAGE OUT) PARADIGM





# Quality issues on sequencing data

The sequences might contain errors such as:

- > **Duplicated Reads** (Low complexity libraries and PCR duplicares)
- > **Reading into the adapters** (short fragments in comparison with read length)
- > **Error Indel** (Deletion or insertion of a base during sequencing )
- > **Undetermined Base** (base calling base was uncertain and it is replaced by an N)
- > **Substitution errors** (wrong base calling base)

Rare

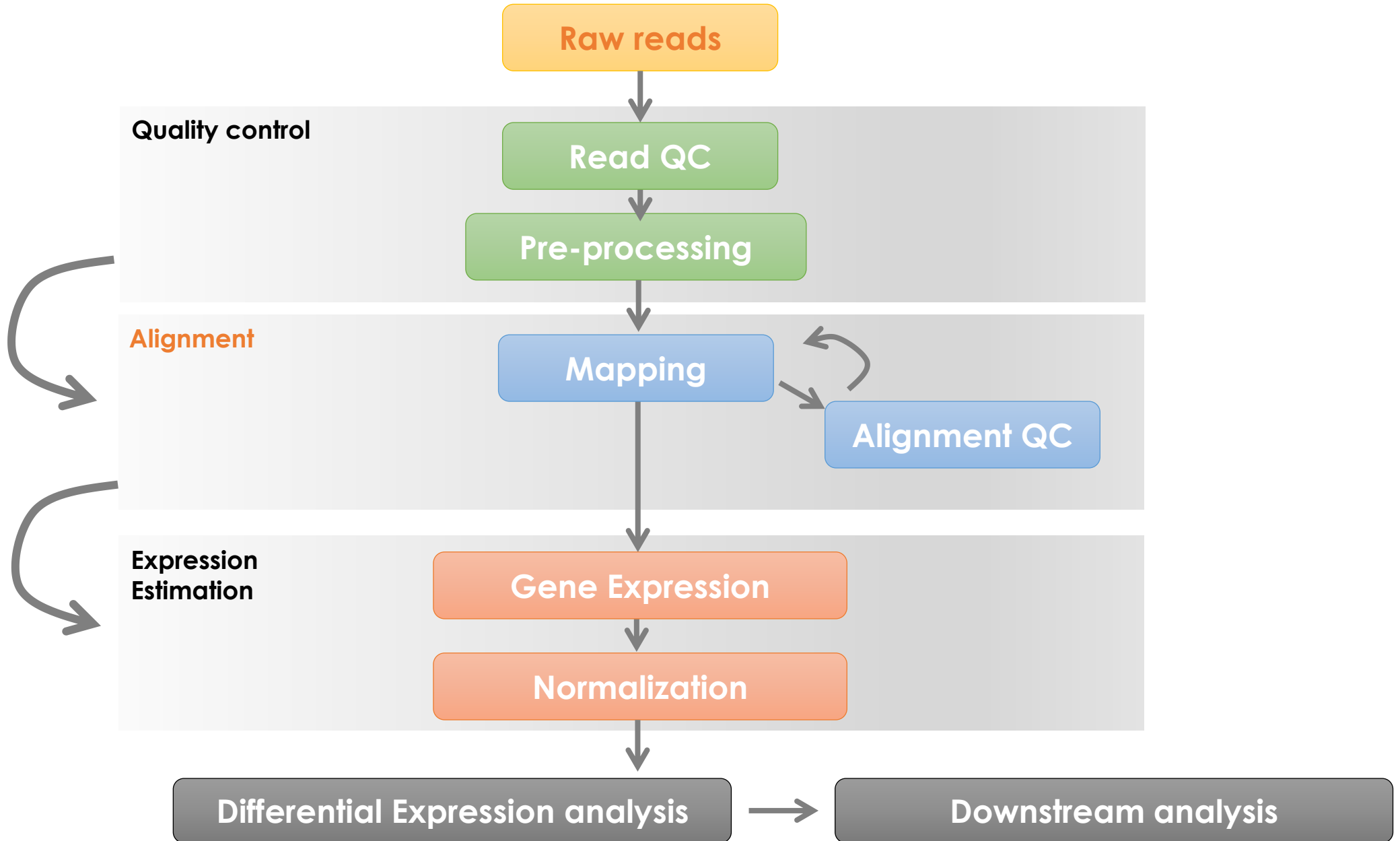


Common

ALL OF THEM CAN BE RECTIFIED OR ELIMINATED THROUGH BIOINFORMATICS ANALYSIS

# Data Analysis General Pipeline

3 main steps



# RNA-sequencing alignment is challenging

- It requires high computational power (Millions of reads being analyzed)
- Intron presence (mapping to the reference genome — alignment must be splice aware)
- Inefficient alignment (Genetic variants, repeat sequences and contaminations)

# Sequence alignment: Two Types

We must align reads to find the correct position in the reference genome from where they were originated.

## Global Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
  ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

- Performed **end-to-end** for both sequences **introducing gaps** if needed
- Aligns all the bases both for query and target
- Ideal for related sequences with similar length

## Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
  ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
```

Query Sequence

```
5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

- Perform alignments only in the **most similar regions**
- **Align pieces** of query and target sequences (substrings/subsequences)
- Can provide more than one alignment

# Features of alignment strategies

- **Genome alignment + Gene model assembly (splice aware alignment)**

Positive:

Detection of new transcripts

Negative:

Alignment is difficult

Insert size and inner distance are difficult to infer due to the intron presence

- **Transcriptome alignment**

Positive:

Do not require spliced alignment

Simplifies the expression estimation by isoform

Insert size and inner distance are informative

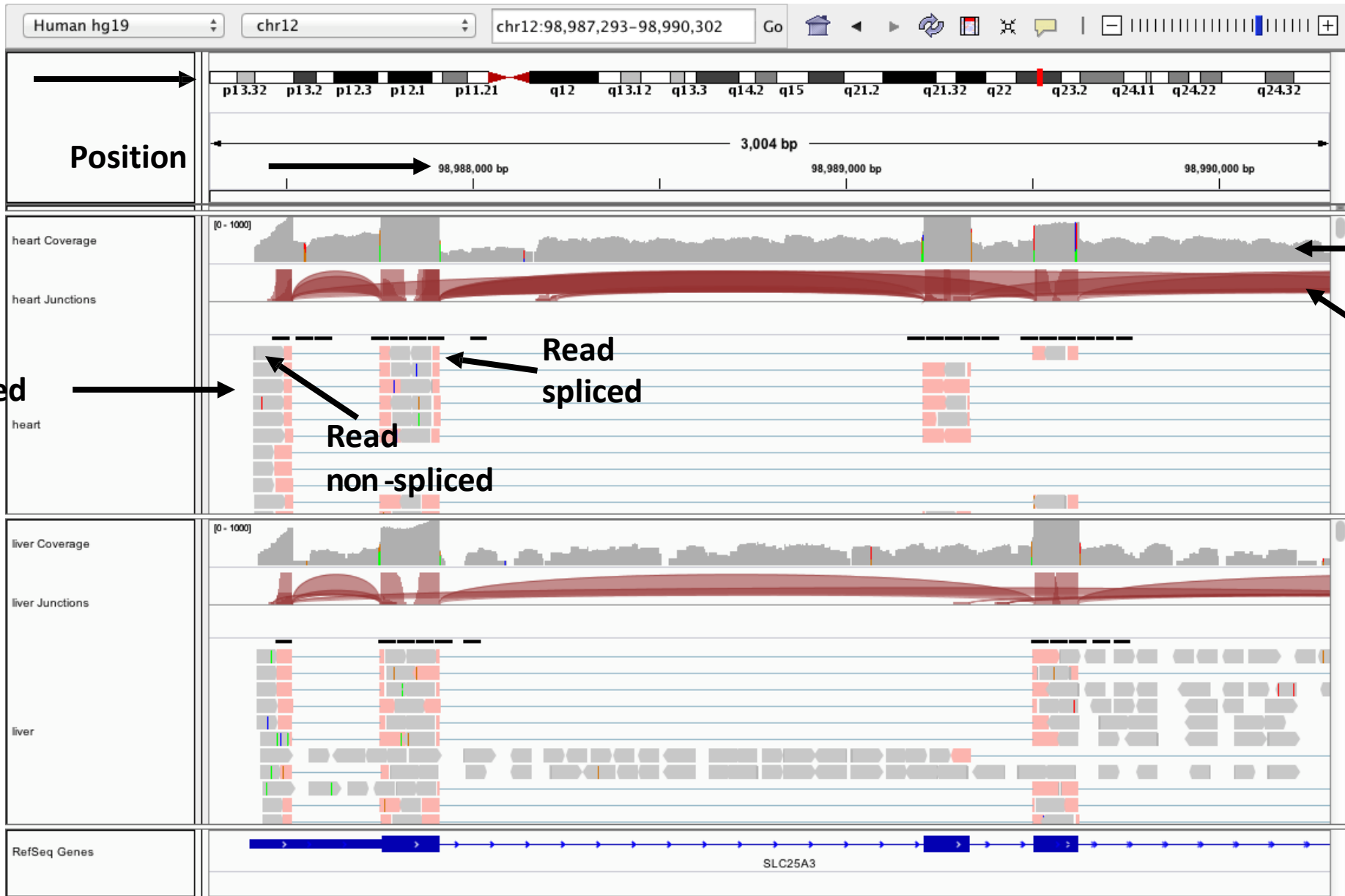
Negative:

Depends on the gene model quality

No discovery of new transcripts 😞

# Alignment visualization with IGV

Chromosome



Read Aligned

Read Coverage

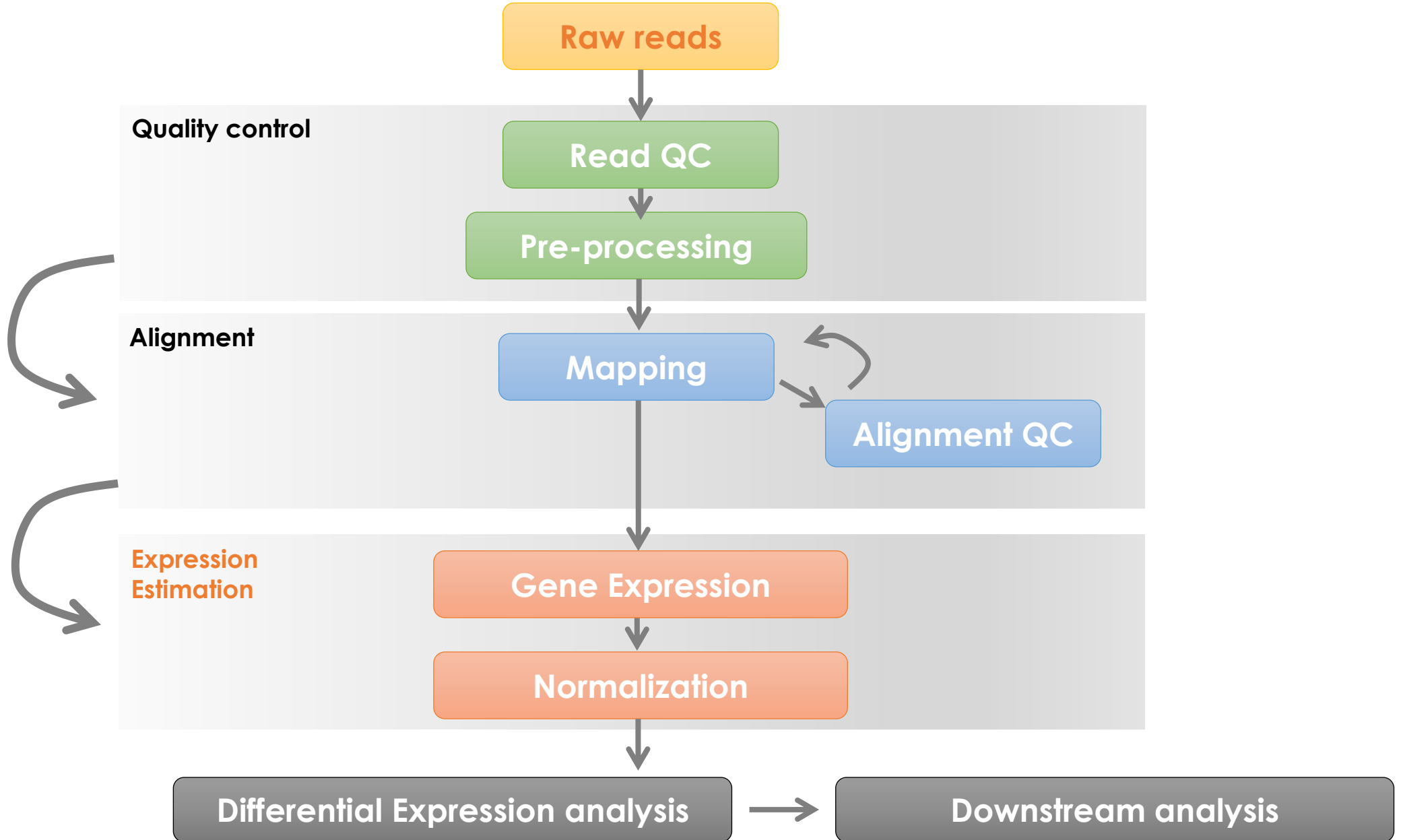
Splicing Junctions

Read non-spliced

Read spliced

# Data Analysis General Pipeline

3 main steps

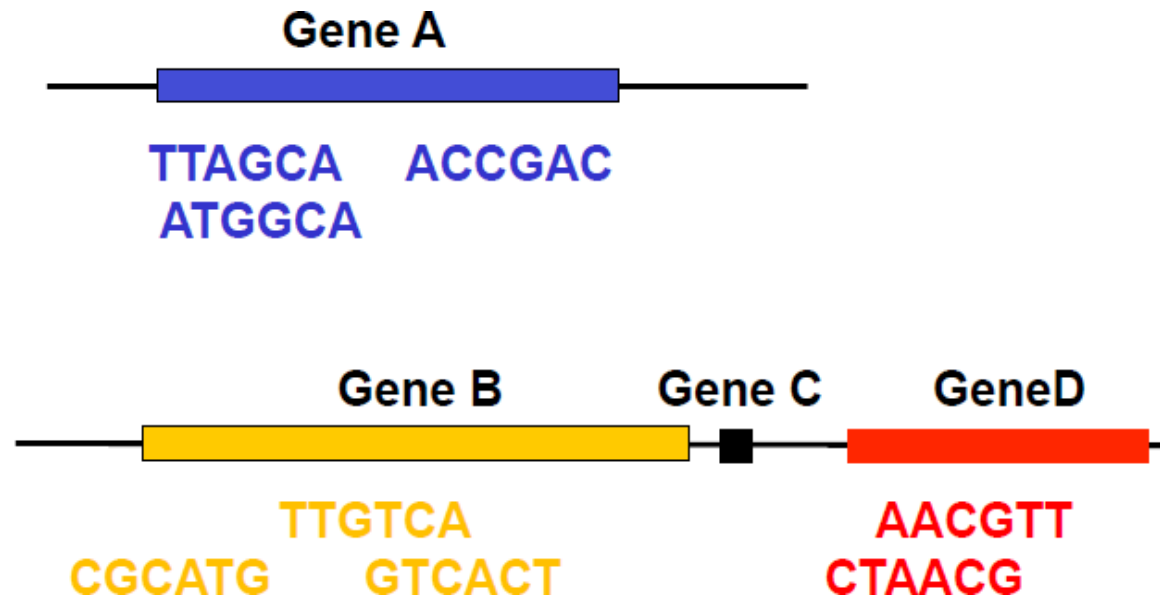




# Expression level estimation

Is a two steps process:

- 1) Count aligned reads to genomic features (exons, genes, transcripts)
- 2) Normalize counts



Count Table

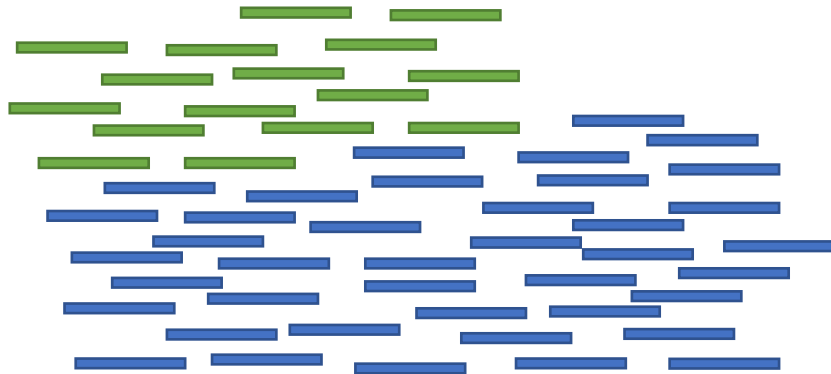
Gene ID	Sample1
A	3
B	3
C	0
D	2

Normalization is the process of scaling raw counts values to make them comparable

# Why do we care about normalization?

Given the number of reads in the draw, **which isoform is more expressed?**

Reads



Isoforms



The blue isoform **has more reads, but it is also longer than the green one.**

As the library preparation includes fragmentation, long isoforms will generate a greater number of reads

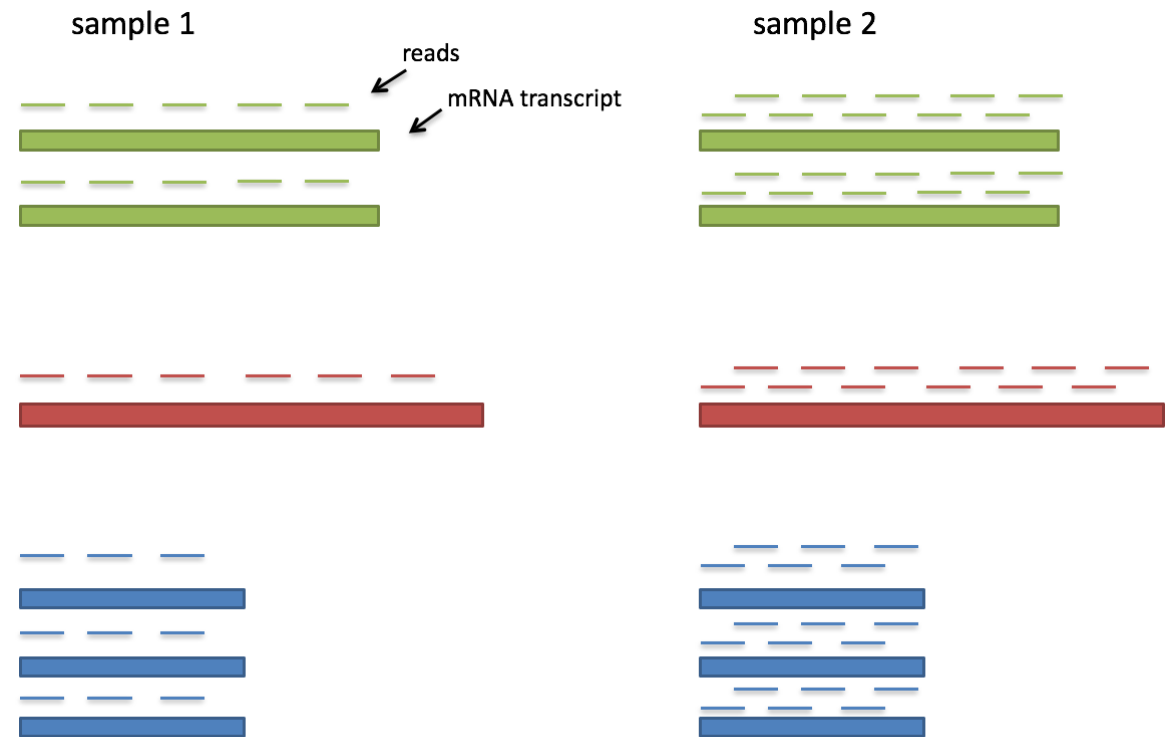
# Principal factors included on the normalization

- Global Sequence Coverage

In this case, it is the total number of mapped reads to the genome or transcriptome. This is very important to normalize when comparing expressions **between samples**.

- Genes Length

- Transcriptome composition



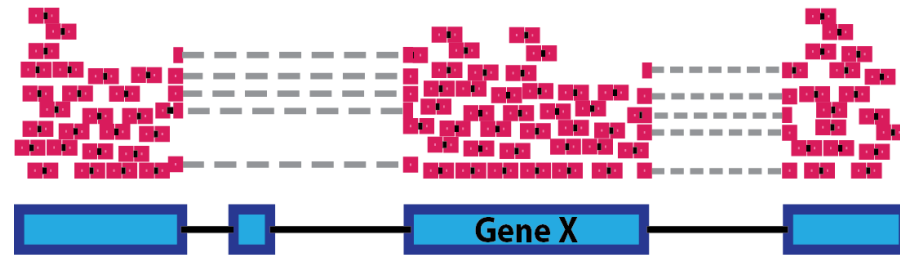
# Principal factors included on the normalization

- Global Sequence Coverage

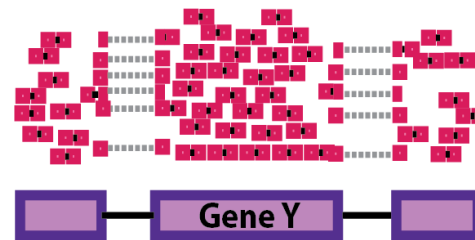
When comparing expression **between genes on the same samples** (let's say comparing the expression of gene A with gene B) the length will tend to overestimate the expression of long genes.

- Genes Length

## Sample A Reads



- Transcriptome composition



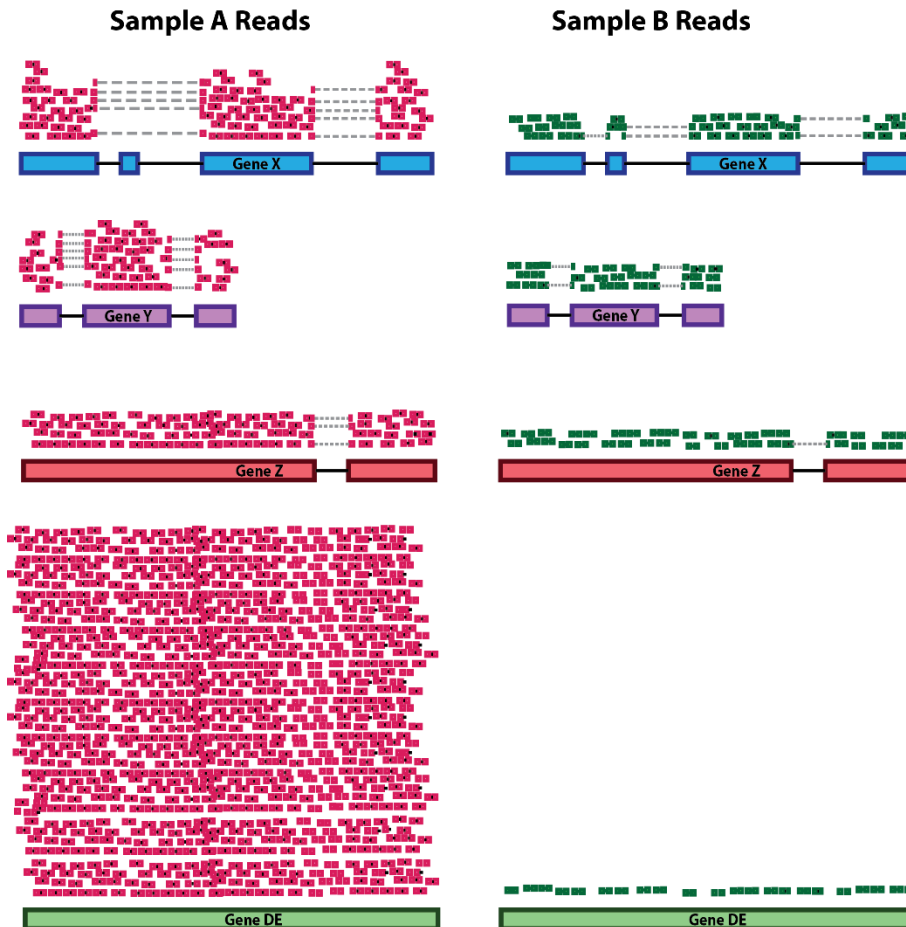
# Principal factors included on the normalization

- Global Sequence Coverage

It is highly recommended that when **comparing two samples with different backgrounds** (e.i. different cell types, different genetic backgrounds, etc) the composition of the transcriptome must be taken in account.

- Genes Length

- Transcriptome composition



# Most popular normalization methods

Counts per million (CPM)

Fragments per kilobase per million reads (FPKM o RPKM)

Transcripts per million reads (TPM)

Trimmed Mean of M-value (TMM - EdgeR)

DESeq's Median of ratios

# Normalization methods (I)

Normalization method	Description	Accounted factors	Recommendations for use
<b>CPM</b> (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same sample group; <b>NOT for within sample comparisons or DE analysis</b>
<b>TPM</b> (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; <b>NOT for DE analysis</b>
<b>RPKM/FPKM</b> (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; <b>NOT for between sample comparisons or DE analysis</b>

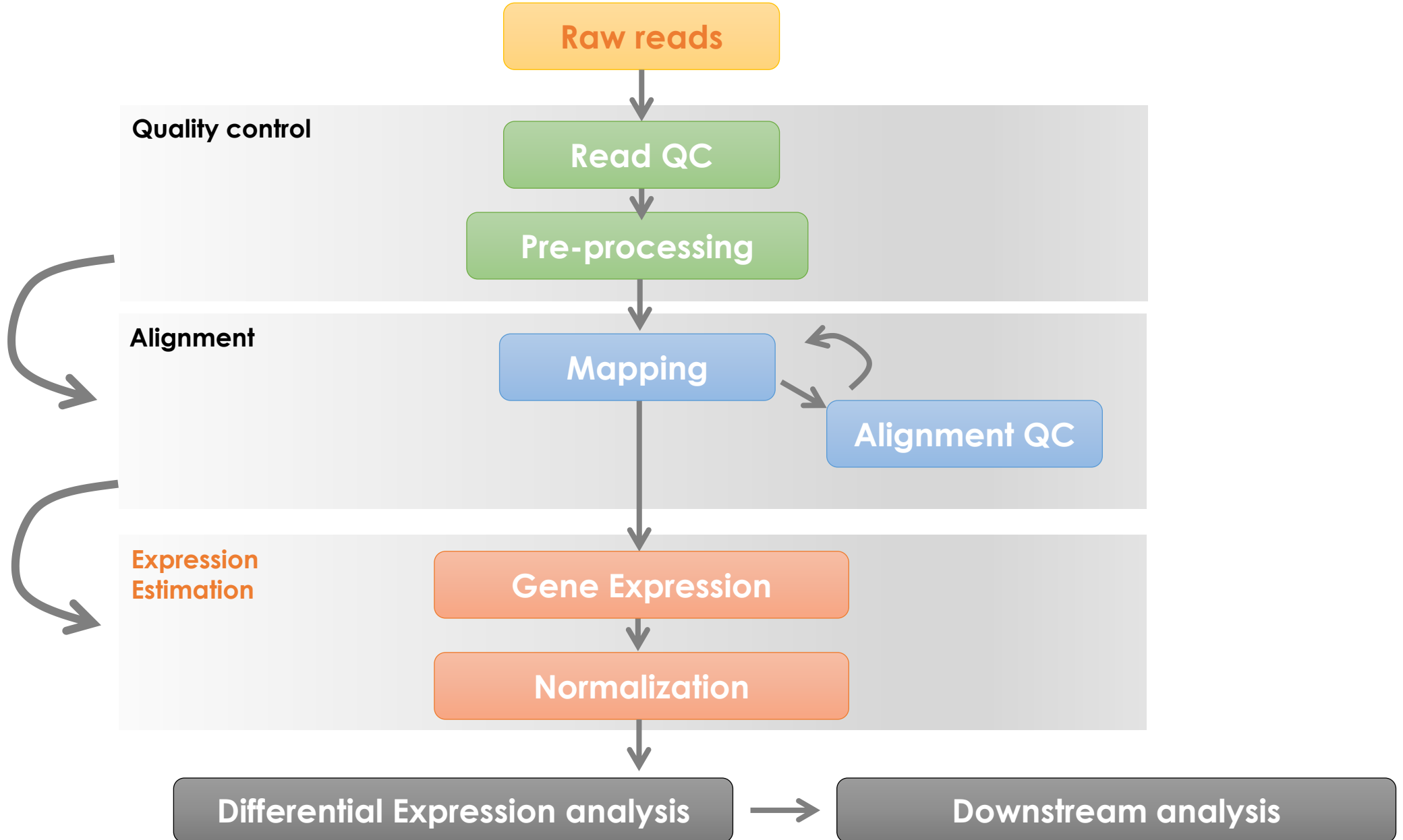


# Normalization methods (II)

Normalization method	Description	Accounted factors	Recommendations for use
DESeq2's <b>median of ratios</b> [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for <b>DE analysis; NOT for within sample comparisons</b>
EdgeR's <b>trimmed mean of M values (TMM)</b> [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for <b>DE analysis</b>

# Data Analysis General Pipeline

3 main steps

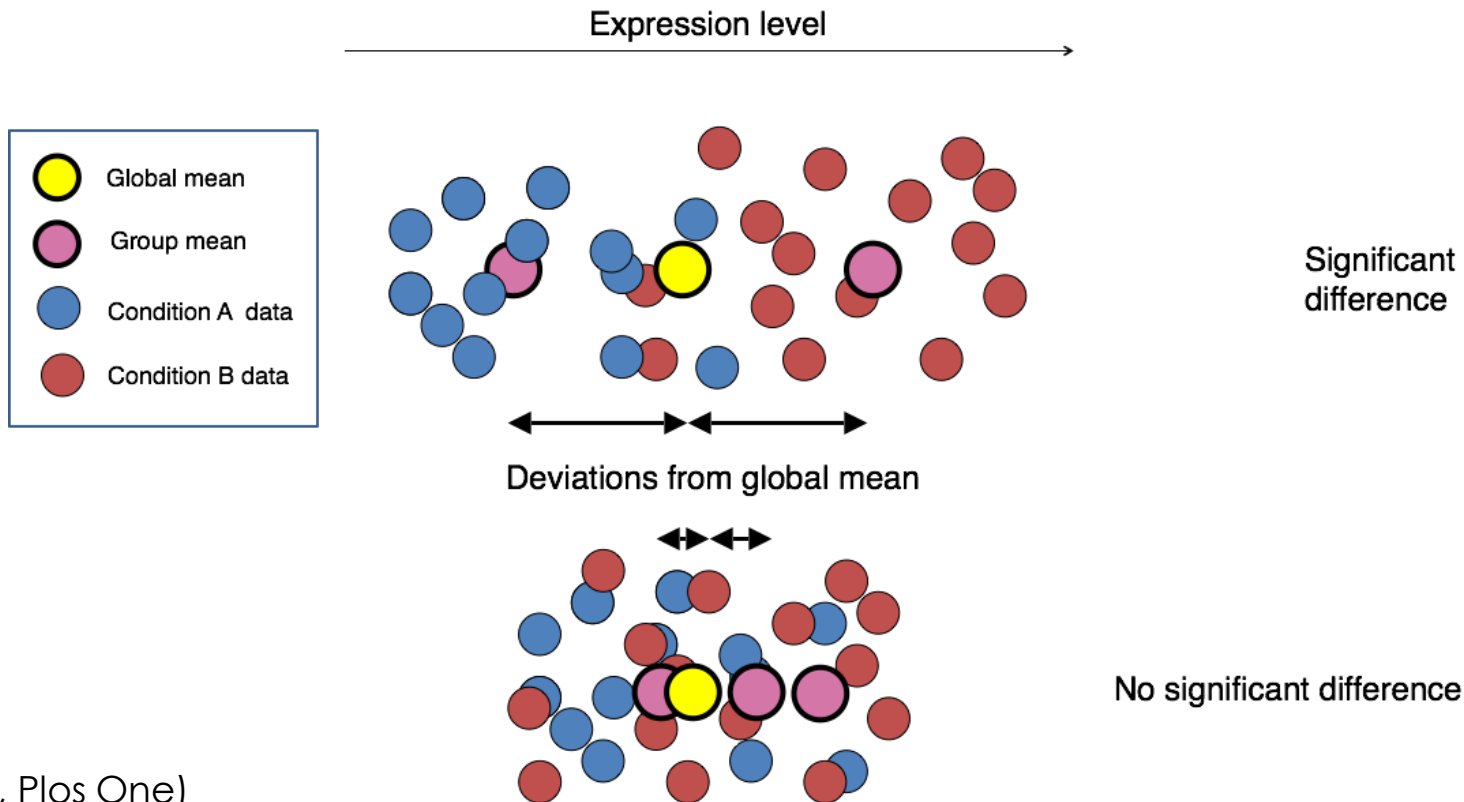


# Differential Expression Estimation

A right estimation of differentially expressed genes between two conditions is key for understanding phenotypical variations

**We should estimate:**

**The magnitude of differential expression**  
**Significance of the differential expression**



# Comparison of DE methods with qRT-PCR

Tool	TPR	SPC	PPV	ACC	$F_1$ measure
edgeR	0.71	0.94	0.90	0.85	0.79
baySeq	0.92	0.40	0.52	0.61	0.66
DESeq	0.44	0.59	0.43	0.53	0.44
NOIseq	<b>0.80</b>	<b>0.95</b>	<b>0.92</b>	<b>0.89</b>	<b>0.86</b>
SAMseq	0.44	0.52	0.39	0.49	0.42
limma+voom	<b>0.81</b>	<b>0.93</b>	<b>0.89</b>	<b>0.88</b>	<b>0.85</b>
EBSeq	0.68	0.55	0.52	0.60	0.59
DESeq2	<b>0.84</b>	<b>0.95</b>	<b>0.92</b>	<b>0.90</b>	<b>0.88</b>
sleuth	0.77	0.54	0.54	0.63	0.64

<https://doi.org/10.1371/journal.pone.0190152.t004>

ACC: Rate of right predictions

SPC: ratio of true detection

TRP: Sensibility or rate of true discovery

**NOIseq, Limma+voom and DESeq2 are the programs highly correlating with qRT-PCR results**

# Visualization of differential expression analysis

The raw out of a DE analysis is a long table of genes/transcripts with stats results and expression information

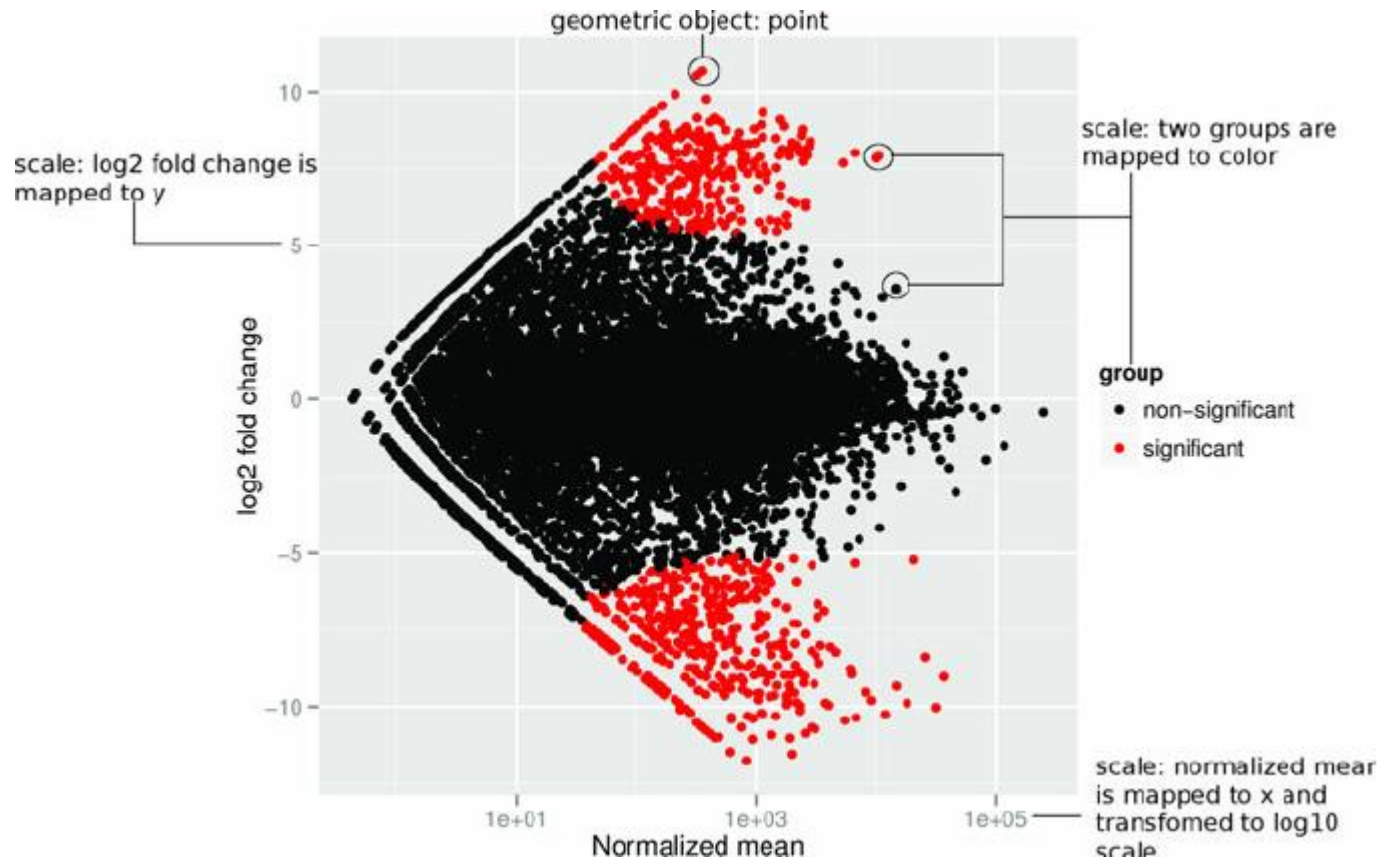
ID	Gene_name	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSMUSG00000024907	Gal	1323	2,0	0	3	0	0,019
ENSMUSG00000050541	Adra1b	174	8,0	2	4	0	0,005
ENSMUSG00000072663	Spef2	127	8,2	2	3	0	0,010
ENSMUSG00000082575	Eef2-ps2	130	8,2	2	3	0	0,009
ENSMUSG00000020325	Fstl3	53	8,8	3	3	0	0,097
ENSMUSG00000064202	Spata6l	53	8,8	3	3	0	0,091
ENSMUSG00000075307	Klh41	106	8,9	3	3	0	0,031
ENSMUSG00000071398	2410004P03Rik	110	8,9	3	3	0	0,038
ENSMUSG00000116735	Gm49555	58	9,0	3	3	0	0,059
ENSMUSG00000026730	Pter	60	9,0	3	3	0	0,093
ENSMUSG00000115569	Gm49169	60	9,0	3	3	0	0,096
ENSMUSG00000018923	Med11	60	9,0	3	3	0	0,086
ENSMUSG00000028840	Zfp593	120	9,1	2	4	0	0,003

We often are interested on:

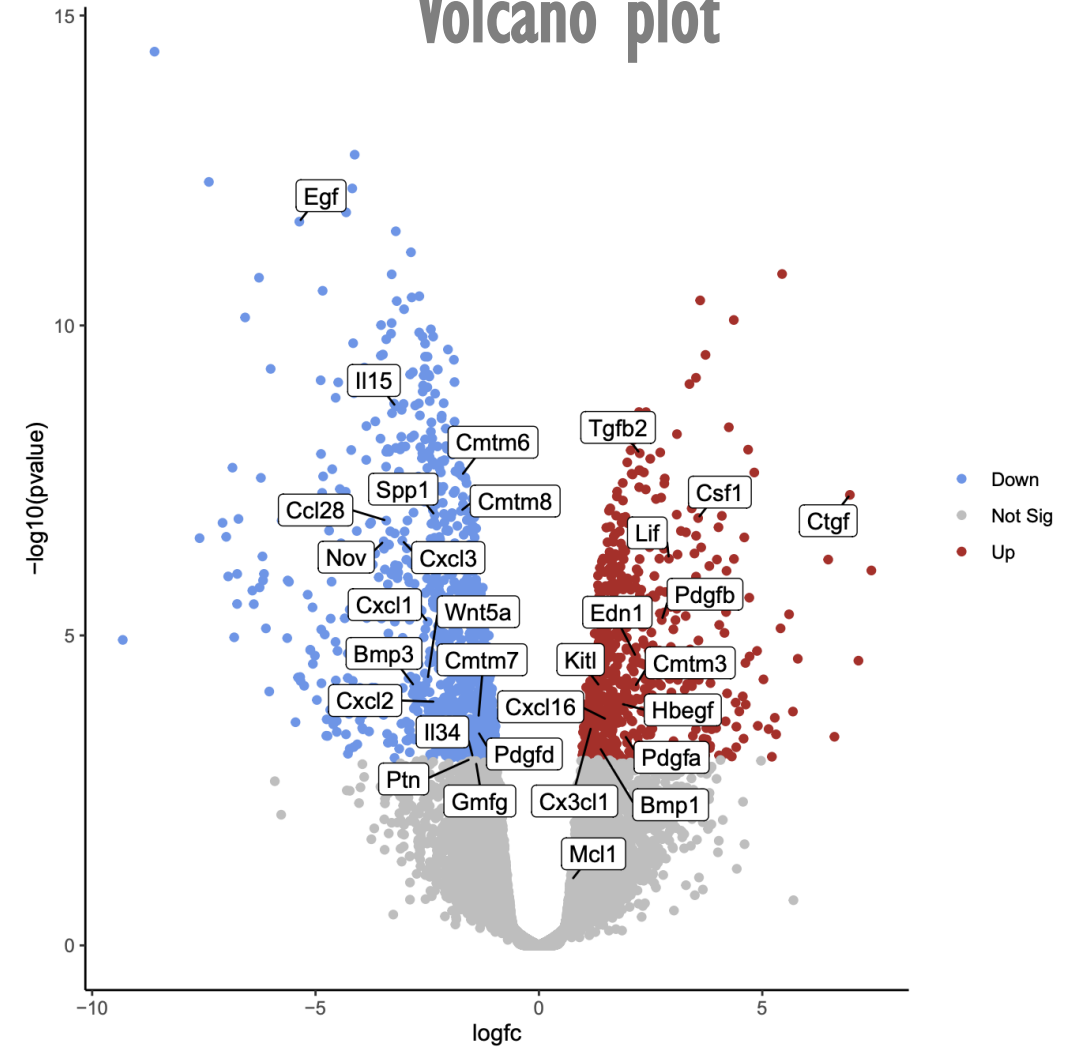
- Up and Downregulated genes (significantly changing)
- Fold of change
- Expression levels of DE genes

# Plotting results (scatterplots)

## MA plot

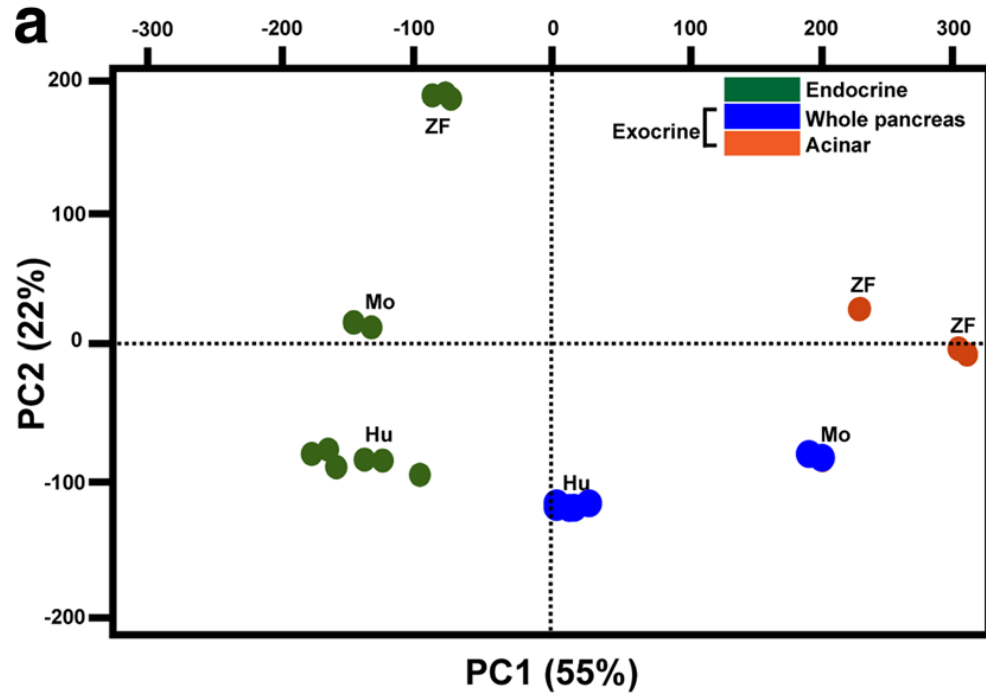


## Volcano plot

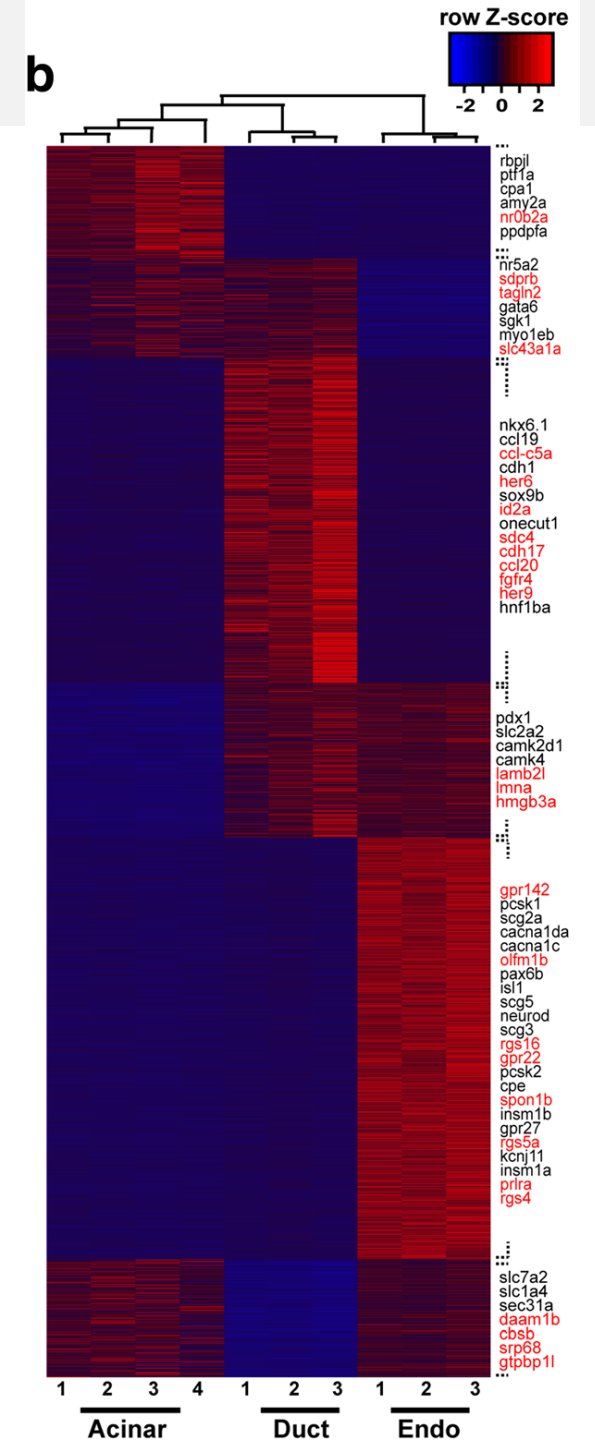


# Plotting results

## Principal Component Analysis (PCA) plot

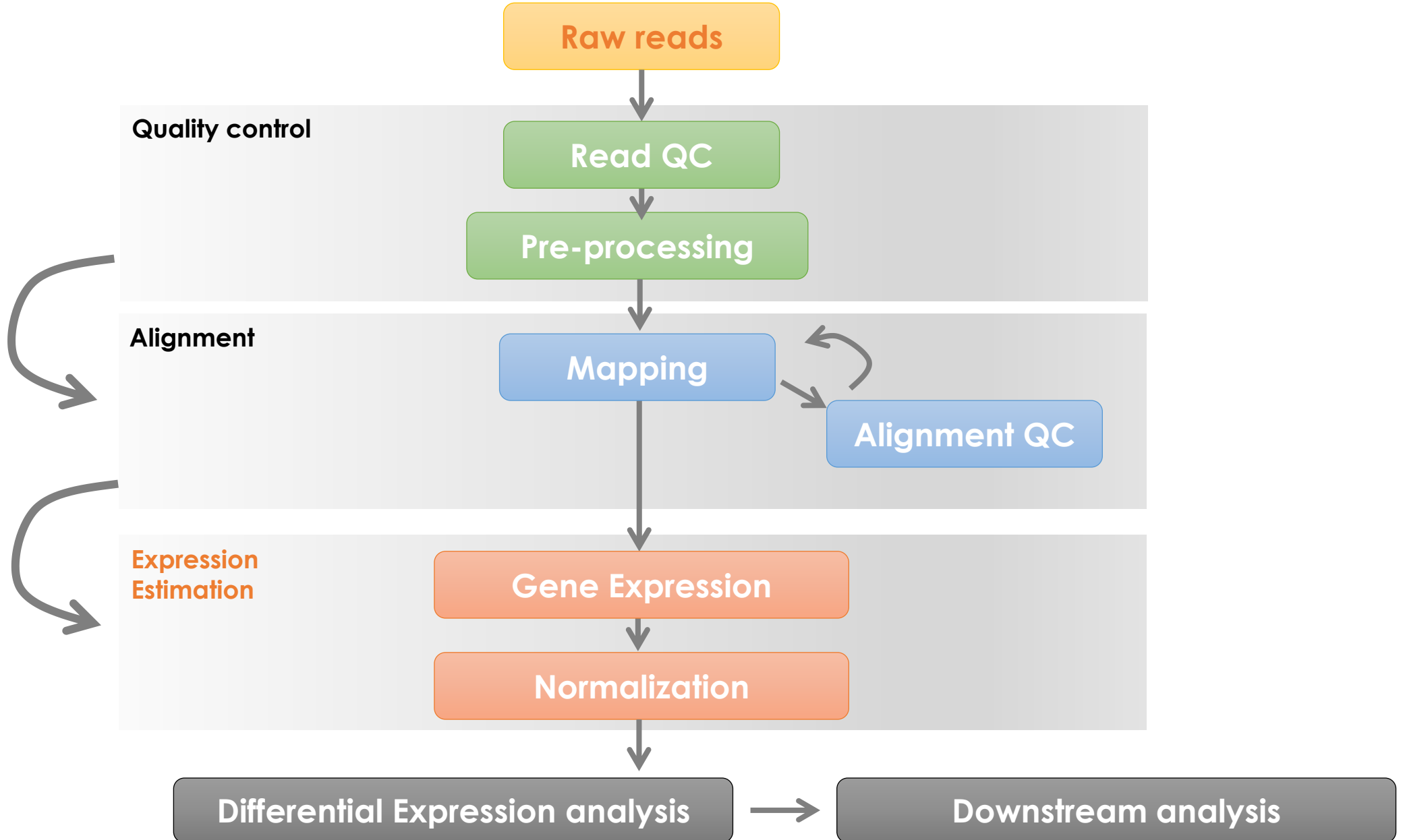


## HeatMap plot



# Data Analysis General Pipeline

3 main steps





# RNA-seq Downstream Analysis

After we got our DE gene list, we need to add biological meaning to this set of genes based on the following questions:

Biological function of modulated genes?

Biological pathway affected by my treatment?

# RNA-seq Downstream Analysis

Then, we can run different analyses to get ideas about the function of the modulated genes:

- Gene Ontology Mapping
- Enrichment Analysis
- Gene Set Enrichment analysis

# Gene Ontology

The Gene Ontology (GO) describes our knowledge of the biological domain with respect to three aspects:

## Molecular Function:

describe activities that occur at the molecular level, such as “catalysis” or “transport”.

## Cellular Component:

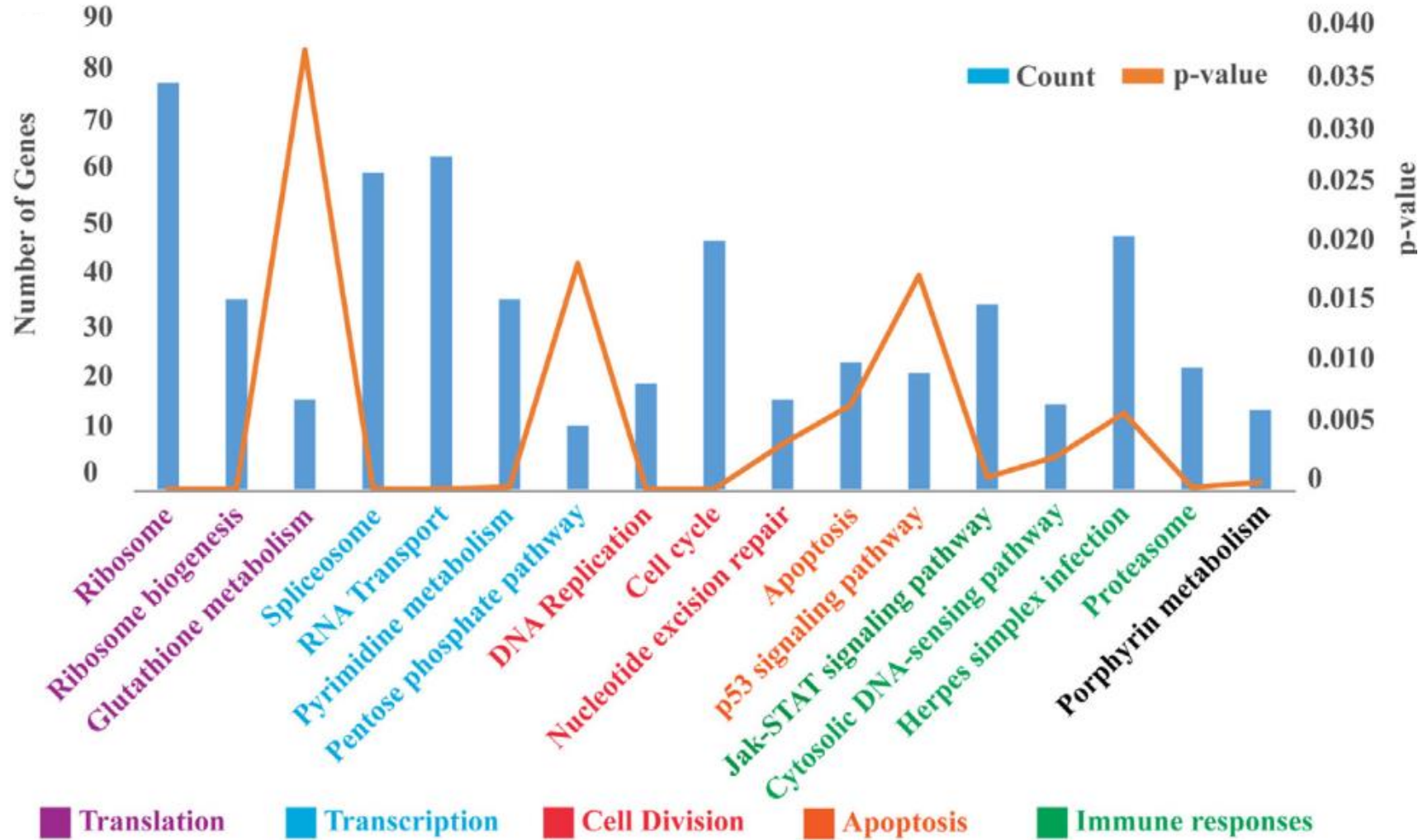
Locations relative to cellular structures in which a gene product performs a function, either cellular compartments (e.g., mitochondrion), or stable macromolecular complexes of which they are parts (e.g., the ribosome).

## Biological Process:

The larger processes, or ‘biological programs’ accomplished by multiple molecular activities. Examples of broad biological process terms are DNA repair or signal transduction.

**They do not represent biological pathways**

# Gene Ontology plot



# The GENEontology Consortium



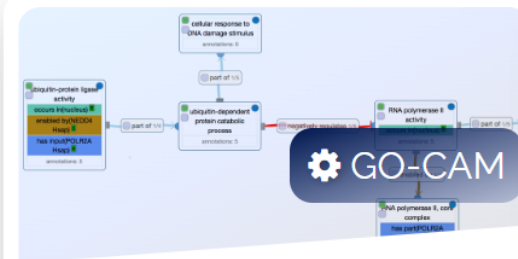
The network of biological classes describing the current best representation of the “universe” of biology: the molecular functions, cellular locations, and processes gene products may carry out.

- [GO Ontology Overview](#)
- [Browse in AmiGO](#)
- [Download](#)



Statements, based on specific, traceable scientific evidence, asserting that a specific gene product is a real exemplar of a particular GO class.

- [GO Annotations Overview](#)
- [Browse in AmiGO](#)
- [Download](#)



GO Causal Activity Model (GO-CAM) provides a structured framework to link standard GO annotations into a more complete model of a biological system.

- [GO-CAM Overview](#)
- [Browse GO-CAMs](#)
- [Download](#)



Tools to curate, browse, search, visualize and download both the ontology and annotations. Includes bioinformatic guides (Notebooks) and simple API access to integrate the GO into your research.

- [GO Tools Overview](#)
- [GO APIs Guide](#)
- [GO GitHub](#)

# Panther: a webtool for GO



[LOGIN](#) [REGISTER](#) [CONTACT US](#)

[Home](#) [About](#) [PANTHER Data](#) [PANTHER Tools](#) [PANTHER Services](#) [Workspace](#) [Downloads](#) [Help/Tutorial](#)

[PANTHER17.0 Released.](#)

## Search

All

## Quick links

[Whole genome function views](#)

[Genome statistics](#)

[Data Version](#)

[PANTHER API](#)

[FAQ](#)

[How to cite PANTHER](#)

[Recent publication describing PANTHER](#)

## News

[PANTHER17.0 Released.](#)

[Click for additional info.](#)

## Newsletter subscription

Enter your Email:



## Gene List Analysis

## Browse

## Sequence Search

## cSNP Scoring

## Keyword Search

Please refer to our article in [Nature Protocols](#) for detailed instructions on how to use this page.

### Help Tips

#### Steps:

- 1. Select list and list type to analyze
- 2. Select Organism
- 3. Select operation

[Using enhancer data](#)

### 1. Enter ids and or select file for batch upload. Else enter ids or select file or list from workspace for comparing to a reference list.

Enter IDs: [Supported IDs](#)

separate IDs by a space or comma

Upload IDs: [File format](#)

No file chosen

Please [login](#) to be able to select lists from your workspace.

Select List Type:

- ID List
- Previously exported text search results
- Workspace list
- PANTHER Generic Mapping
- ID's from Reference Proteome Genome

Organism for id list

- VCF File
- Flanking region
- Search Enhancer Data

### 2. Select organism.

- Homo sapiens
- Mus musculus
- Rattus norvegicus
- Gallus gallus
- Danio rerio

### 3. Select Analysis.

- Functional classification viewed in gene list
- Functional classification viewed in graphic charts
- Bar chart
- Pie chart
- Statistical overrepresentation test
- Statistical enrichment test

# Enrichment Analysis

- It characterizes a gene list by **looking at classes of genes** representing functions that are **overrepresented** on the list and associated with your study
- The analysis test **statistically the overrepresentation** of these gene classes and estimate if they are **significant**
- For this analysis, the **gene background used is essential**. Your background must respond to the classes of genes used as input.
  - For transcriptome-wide modulated gene set the perfect background would be all the genes expressed in your data set.
  - For regulated kinases gene set a “kinome” background (all kinases annotated in the genome)

# DAVID: Webtool for pathways analysis

**DAVID BIOINFORMATICS RESOURCE**

Analysis V

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads

**Upload List Background**

### Upload Gene List

Demolist 1 Demolist 2 Upload Help

**Step 1: Enter Gene List**

A: Paste a list

Or

B: Choose From a File

No file chosen

Multi-List File ?

**Step 2: Select Identifier**

**Step 3: List Type**

Gene List

Background

**Step 4: Submit List**

← **Step 1. Submit your gene list**

An example:

Copy/paste IDs to "box A" -> Select Identifier

- 1007\_s\_at
- 1053\_at
- 117\_at
- 121\_at
- 1255\_g\_at
- 1294\_at
- 1316\_at
- 1320\_at
- 1405\_i\_at
- 1431\_at
- 1438\_at
- 1487\_at
- 1494\_f\_at
- 1598\_g\_at

### Gene List Manager

Select to limit annotations by one or more species [Help](#)

- Use All Species -

Danio rerio(240)

Unknown(2)

List Manager [Help](#)

List\_1

Downregulated

**Select List to:**

[View Unmapped Ids](#)

### Annotation Summary Results

[Help and Tool Manual](#)

**Current Gene List: Downregulated**      **240 DAVID IDs**

**Current Background: Danio rerio**      **Check Defaults**

**Functional Annotations (6 selected)**

<input checked="" type="checkbox"/> <b>COG_ONTOLOGY</b>	5.4%	13	<input type="button" value="Chart"/>	
<input type="checkbox"/> PIR_SEQ_FEATURE	0.4%	1	<input type="button" value="Chart"/>	
<input checked="" type="checkbox"/> <b>UP_KW_BIOLOGICAL_PROCESS</b>	23.8%	57	<input type="button" value="Chart"/>	
<input checked="" type="checkbox"/> <b>UP_KW_CELLULAR_COMPONENT</b>	51.2%	123	<input type="button" value="Chart"/>	
<input checked="" type="checkbox"/> <b>UP_KW_MOLECULAR_FUNCTION</b>	44.6%	107	<input type="button" value="Chart"/>	
<input checked="" type="checkbox"/> <b>UP_KW_PTM</b>	21.7%	52	<input type="button" value="Chart"/>	
<input checked="" type="checkbox"/> <b>UP_SEQ_FEATURE</b>	87.9%	211	<input type="button" value="Chart"/>	

**Gene\_Ontology (3 selected)**

<input type="checkbox"/> BIOGRID_INTERACTION	0.8%	2	<input type="button" value="Chart"/>	
<input type="checkbox"/> INTACT	2.5%	6	<input type="button" value="Chart"/>	
<input type="checkbox"/> MINT	0.8%	2	<input type="button" value="Chart"/>	
<input checked="" type="checkbox"/> <b>UP_KW_LIGAND</b>	25.8%	62	<input type="button" value="Chart"/>	

**General Annotations (0 selected)**

**Interactions (1 selected)**

<input checked="" type="checkbox"/> <b>UP_KW_LIGAND</b>	25.8%	62	<input type="button" value="Chart"/>	
---	-------	----	--------------------------------------	--

**Literature (0 selected)**

**Pathways (0 selected)**

<input type="checkbox"/> EC_NUMBER	15.0%	36	<input type="button" value="Chart"/>	
<input type="checkbox"/> <b>KEGG_PATHWAY</b>	42.5%	102	<input type="button" value="Chart"/>	
<input type="checkbox"/> REACTOME_PATHWAY	30.8%	74	<input type="button" value="Chart"/>	
<input type="checkbox"/> WIKIPATHWAYS	19.2%	46	<input type="button" value="Chart"/>	

**Protein\_Domains (4 selected)**

**Tissue\_Expression (0 selected)**

\*\*\*Red annotation categories denote DAVID defined defaults\*\*\*

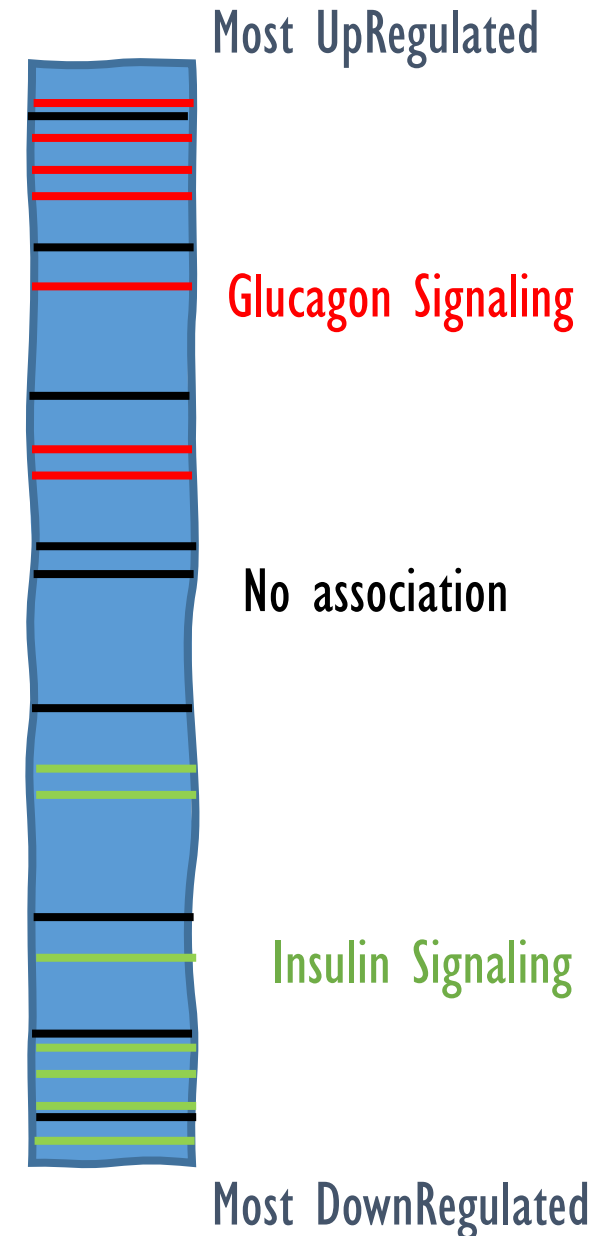
### Combined View for Selected Annotation



# GENE SET ENRICHMENT ANALYSIS (GSEA)

- Instead of comparing modulated genes list with a background list, we use a **ranked list**.
- This list will be organized in descending order based on the fold of change, p-value, etc
- Then, “functional terms” (GO, disease, etc) are mapped to the ranked list.
  - Genes upregulated that are enriched for a certain functional term will be at the top of the list
  - Genes Downregulated enriched for a certain term will be found at the bottom of the list
  - Terms not enriched will be mapped all over the list

**As results we will get enrichment plots by pathways**

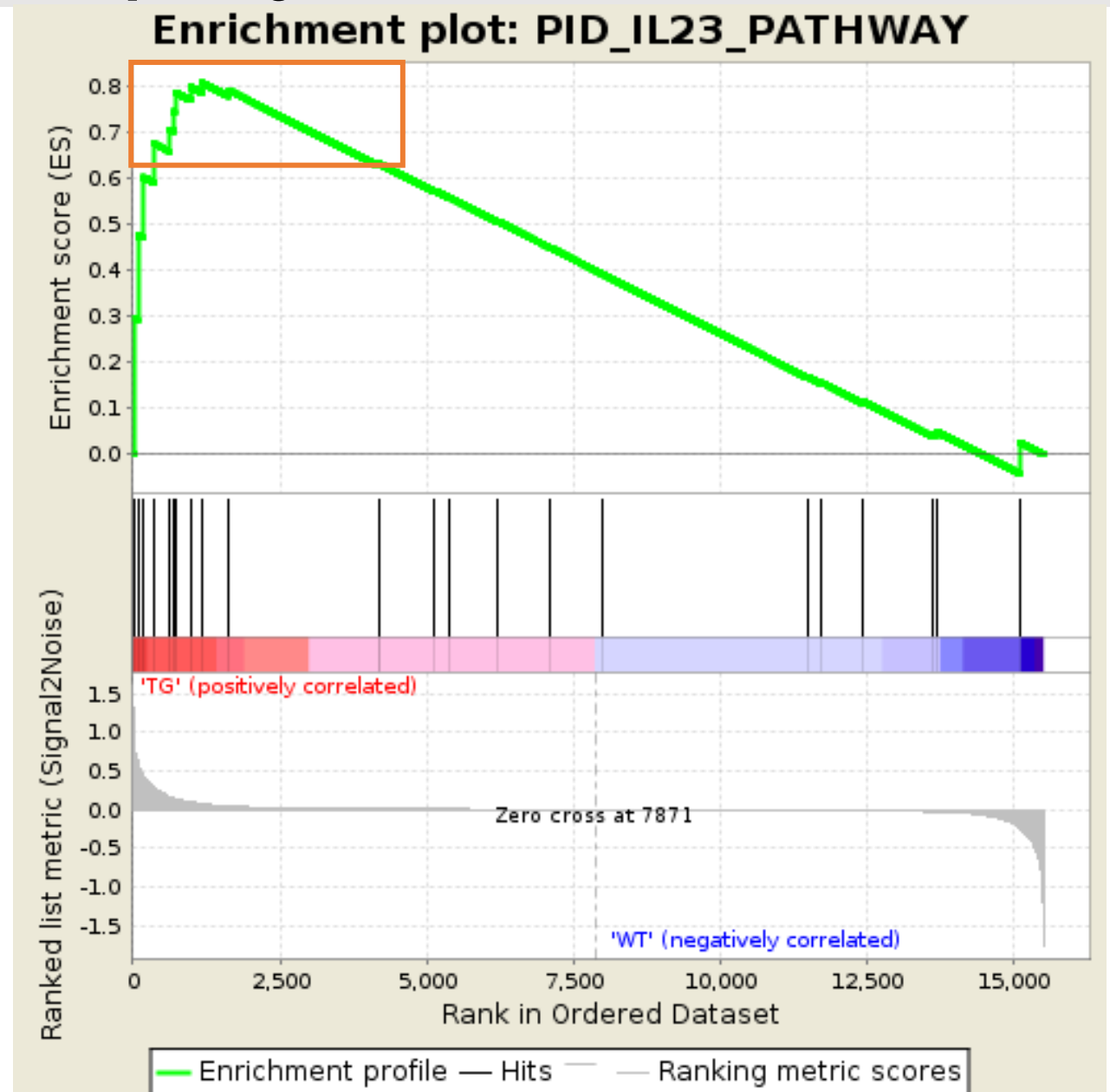


# Enrichment Plots: Interpreting Results from GSEA

## Enrichment Score:

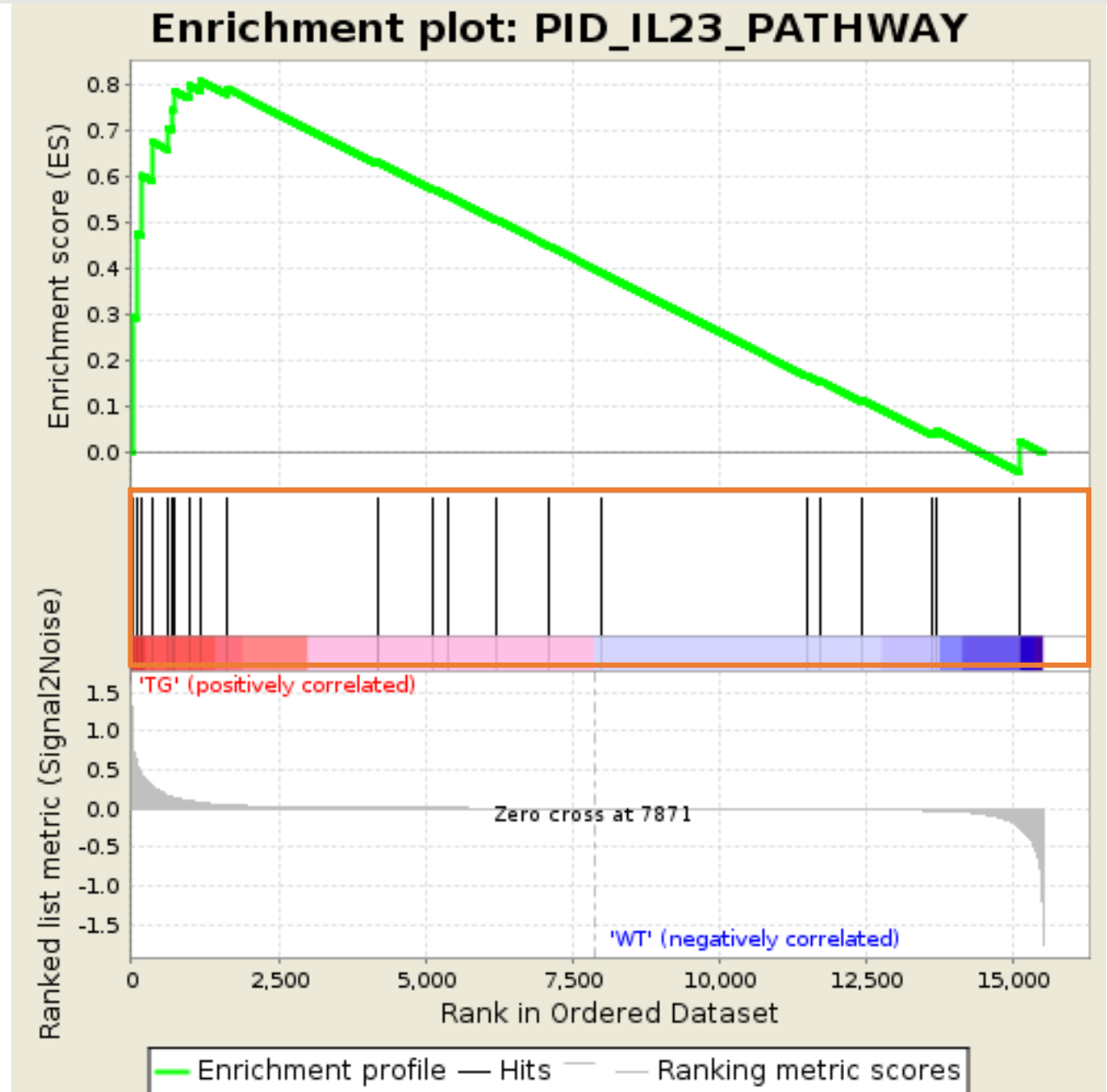
Which reflects the **degree** to which a gene set is **overrepresented** at the top or bottom of a ranked list of genes.

The score at the peak of the plot (the score furthest from 0.0) is the **ES** for the gene set. Gene sets with a **distinct peak** at the beginning (such as the one shown here) or end of the ranked list are generally the most interesting.



# Interpreting Results from GSEA

Shows where the members of the gene list appears in the ranked list of genes



# Interpreting Results from GSEA

## Leading Edge Subset:

is the subset of members that contribute most to the ES. For a positive ES (such as the one shown here), the leading-edge subset is the set of members that appear in the ranked list prior to the peak score.

