DNA Sequencing and Data Analysis

Prof Noam Shomron

Hadas Volkov

Lecture 9, December 30, 2022

DNA Sequencing and Data Analysis

Friday 8:45 AM to 11:15 AM

Arazi-Ofer Building, C.L03

nshomron@gmail.com

hadas.volkov@post.runi.ac.il

DNA Sequencing and Data Analysis

RNA-Seq

Class	Title	Content/assignments	Activity, location
1, 4.11	Introduction to Cells and DNA	Basic knowledge of biology	In the lecture hall, Noam
2, 11.11	DNA Sequencing past and present	Basic knowledge of molecular DNA	In the lecture hall, Noam
3, 18.11	Genomics technologies	DNA, RNA, technologies	In the lecture hall, Noam
4, 25.11	Introduction to Bioinformatics challenges in reading DNA	Focus on three methods: WES/WGS, RNA-seq, cell-free DNA	In the lecture hall, Noam
5, 2.12	Modern DNA Sequencing, 2nd wave File Formats, tools.	Analysis approaches for WES/WGS, RNA-seq, cell-free DNA	In the lecture hall, Hadas and Noam
6, 9.12	De novo Shotgun Assembly	The algorithms and methods behind the assembly problem	In computer class, Hadas and Noam
7, 16.12	Sequence Mapping and Alignment	The algorithms behind mapping and alignment, fast and heuristics	In computer class, Hadas and Noam
8, 23.12	Variant Calling and Somatic Variant Analysis	The bioinformatics behind discovery of novel mutations in cancer	In computer class, Hadas and Noam
9, 30.12	**RNA-Seq**	**The bioinformatics behind RNA-Seq and Differential Gene Expression**	**In computer class, Hadas and Noam**
10, 6.1	Nanopore data analysis introduction Practice molecular biology techniques	Pipetting, transferring small amounts of fluids, running a dry Nanopore experiment	In biology class, Meitar and Noam
11, 13.1	Nanopore DNA sequencing	Nanopore DNA sequencing, experimental run	In biology class, Meitar, Hadas, Assaf
12, 20.1	Nanopore data analysis	Nanopore DNA analysis, experimental run	In computer class, Hadas and Noam
13, 27.1	Nanopore data analysis and presentations	Groups present their results	In the lecture hall, Hadas and Noam

Lesson Goals

Understand some common RNA-seq uses and protocols

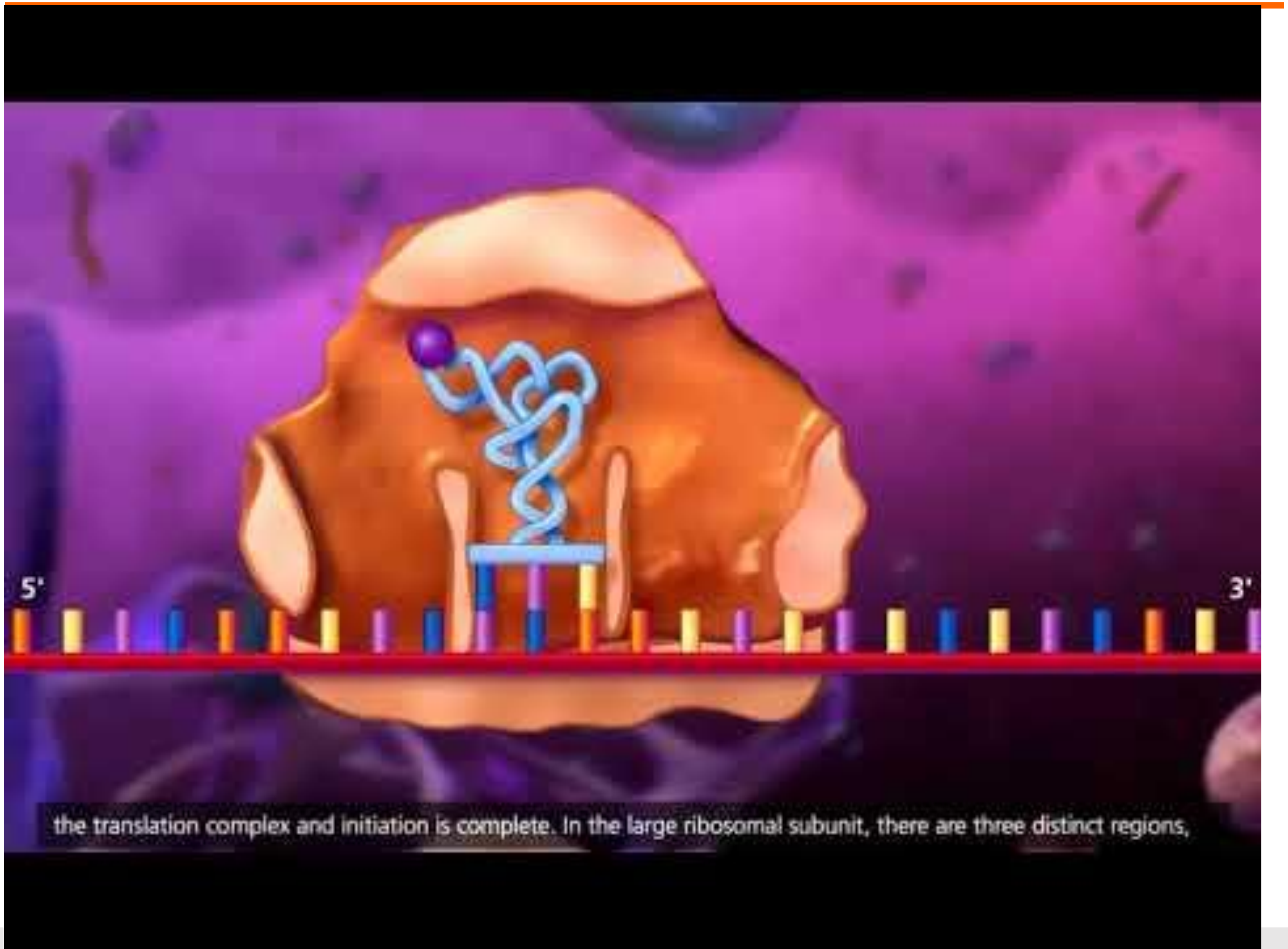
Be familiar with the basic workflow of gene expression data analysis

- Specifically differential gene expression analysis

Know how to map RNA-seq reads to a reference genome using STAR

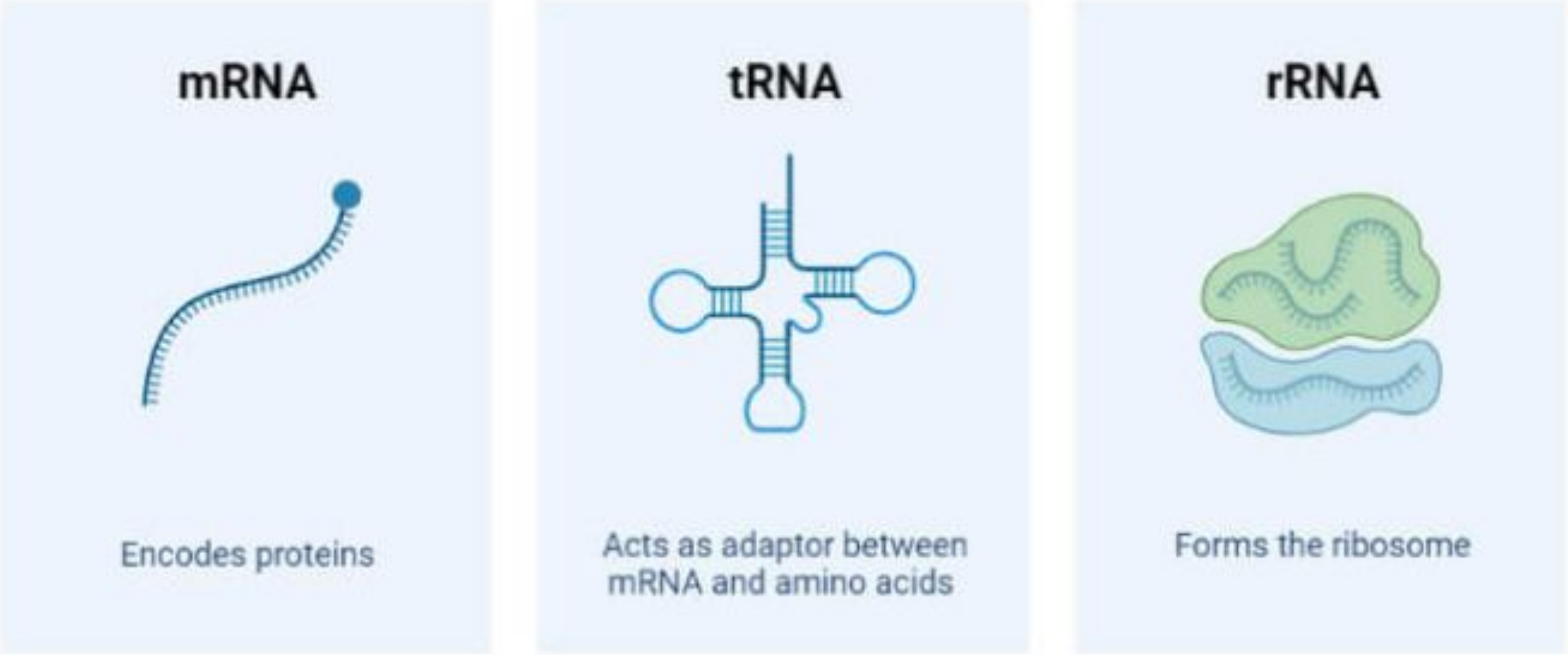
Perform Differential Gene Expression (DGE) analysis

RNA



the translation complex and initiation is complete. In the large ribosomal subunit, there are three distinct regions,

RNA



mRNA

Encodes proteins

tRNA

Acts as adaptor between mRNA and amino acids

rRNA

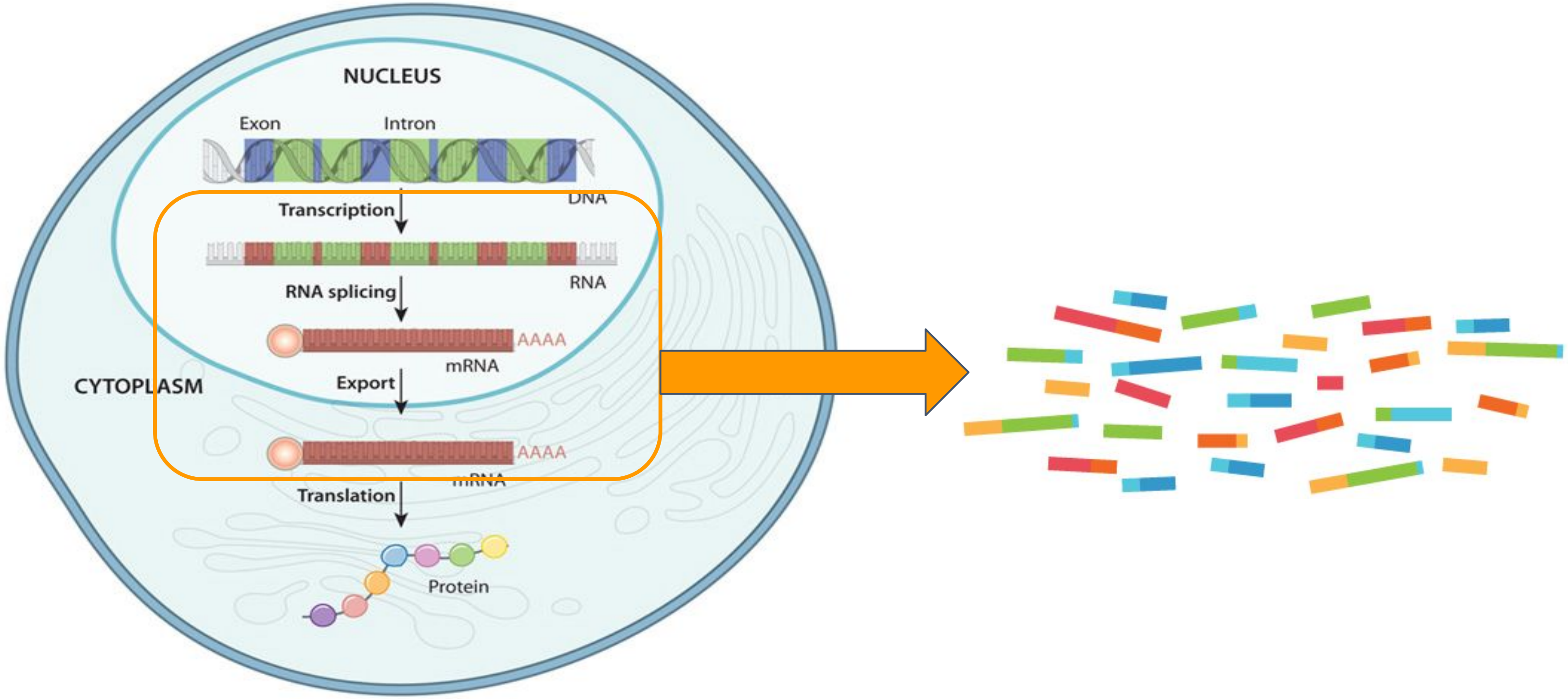
Forms the ribosome

What is RNA-Sequencing?

Sequencing of RNA using NGS technology

Allows assessment of presence and quantity or RNA in a sample

Take a snapshot of expression in a sample



Why Study the Transcriptome?

The full range of messenger RNA molecules expressed by an organism

Indication of cell physiology

Dynamic - responds to the environment

- Changes over time
- Responds to external stimuli
- Controls cellular processes

Reduced representation of the genome

- Smaller = cheaper
- Only the “functional” parts of the genome

Types of Expression Analysis

Expression quantification

Differential gene expression

Assemble whole transcriptomes

Detect new transcripts

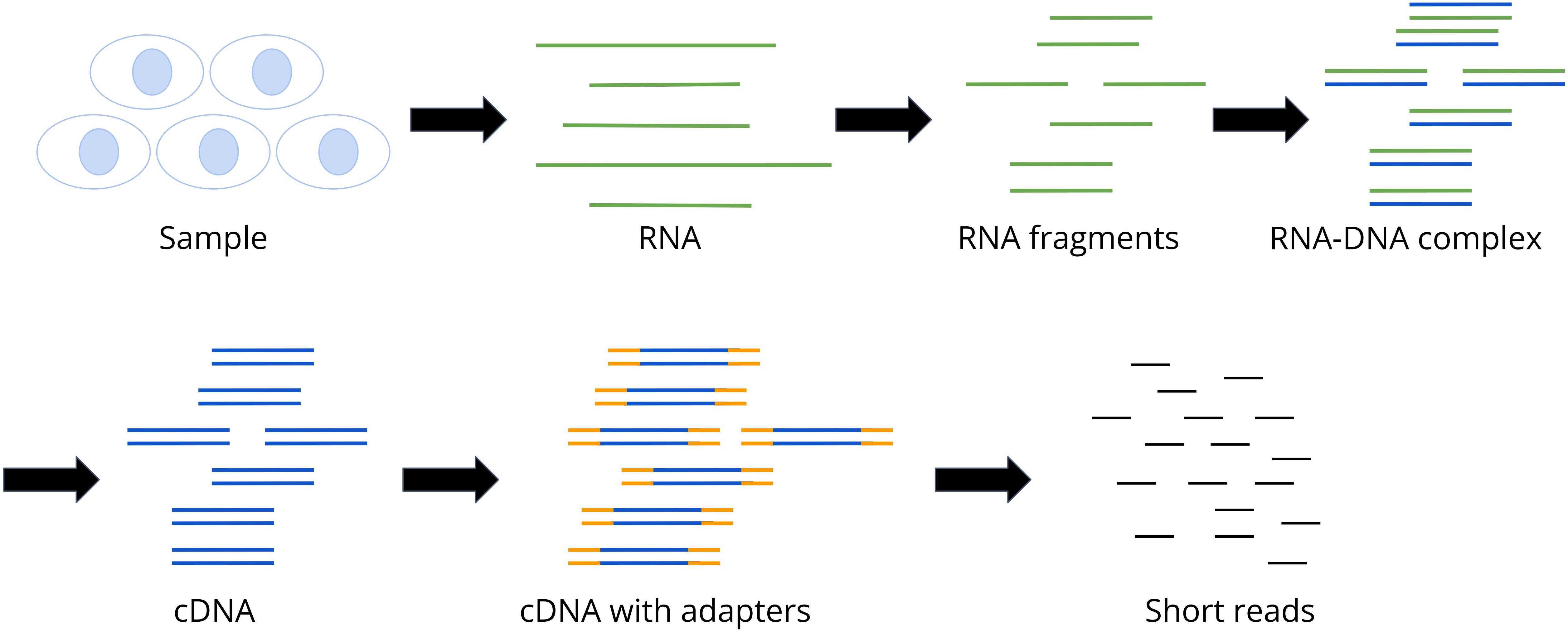
Detect splicing variants

Detect allele-specific expression

Gene co-expression

Single-cell analysis

RNA-Seq Basic Protocol



Sample

RNA

RNA fragments

RNA-DNA complex

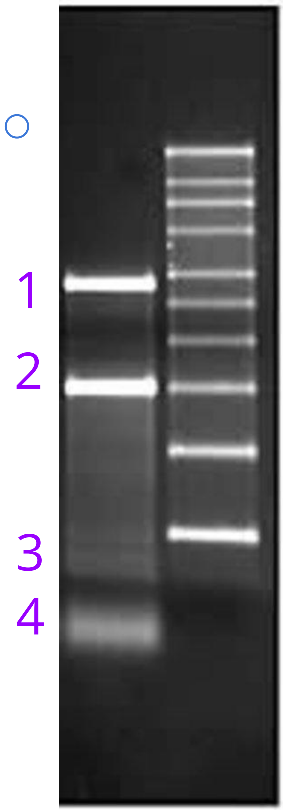
cDNA

cDNA with adapters

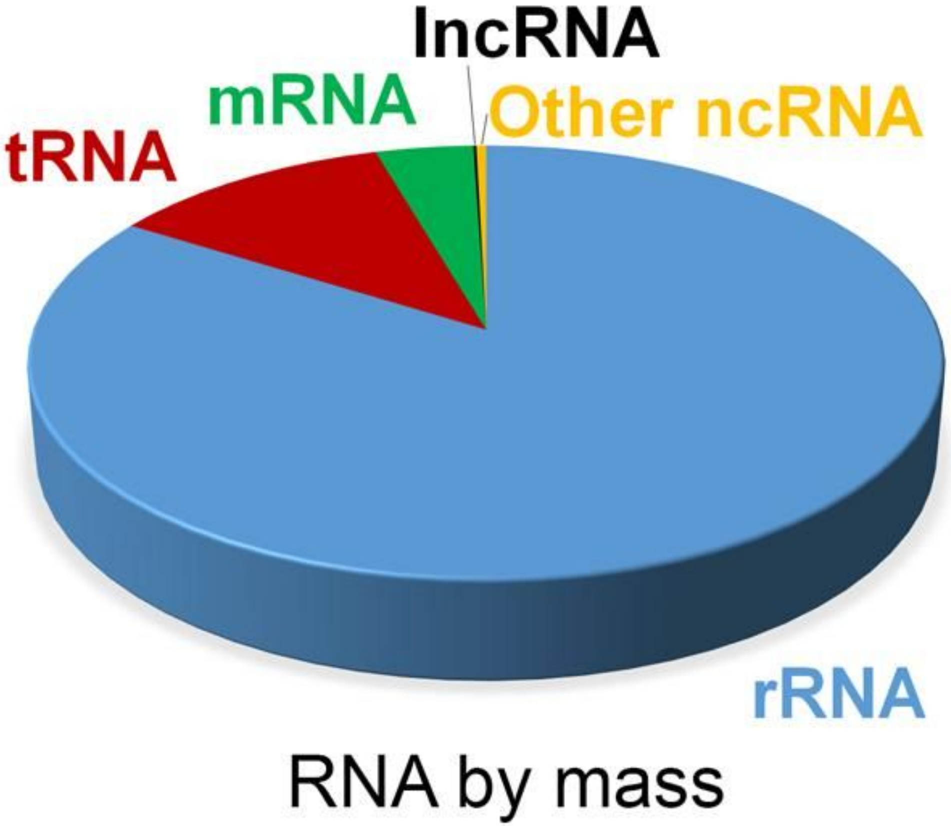
Short reads

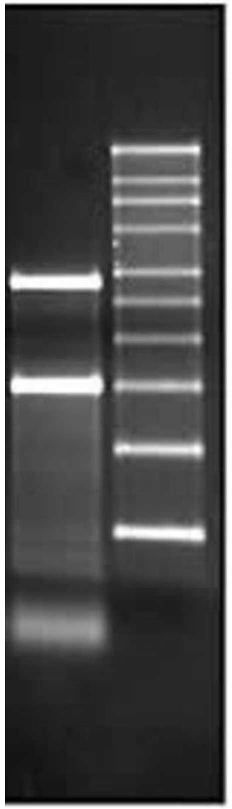
Total RNA

Which band contains the mRNA?



Total RNA (Mammalian)





rRNA large subunit

rRNA small subunit

mRNA

tRNA

Palazzo, Alexander F., and Eliza S. Lee. "Non-coding RNA: what is functional and what is junk?." *Frontiers in genetics* 6 (2015): 2.

Enrichment for Mature RNA

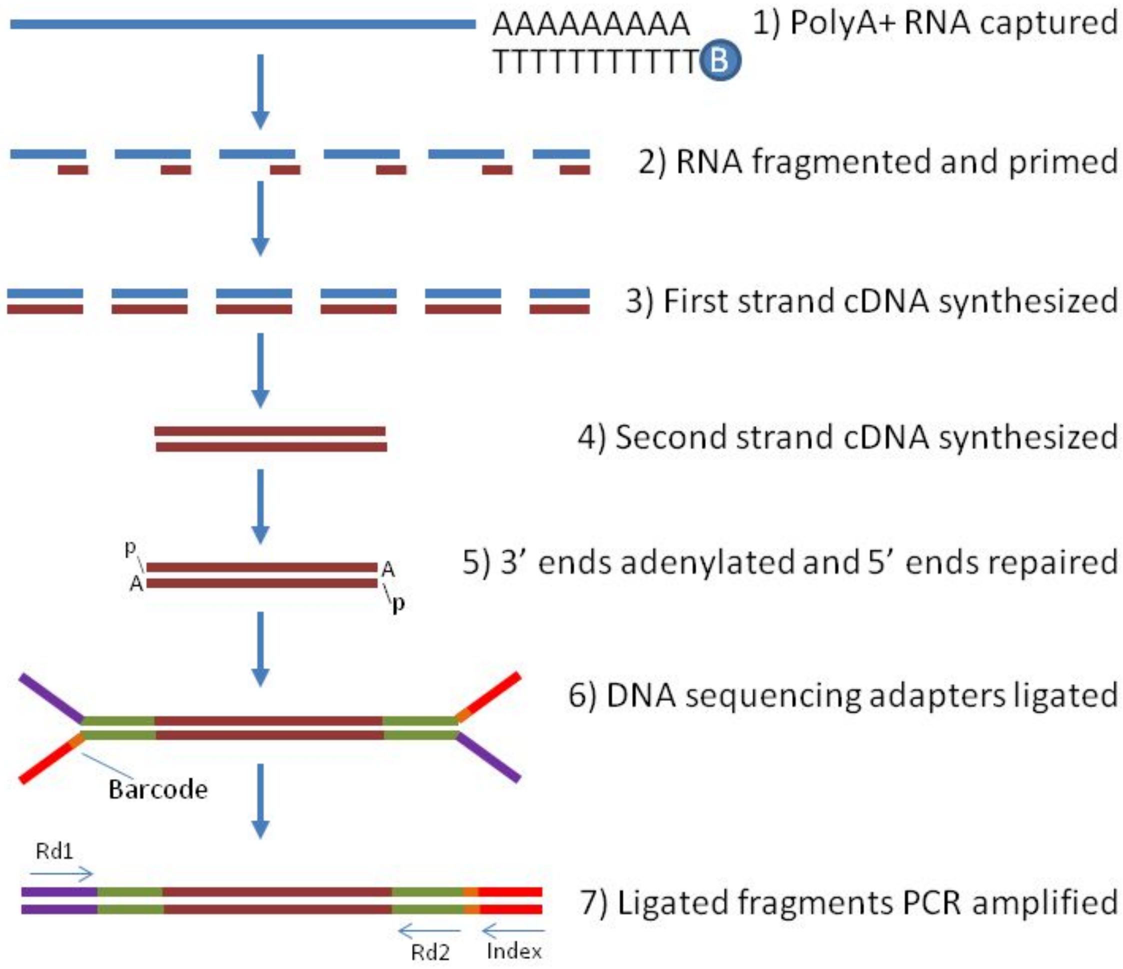
Poly-A selection method

- use a poly-dT baits to bind mRNAs and discard the rest
- Removes rRNA, tRNA and others
- Enriches for mature mRNA (containing poly-A tail)
- Not all mRNAs have poly-A tails
  - Histones mRNA in Metazoans
  - Mitochondrial mRNA

rRNA depletion method

- Use baits designed specifically for rRNA
- Does not remove tRNA
- Only option in bacteria (no poly-A tail)
- Only option when extracting small RNAs

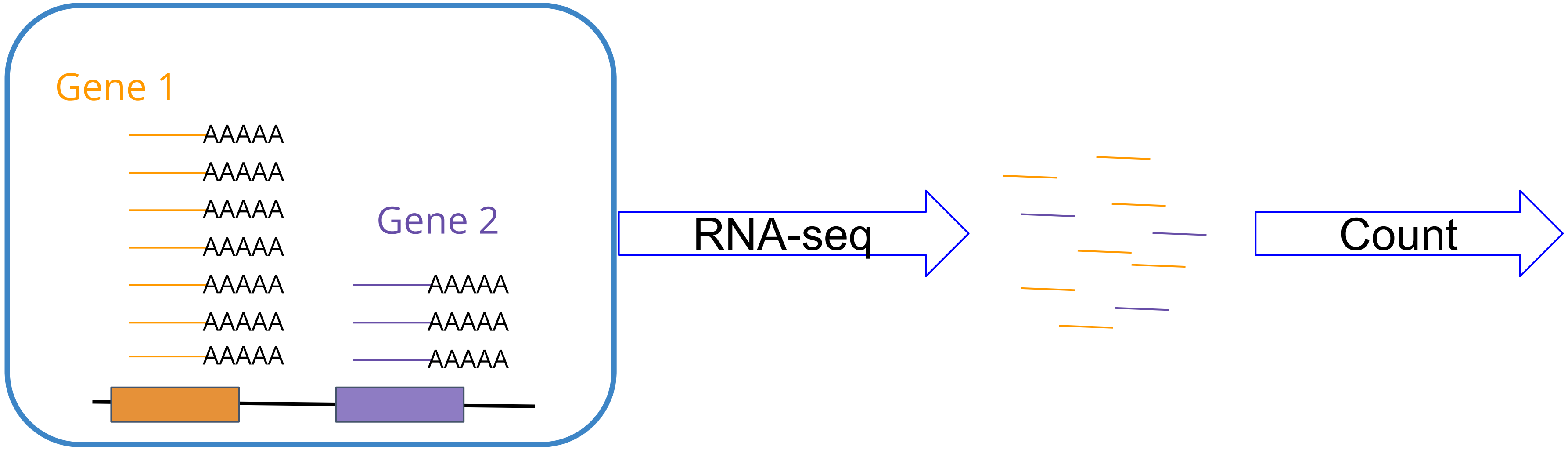
RNA-Seq Library Prep


AAAAAAAAA
TTTTTTTTTTT
B
1) PolyA+ RNA captured
2) RNA fragmented and primed
3) First strand cDNA synthesized
4) Second strand cDNA synthesized
p
A
A
p
5) 3' ends adenylated and 5' ends repaired
6) DNA sequencing adapters ligated
Barcode
Rd1
7) Ligated fragments PCR amplified
Rd2
Index

https://www.labome.com/method/RNA-seq.html

Expression Levels

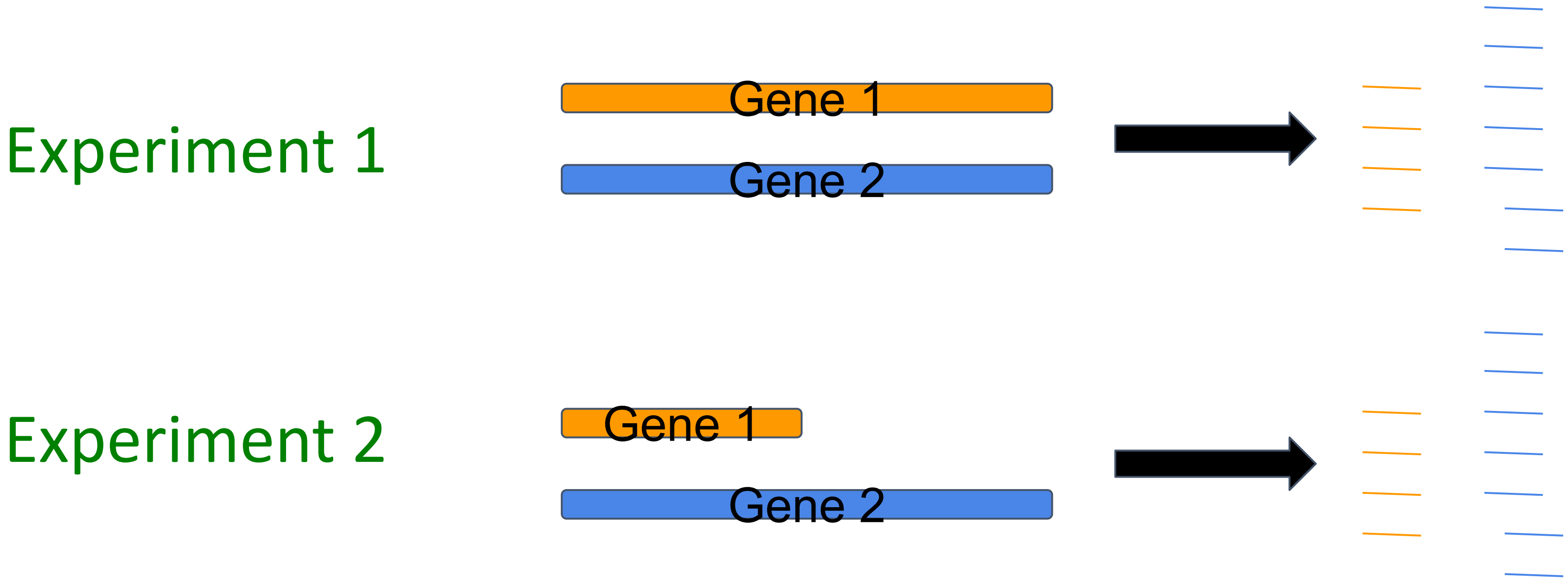
Higher expression → more transcripts → more RNA-seq reads



Gene	Reads
Gene 1	7
Gene 2	3

Are Read Counts a Good Measure?

Variable length of transcript of interest



Experiment 1

Gene 1

Gene 2

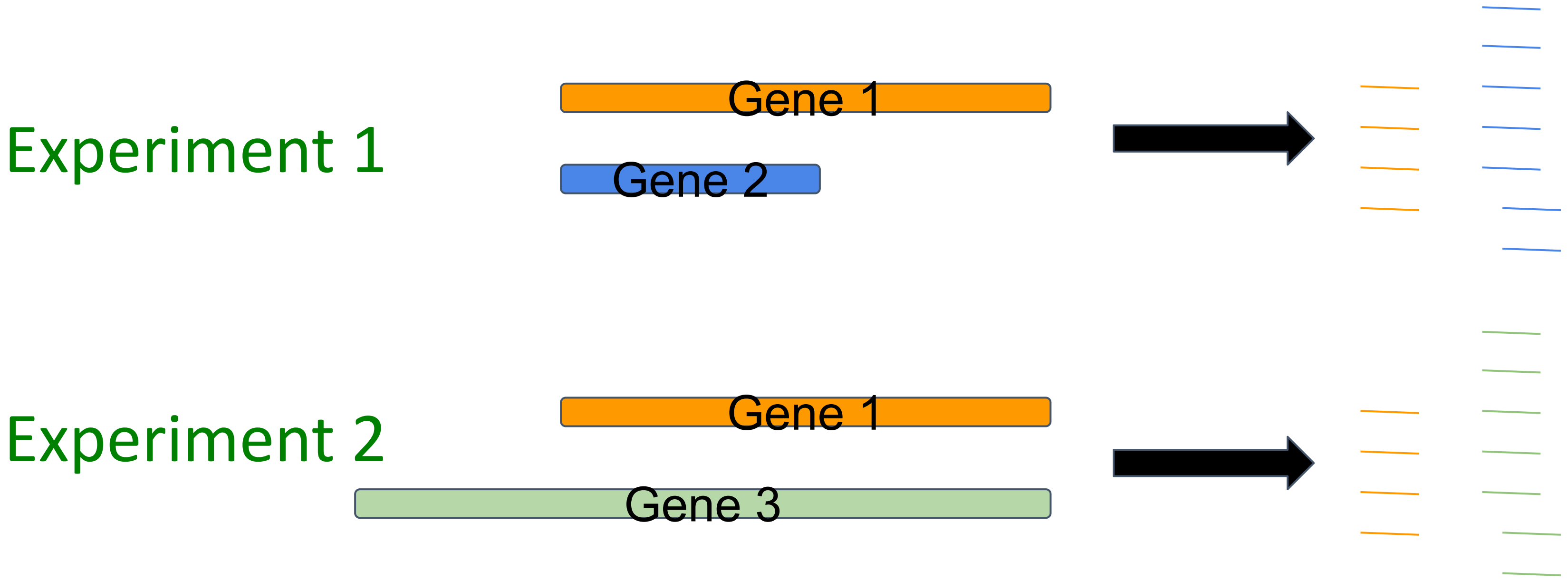
Experiment 2

Gene 1

Gene 2

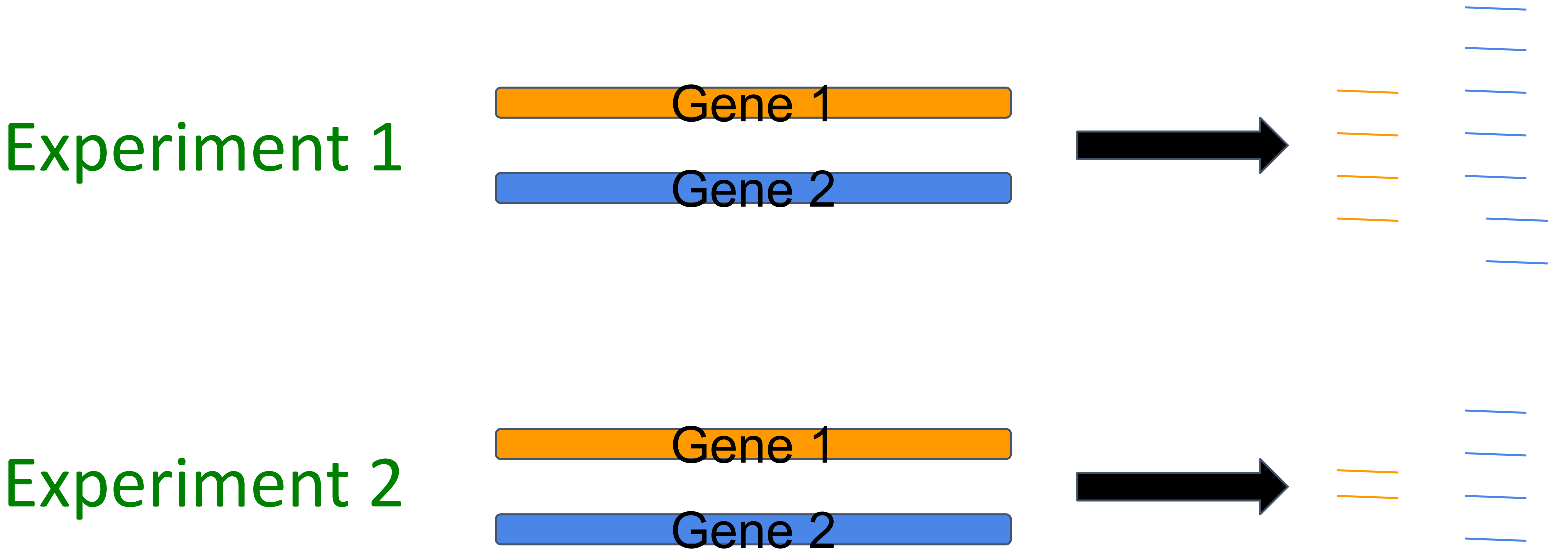
Are Read Counts a Good Measure?

Variable length of other transcripts



Are Read Counts a Good Measure?

Variable number of reads between experiments



Read Count Normalization

Normalize for transcript length

RPK - **reads per kilobase** (of transcript)

$$RPK_i = 10^3 \cdot \frac{n_i}{l_i}$$

Normalize for sequencing depth

RPKM - **reads per kilobase** (of transcript) **per million** (reads)

$$RPKM_i = 10^9 \cdot \frac{n_i}{l_i \cdot \sum_j n_j}$$

Read Count Normalization

Experiment 1 - total reads: 100,000

Gene	**Transcript length**	**Reads**	**RPK**	**RPKM**
Gene1	500	10	20	200
Gene2	1000	20	20	200

Experiment 2 - total reads: 1,000,000

Gene	**Transcript length**	**Reads**	**RPK**	**RPKM**
Gene1	500	10	20	20
Gene2	1000	50	50	50

Differential Gene Expression (DGE)

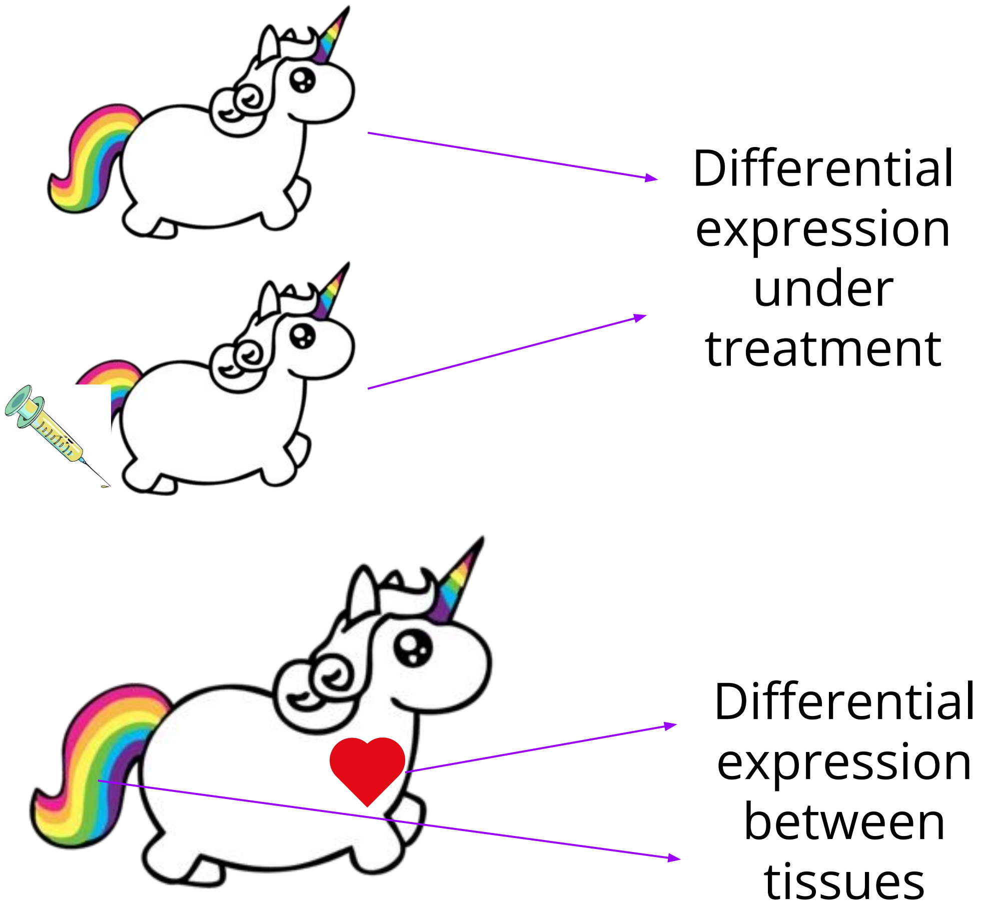
Compare gene expression levels across all genes between two (or more) samples

Samples of the same cell type

- ○ Different physiological conditions
- ○ Different environmental conditions
  - ■ Growth medium
  - ■ Treatment

Samples of different cell types

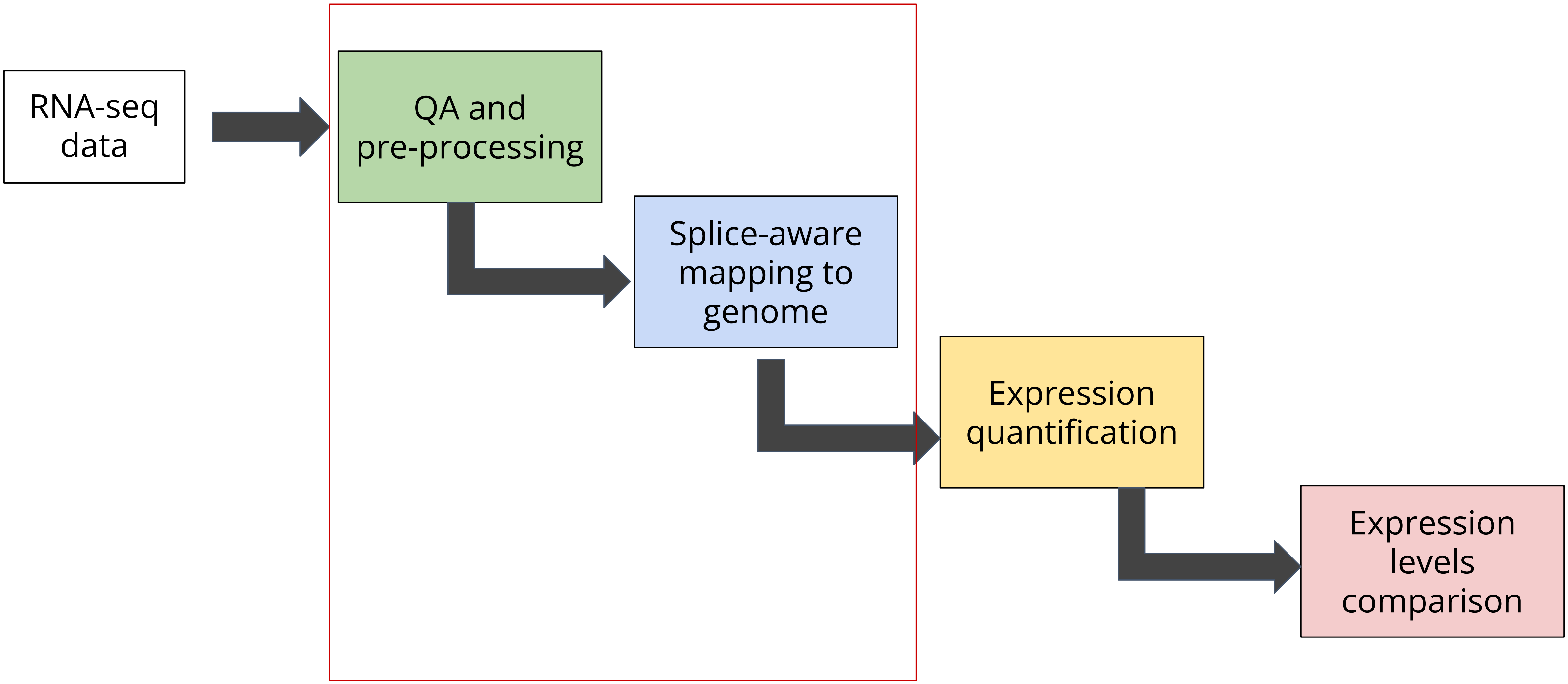
- ○ Different tissues
- ○ WT vs. mutant
- ○ Different strains
- ○ Normal vs. tumor



Differential expression under treatment

Differential expression between tissues

DGE Workflow



RNA-seq data

QA and pre-processing

Splice-aware mapping to genome

Expression quantification

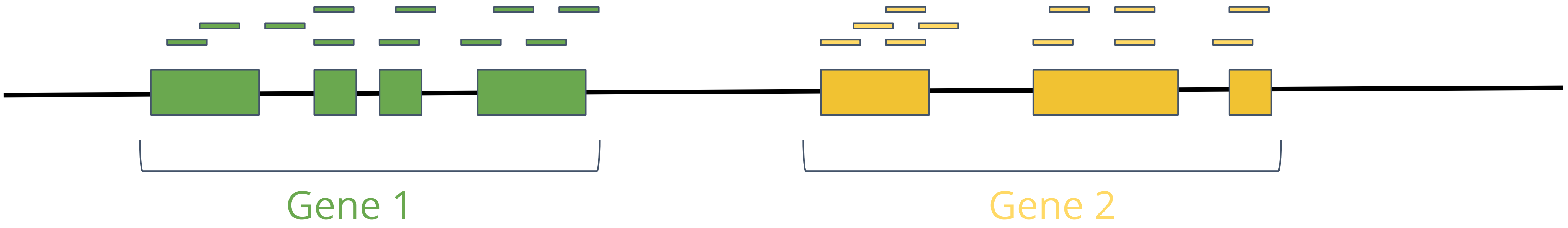
Expression levels comparison

Splicing

We need to assign RNA-seq reads to specific genes

The simplest way is by read mapping

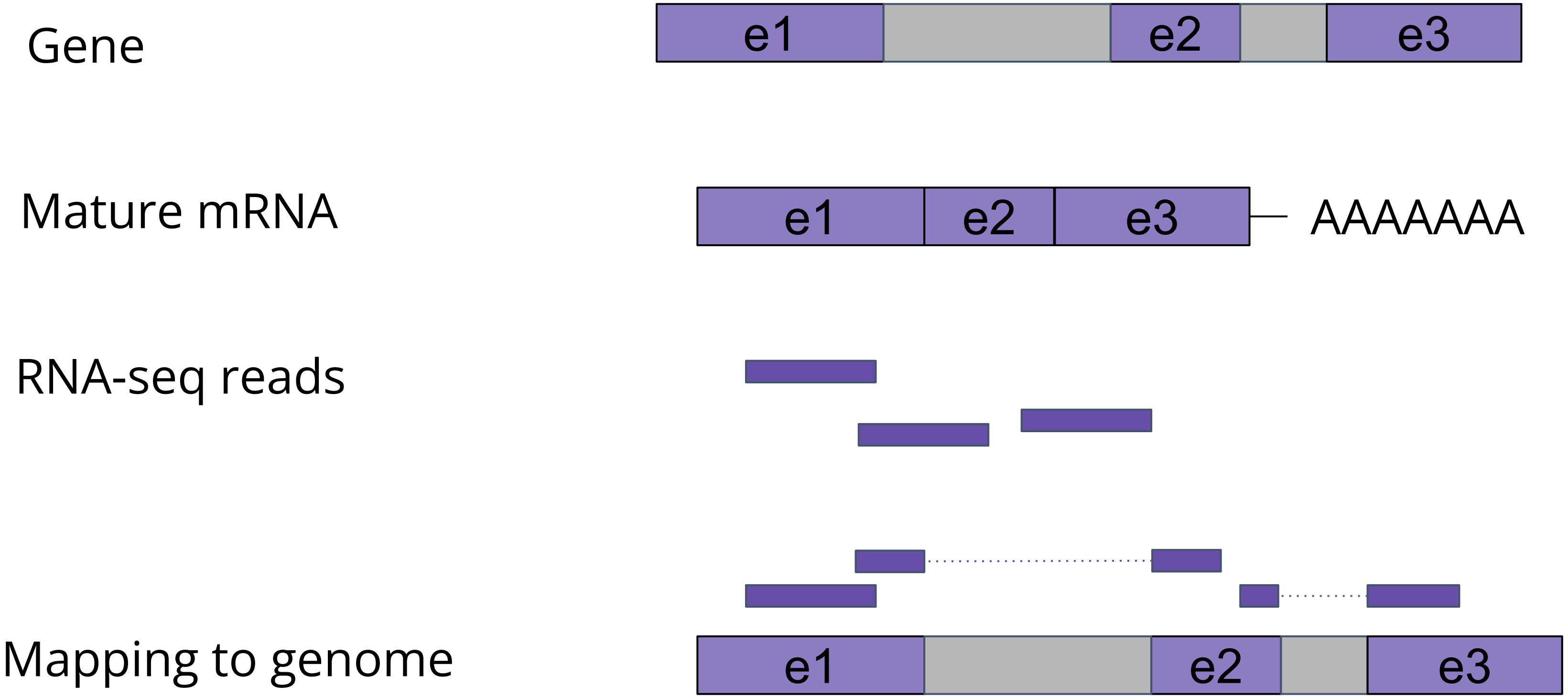
We must have a reference genome and annotation



Splice Aware Mapping

We need to consider intron-exon gene structure

Allow large gaps in read alignment



Gene

Mature mRNA

RNA-seq reads

Mapping to genome

STAR

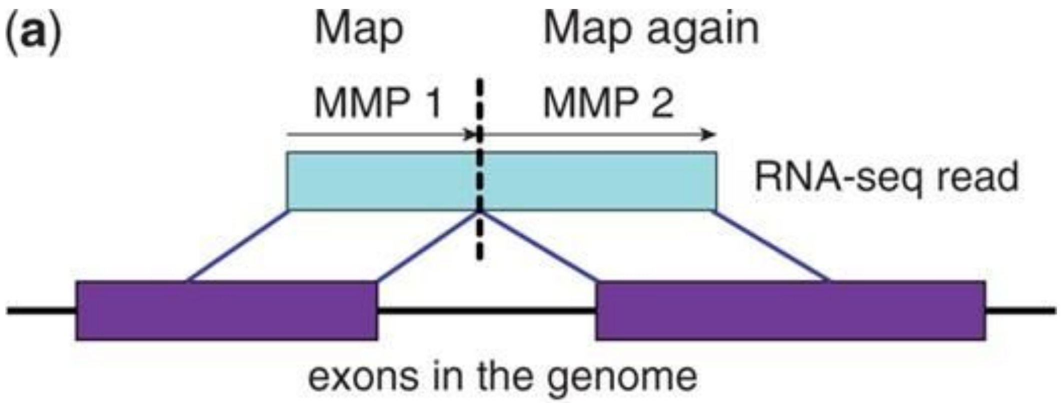
BWA won’t work so well

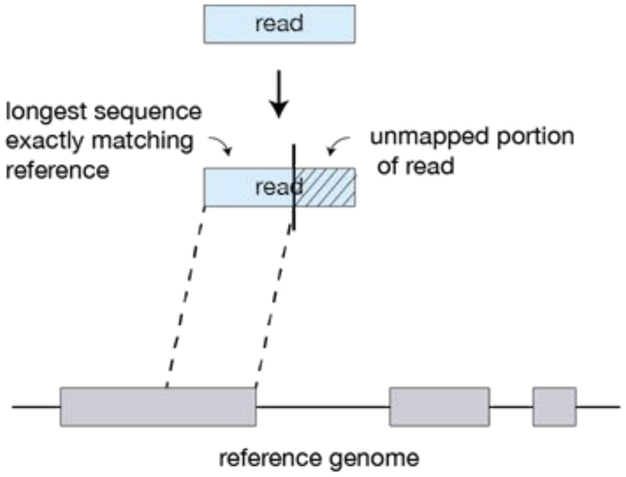
STAR - a very popular choice

Very fast and memory-efficient

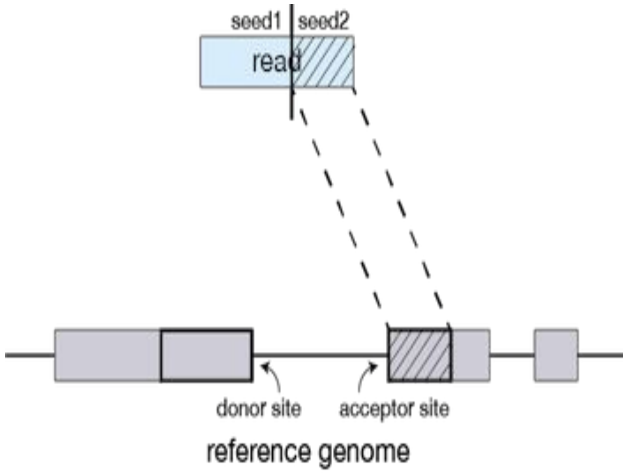
The algorithm works in two steps:

1. Find splice junctions by allowing partial read mapping
2. Stitch together parts of reads mapped to proximal genomic positions

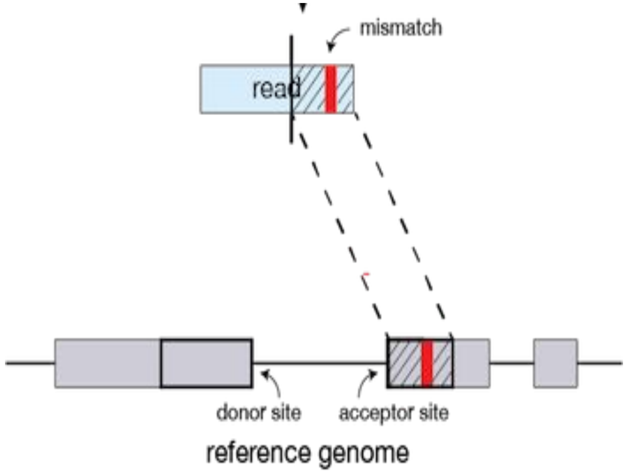




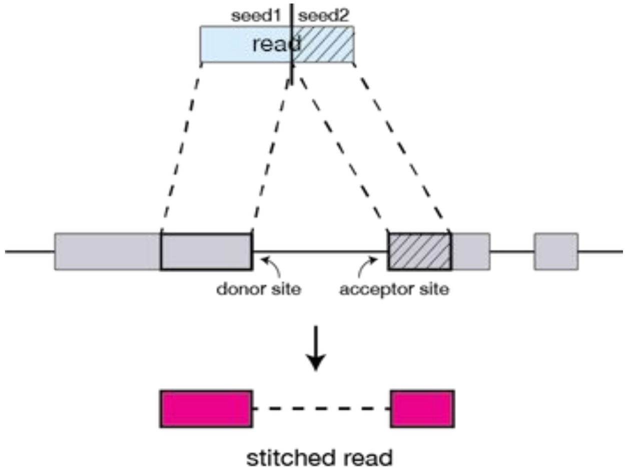
Find best exact mapping



Find best exact mapping for unmapped part



Try to extend mapping over mismatches



Cluster and stitch mappings based on proximity

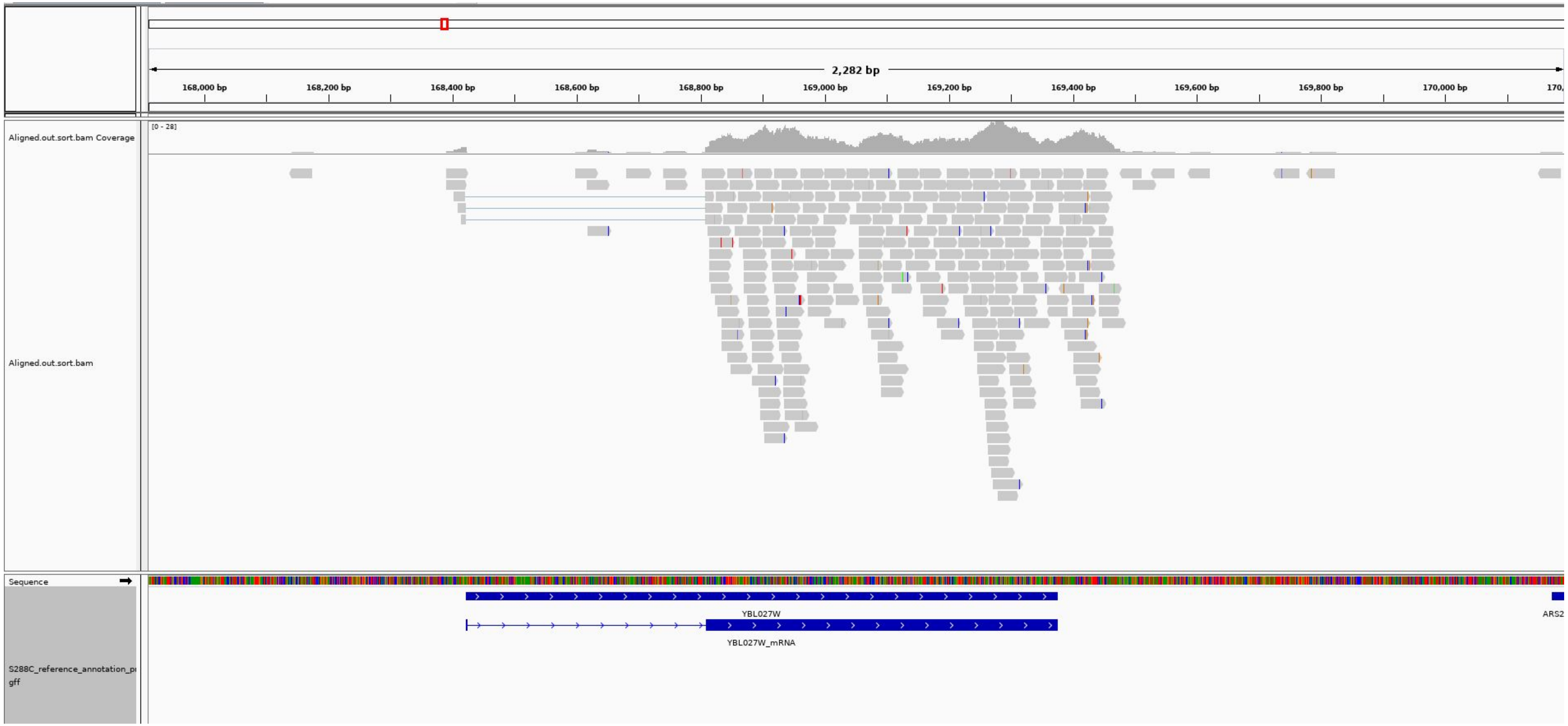
STAR

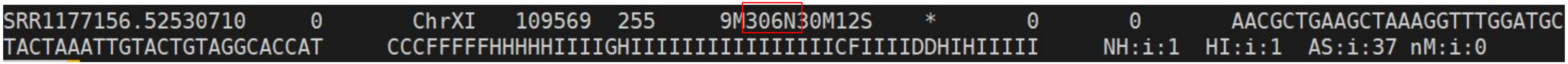
Inputs:

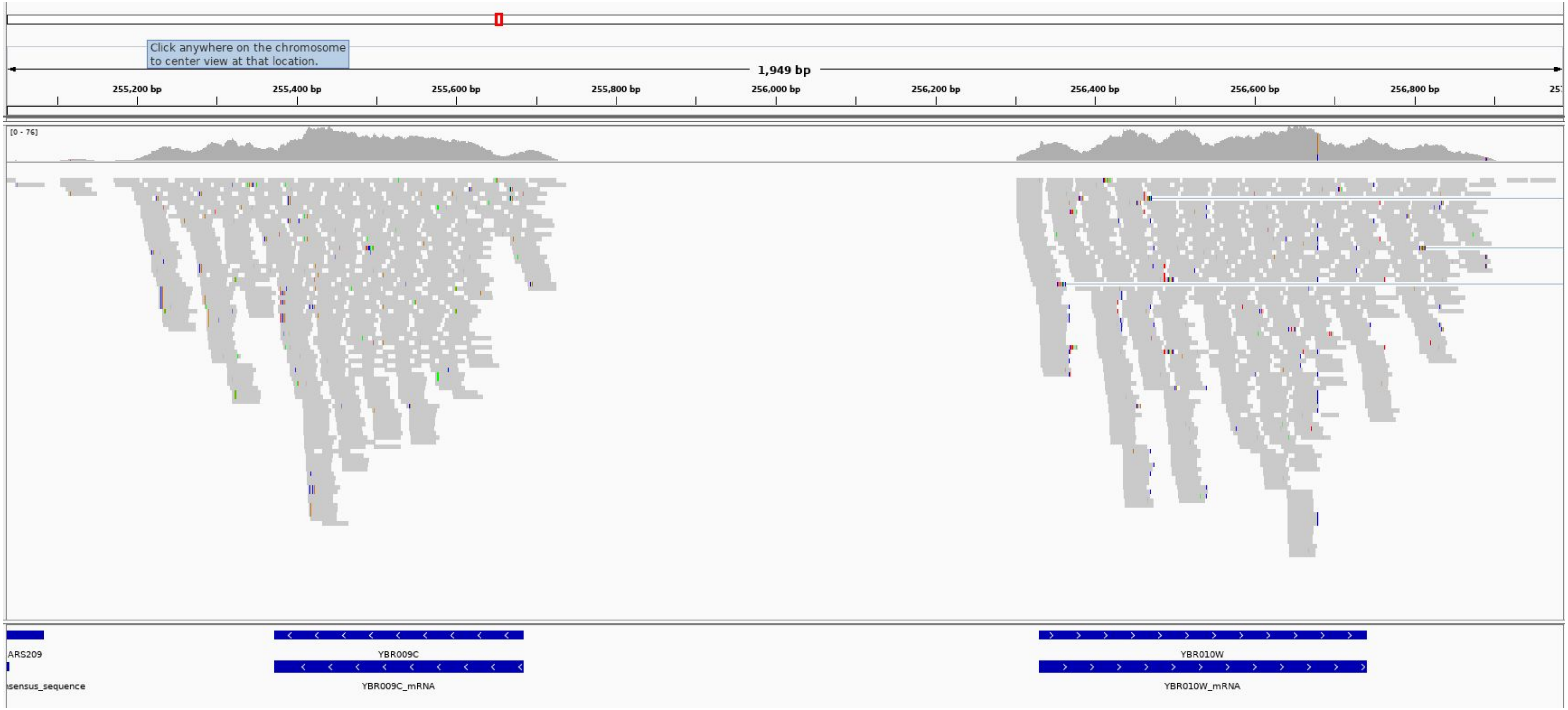
- Reference genome - fasta
- RNA-seq reads - fastq (SE or PE)
- Optional: genome annotation - GFF/GTF

Output: reads alignment in sam/bam format

STAR

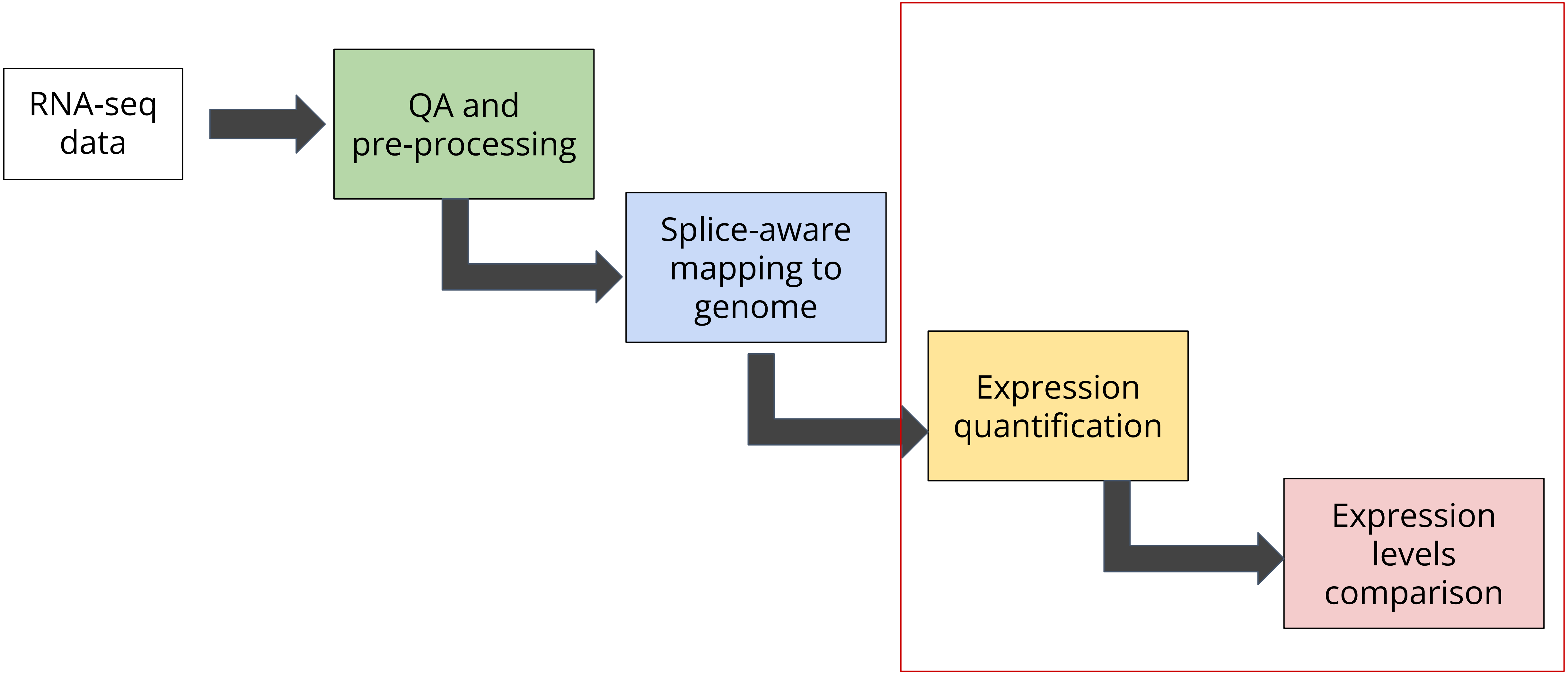



```
SRR1177156.52530710     0       ChrXI   109569  255     9M306N30M12S    *       0       0       AACGCTGAAGCTAAAGGTTTGGATGC
TACTAAATTGTACTGTAGGCACCAT      CCCFFFFFHHHHHIIIIGHIIIIIIIIIIIIIIIICFIIIIDDHIHIIIII      NH:i:1  HI:i:1  AS:i:37 nM:i:0
```



Click anywhere on the chromosome
to center view at that location.

DGE Workflow



RNA-seq data

QA and pre-processing

Splice-aware mapping to genome

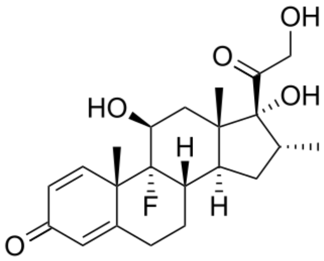
Expression quantification

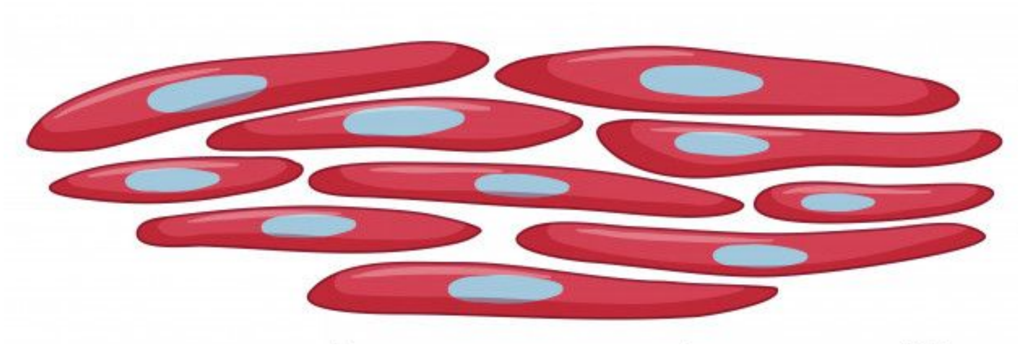
Expression levels comparison

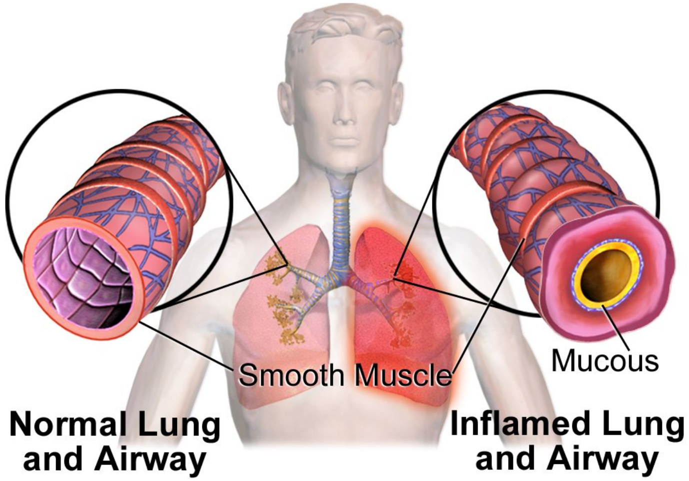
Experimental Data

The effect of the steroid Dexamethasone on human airway smooth muscle cells

Four cell lines - treated/untreated





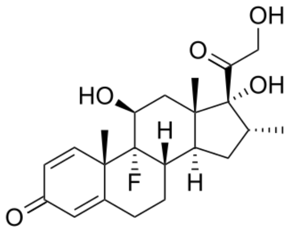


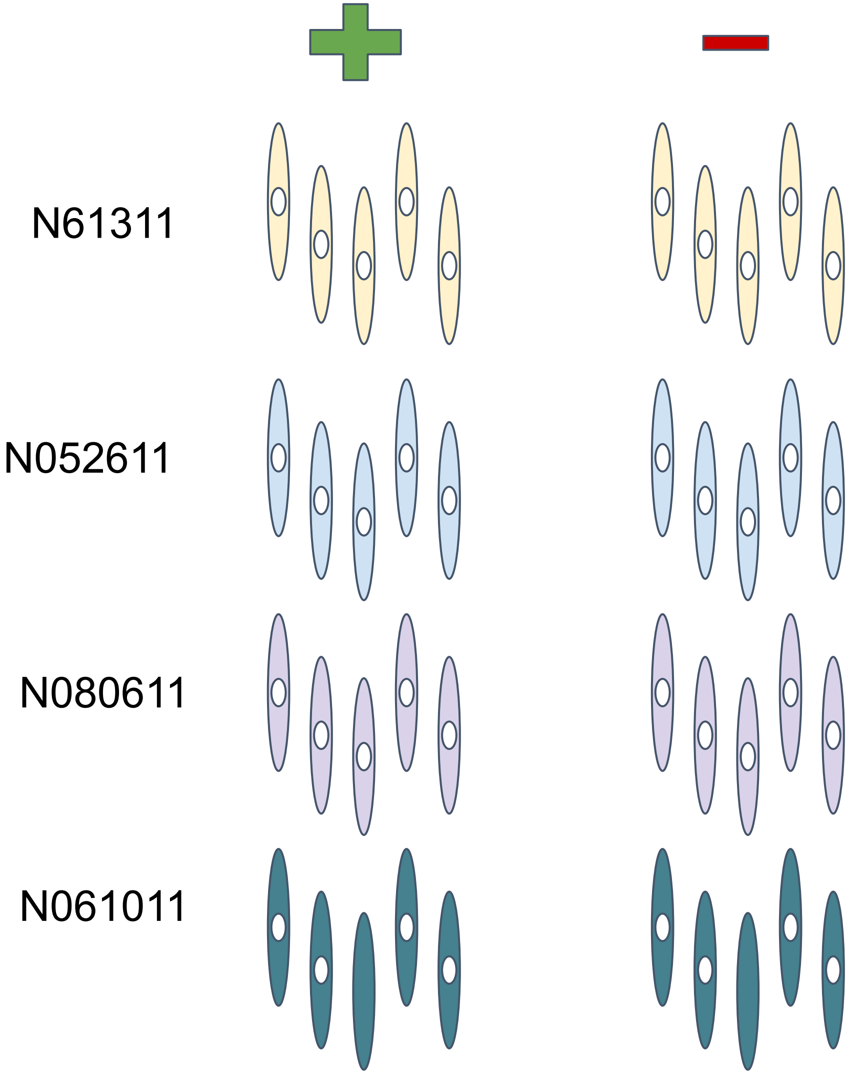
Himes, Blanca E., et al. "RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells." *PloS one* 9.6 (2014).

Experimental Data

Goal: Study the mechanism of dexamethasone action

Method: Which genes are differentially-expressed between treated and untreated samples?





+ −

N61311

N052611

N080611

N061011

Expression Quantification of Mapped Reads

Goal: determine how many RNA-seq reads mapped to each gene

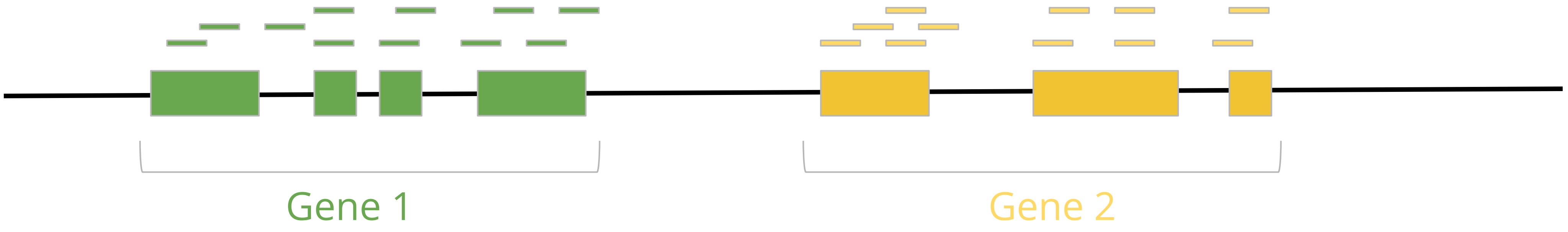
Reason: this is our proxy for gene expression level

Input:

- RNA-seq reads (spliced) mapping - BAM
- Gene annotation - GFF

Output: expression matrix M

M_{ij} = number of reads mapped to gene i in sample j



Count Matrix

Cell line	**N61311**	**N052611**	**N080611**	**N061011**	**N61311**	**N052611**	**N080611**	**N061011**
Dex treatment	+	+	+	+	-	-	-	-
C7	34	512	66	121	25	344	297	76
CCDC69	5	8	8	3	12	7	10	7
DUSP1	1112	985	1003	898	214	128	188	203
FKBP5	33	94	111	42	46	98	57	85
	...							

Total: ~64k transcripts (mRNAs)

Exploring Expression Levels Data

Useful as preparation for differential gene expression analysis

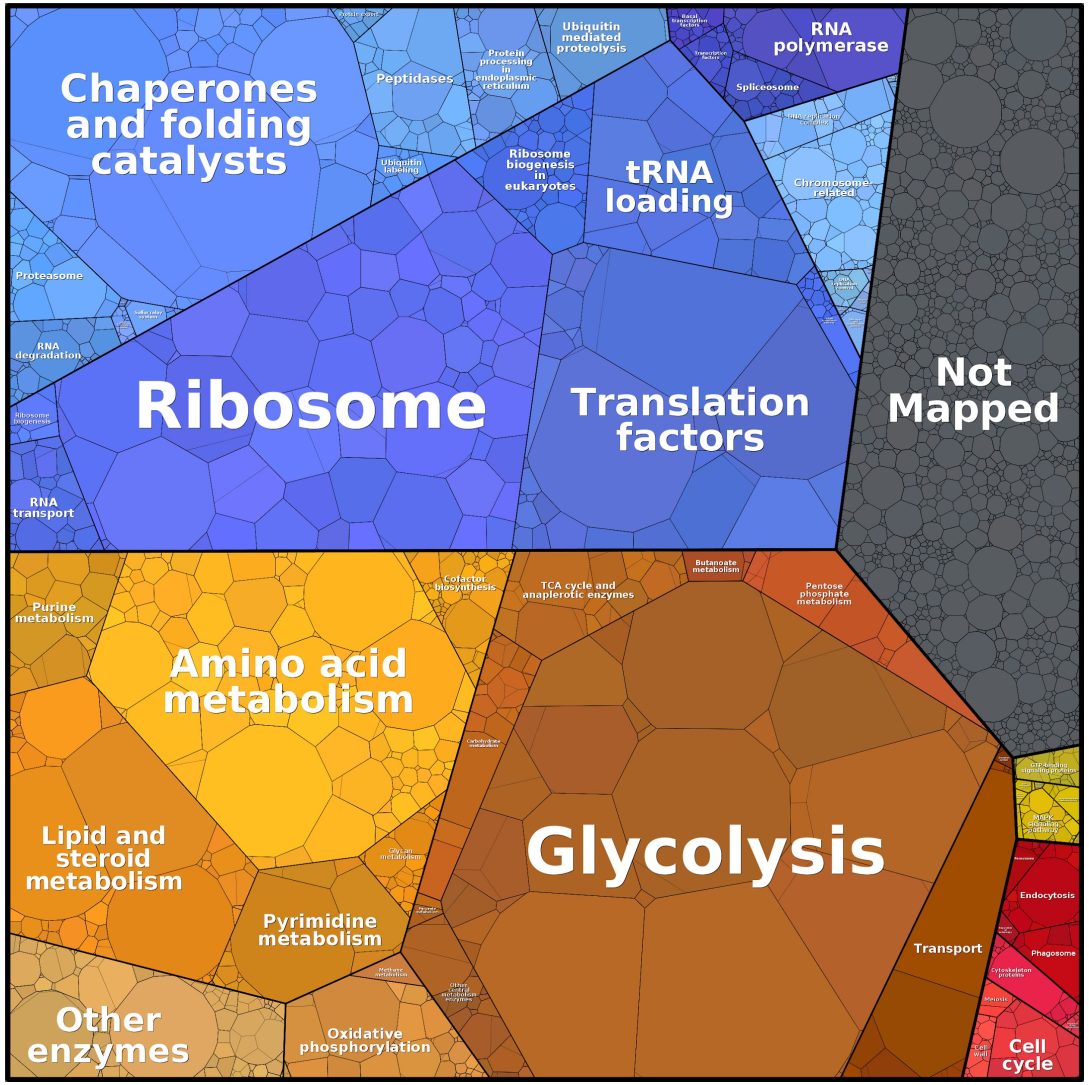
Allows detection of trends in the counts data

Basic QA

Main goal: determine overall similarity between samples

Exploring Expression Levels Data

Expression levels differ in orders of magnitude between genes

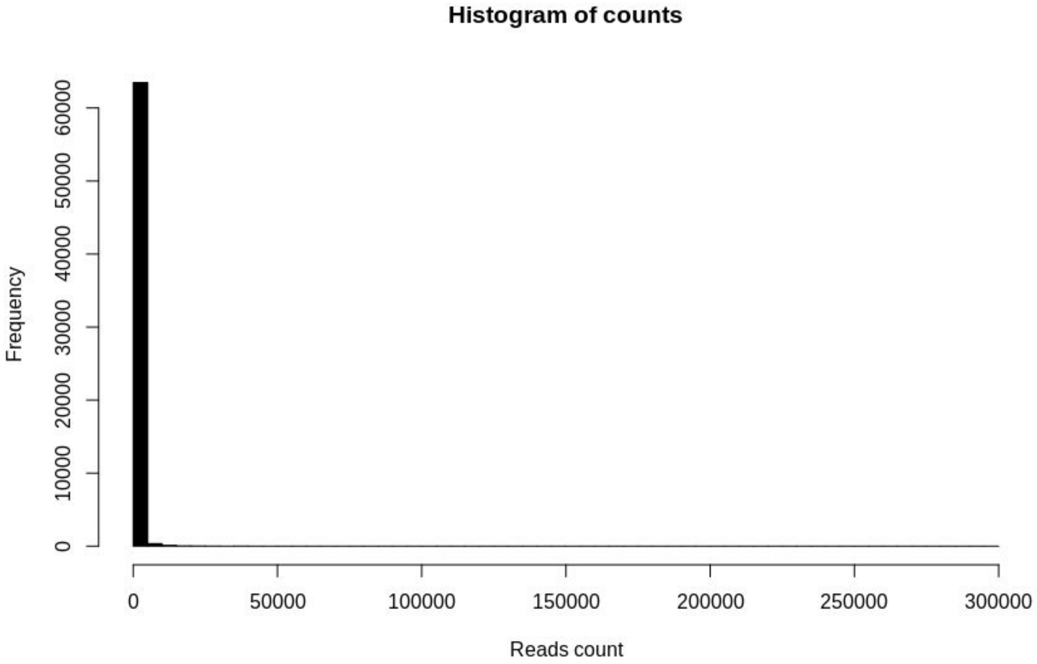


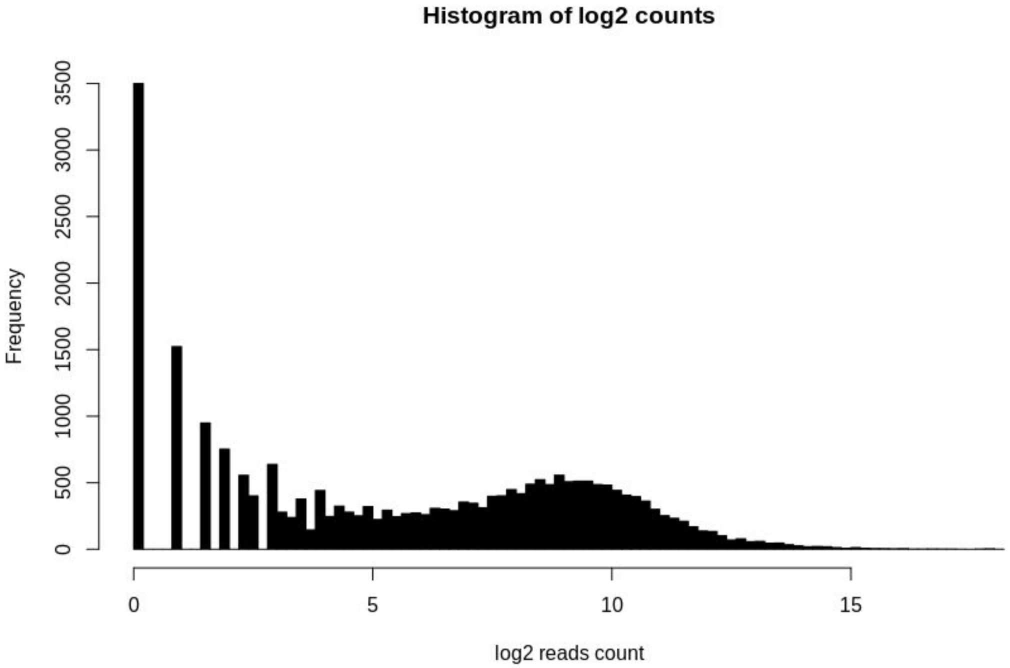
Liebermeister W., Noor E., Flamholz A., Davidi D., Bernhardt J., and Milo R. (2014), Visual account of protein investment in cellular functions. PNAS 111 (23), 8488-8493.

Log Transformation

We usually apply a $\log_2$ transformation to read counts

Makes it easier to explore the data





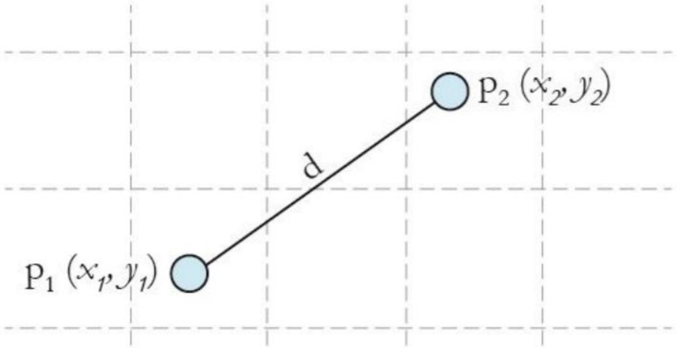
QA Expression Quantification

Which samples are overall similar/different from one another?

Does it match our expectations, given the experimental design?

We can use Euclidean distances:

In 2 dimensions



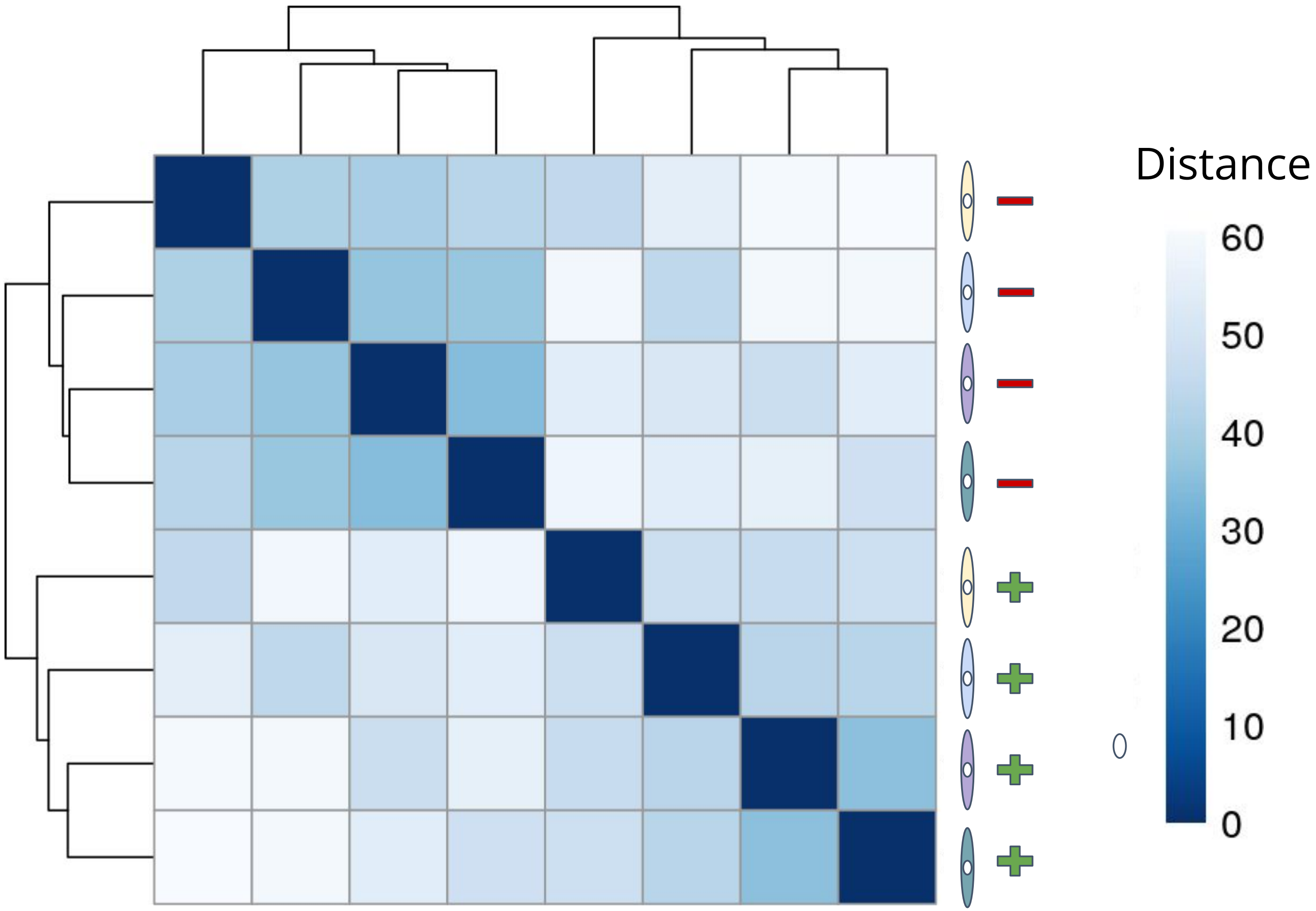
Euclidean distance (d) $= \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$

In n dimensions

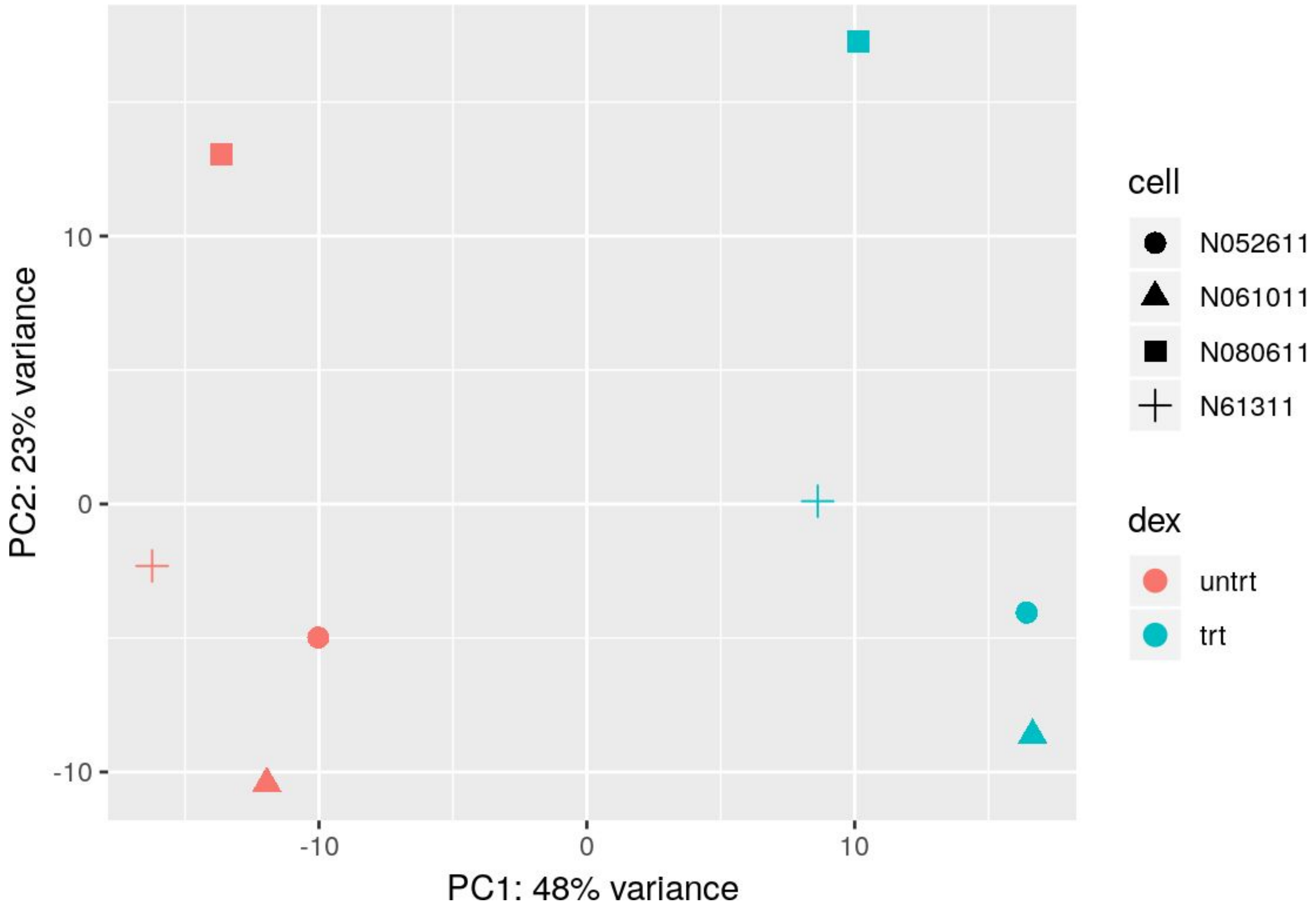
	sample 1	sample2
gene1	m_{11}	m_{12}
gene2	m_{21}	m_{22}
...	...	...
gene n	m_{n1}	m_{n2}

$$d = \sqrt{(m_{12} - m_{11})^2 + (m_{22} - m_{21})^2 + ... + (m_{n2} - m_{n1})^2}$$

Hierarchical Clustering



PCA



10
0
-10
PC2: 23% variance
-10
0
10
PC1: 48% variance
cell
N052611
N061011
N080611
N61311
dex
untrt
trt

Filtering Counts Data

It is useful to remove genes with very few reads from the analysis

- Slow down the analysis
- Reduce detection power for other genes

We won’t be able to detect DE anyway

We can choose a count cutoff

Or we can remove the Xth percentile

In the Himes et. al data:

~64k transcripts ⟶ Require >= 10 reads ⟶ ~20k transcripts

Differential Expression Analysis

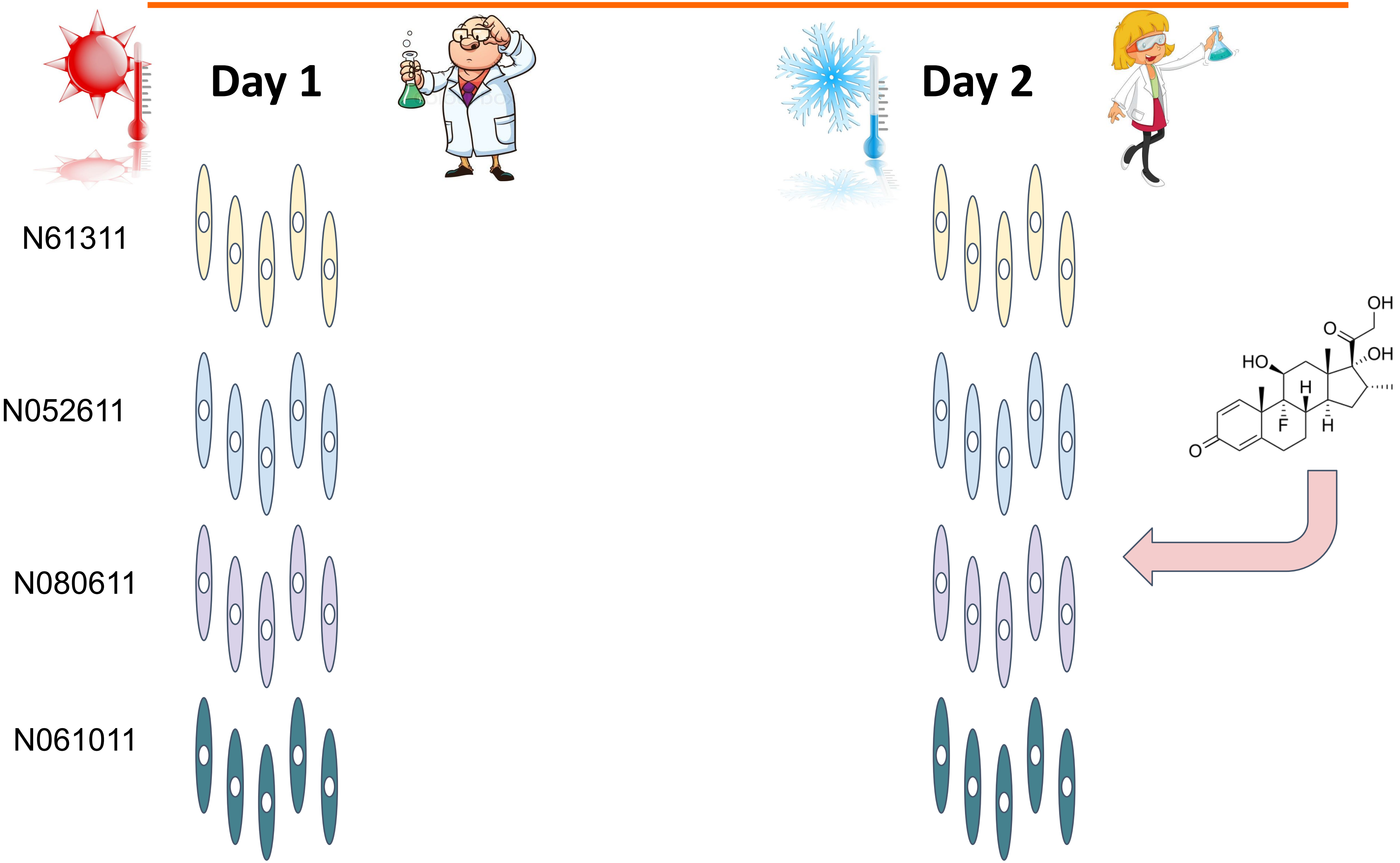
Goal: detect genes that significantly differ in their expression levels between samples

Input:

- Normalized, filtered expression quantification matrix
- Description of the experimental design

Output: per gene - estimated difference between samples and **significance level**

Batch Effects



Day 1

Day 2

N61311

N052611

N080611

N061011

Batch Effects

Introduced during sample handling and preparation

- Technical factors
- External factors

Minimize by:

- Use the same protocol for all samples
- Prepare all samples together

Not always possible

We must test for batch effects when performing DE statistical tests

Batch Effects

Sample	Cell line	Dex	Batch
SRR1039508	N61311	Untreated	1
SRR1039509	N61311	Treated	1
SRR1039512	N052611	Untreated	1
SRR1039513	N052611	Treated	1
SRR1039516	N080611	Untreated	2
SRR1039517	N080611	Treated	2
SRR1039520	N061011	Untreated	2
SRR1039521	N061011	Treated	2

Fold Change

The main measure used in DGE analysis is **fold change** - a.k.a ratio

Ratios are highly non-symmetric

$$R = \frac{Count_{sample1}}{Count_{sample2}}$$

Therefore we use log scaling - **log2 fold change (L2FC)**

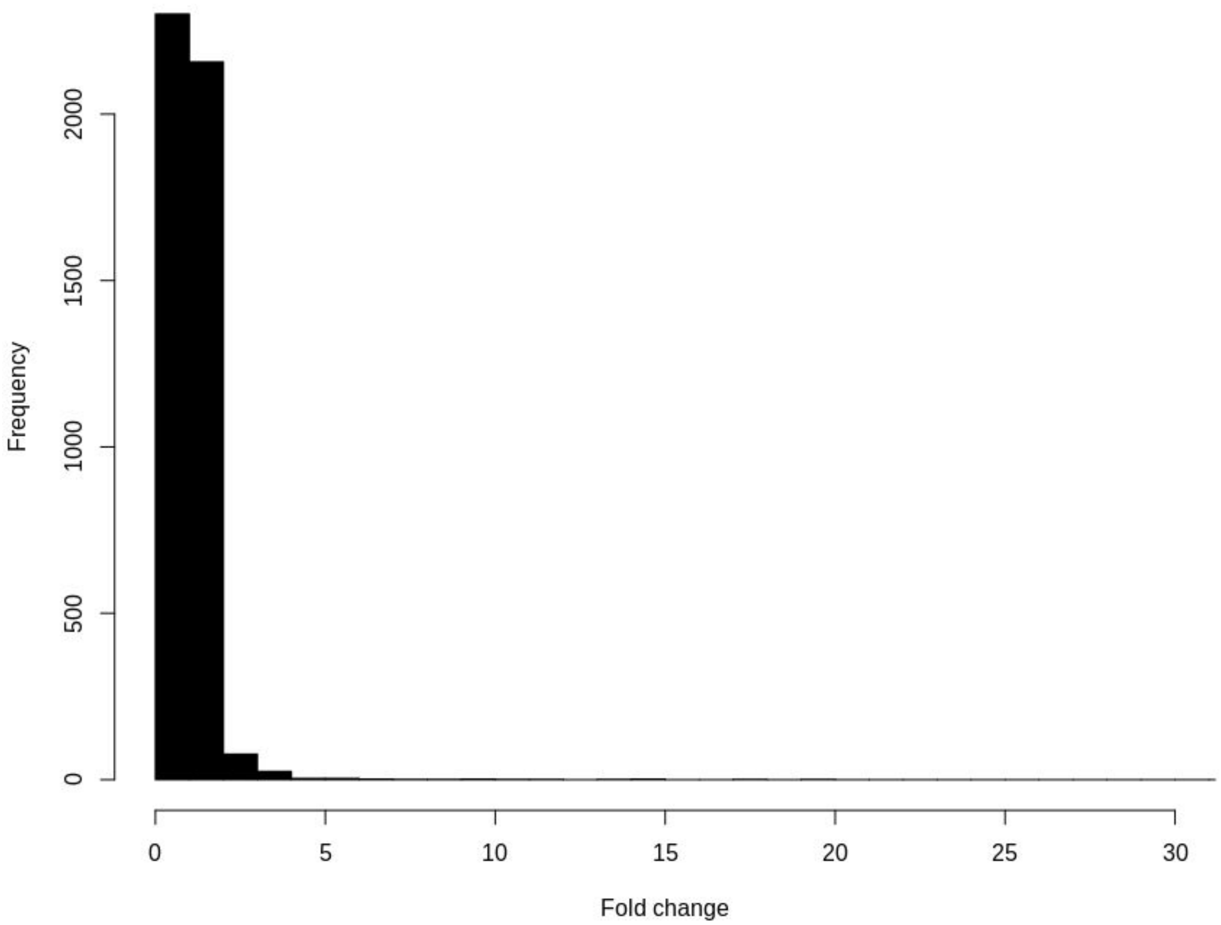
$$L2FC = \log_2(\frac{Count_{sample1}}{Count_{sample2}}) = \log_2 Count_{sample1} - \log_2 Count_{sample2}$$

Fold Change

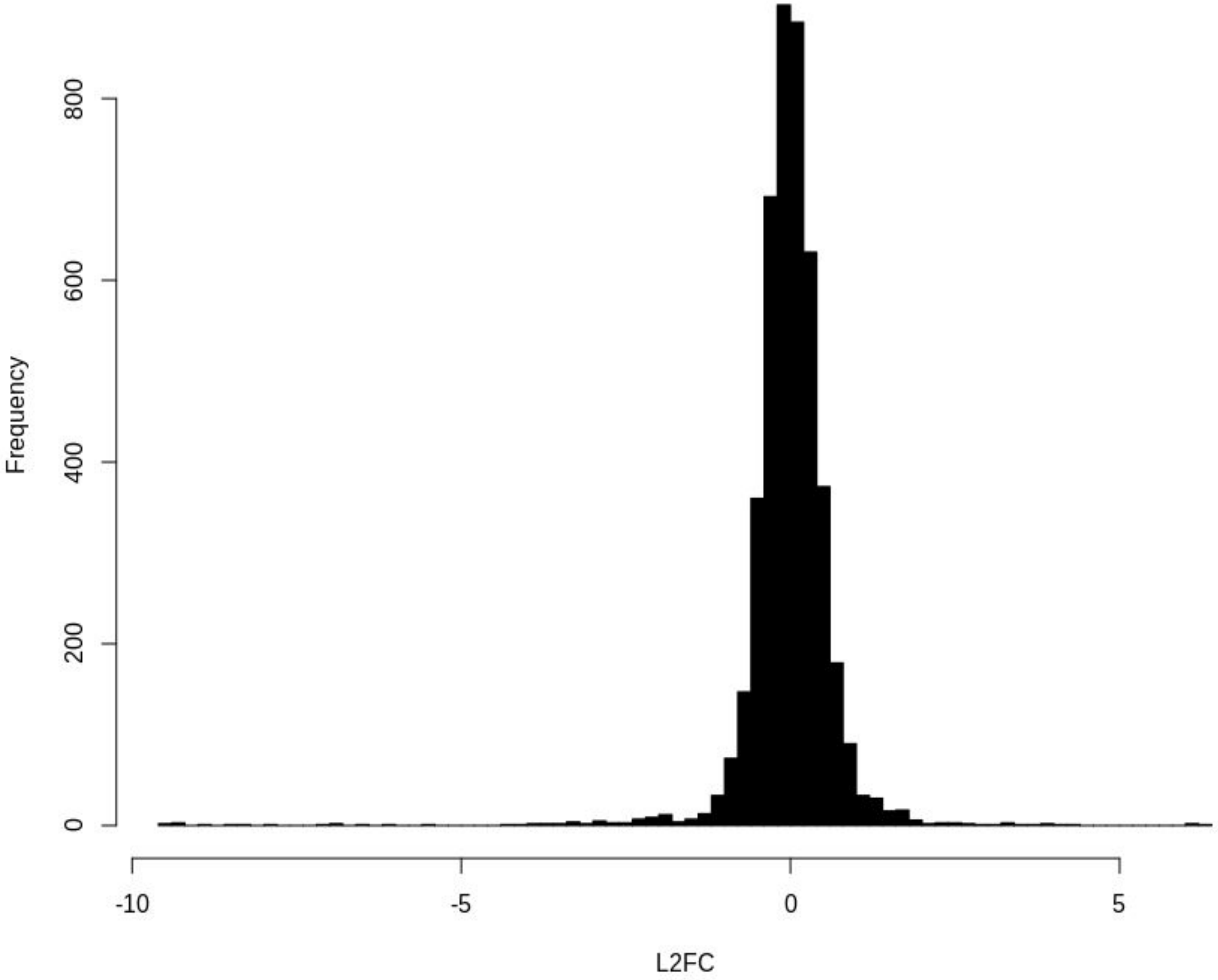
	N61311 Dex	N61311 Dex	N052611 Dex	N080611 Dex	N61311 Unt	N61311 Unt	N052611 Unt	N080611 Unt	Mean Dex	Mean Unt	FC	L2FC
Gene 1	34	512	66	121	25	344	297	76	183.25	185.50	0.99	-0.02
Gene 2	1112	985	1003	898	214	128	188	203	999.50	183.25	5.45	2.45
Gene 3	6	9	4	6	12	9	15	16	6.25	13.00	0.48	-1.06

L2FC = 0 : no difference
L2FC > 0 : sample1 expression > sample 2 expression
L2FC < 0 : sample1 expression < sample 2 expression

Fold Change



Frequency
0
500
1000
1500
2000
0
5
10
15
20
25
30
Fold change



Frequency
0
200
400
600
800
-10
-5
0
5
L2FC

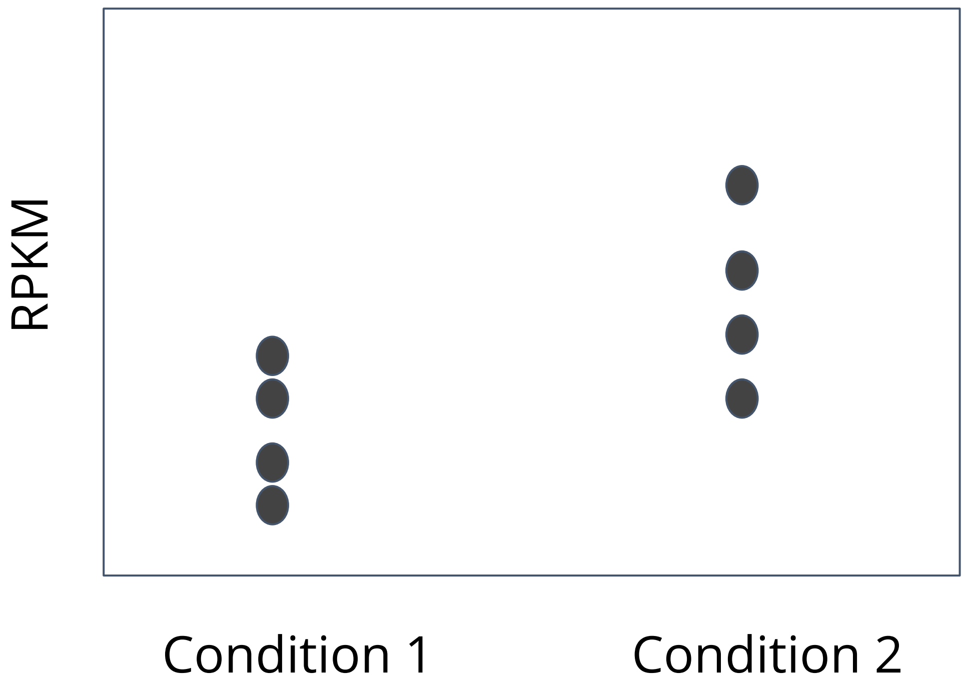
Log2 Fold Changes

Can we tell just by looking at L2FC values?

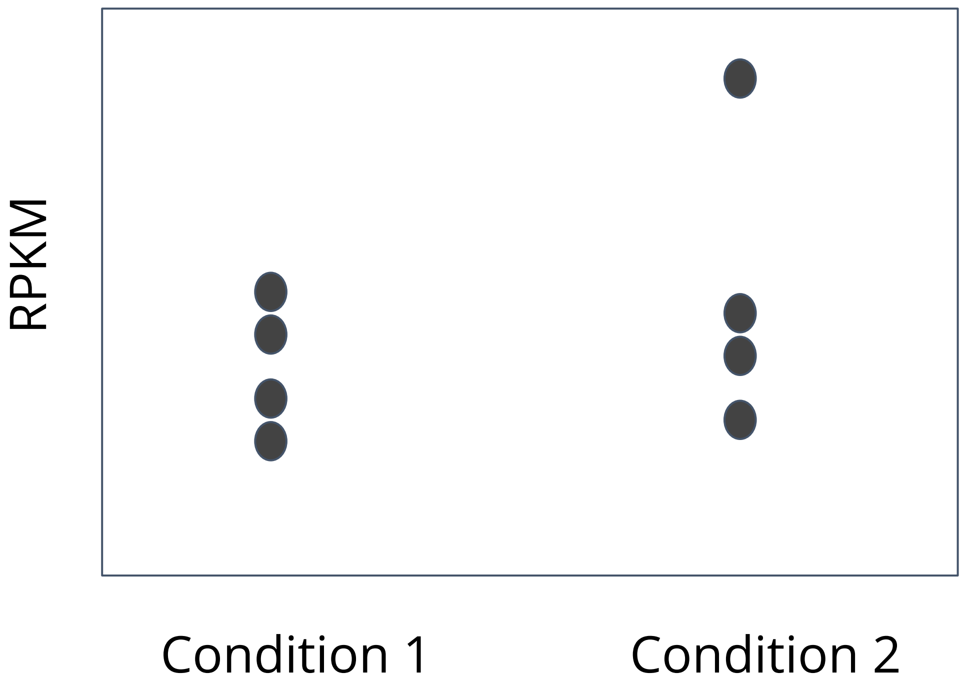
Maybe it’s just random noise?

Replicates can help

Gene X - L2FC = 0.5



Gene Y - L2FC = 0.4



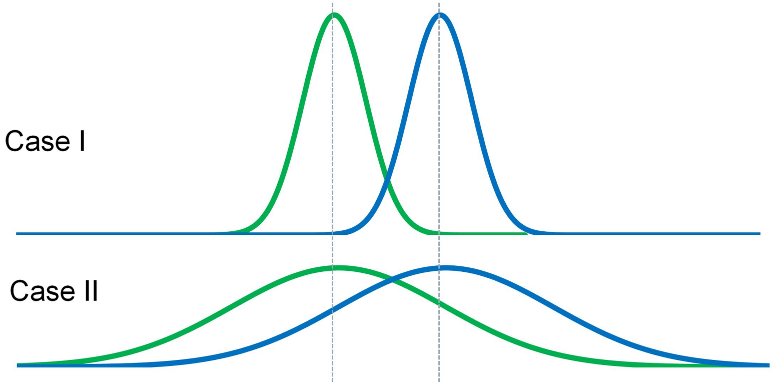
Hypothesis Testing

H_0: there is no difference in expression levels between samples
H_1: expression levels differ between samples

We try and reject H_0 with an appropriate statistical test, e.g.:

- Parametric tests: *t*-test, ANOVA
- Non-parametric tests: Mann-Whitney U test
- Other modeling methods: linear models, GLM

The result is a significance score - **per gene p-value**

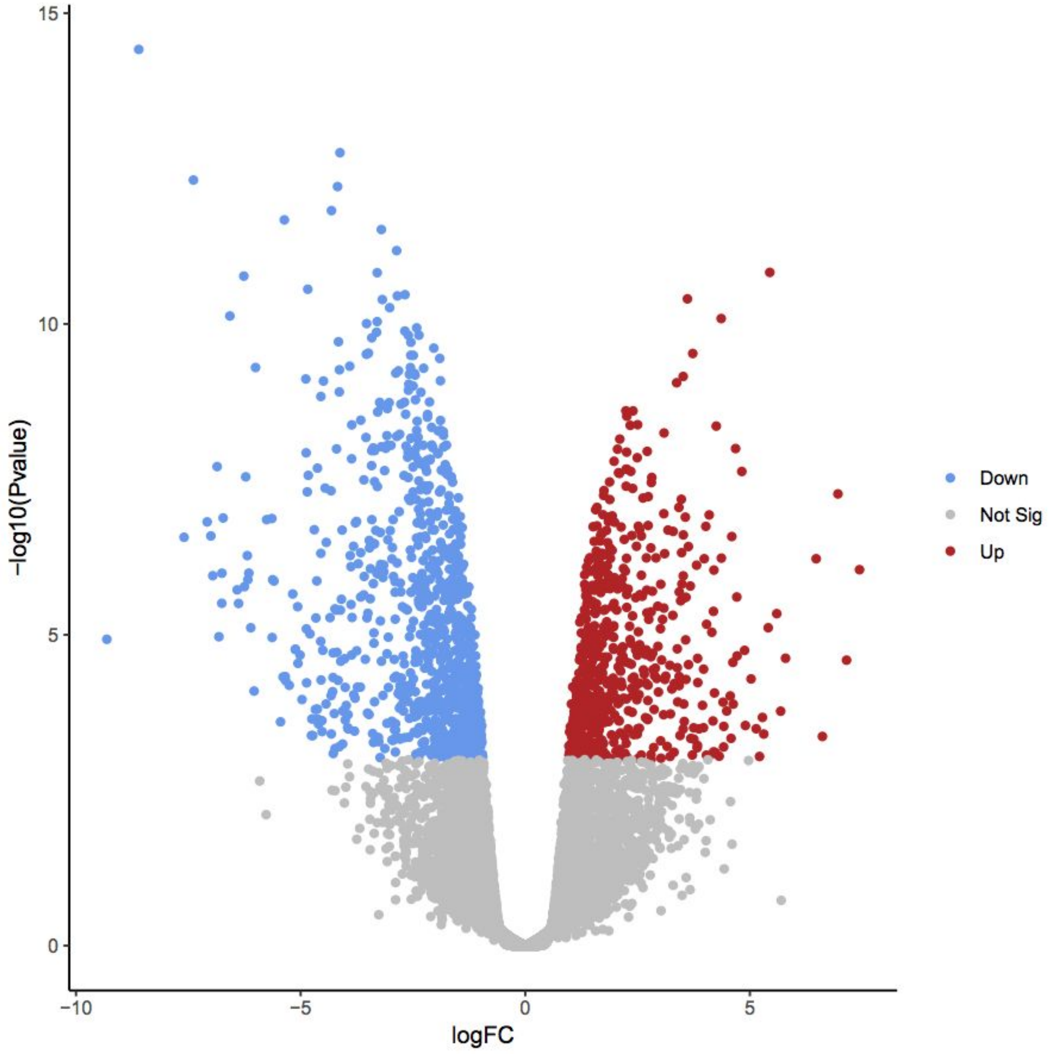


Volcano Plots

For each gene, we must consider both L2FC and p-value

To get a global view - use a **volcano plot**

We can choose a p-value cutoff, e.g. 0.05 or 0.01



Cutoff

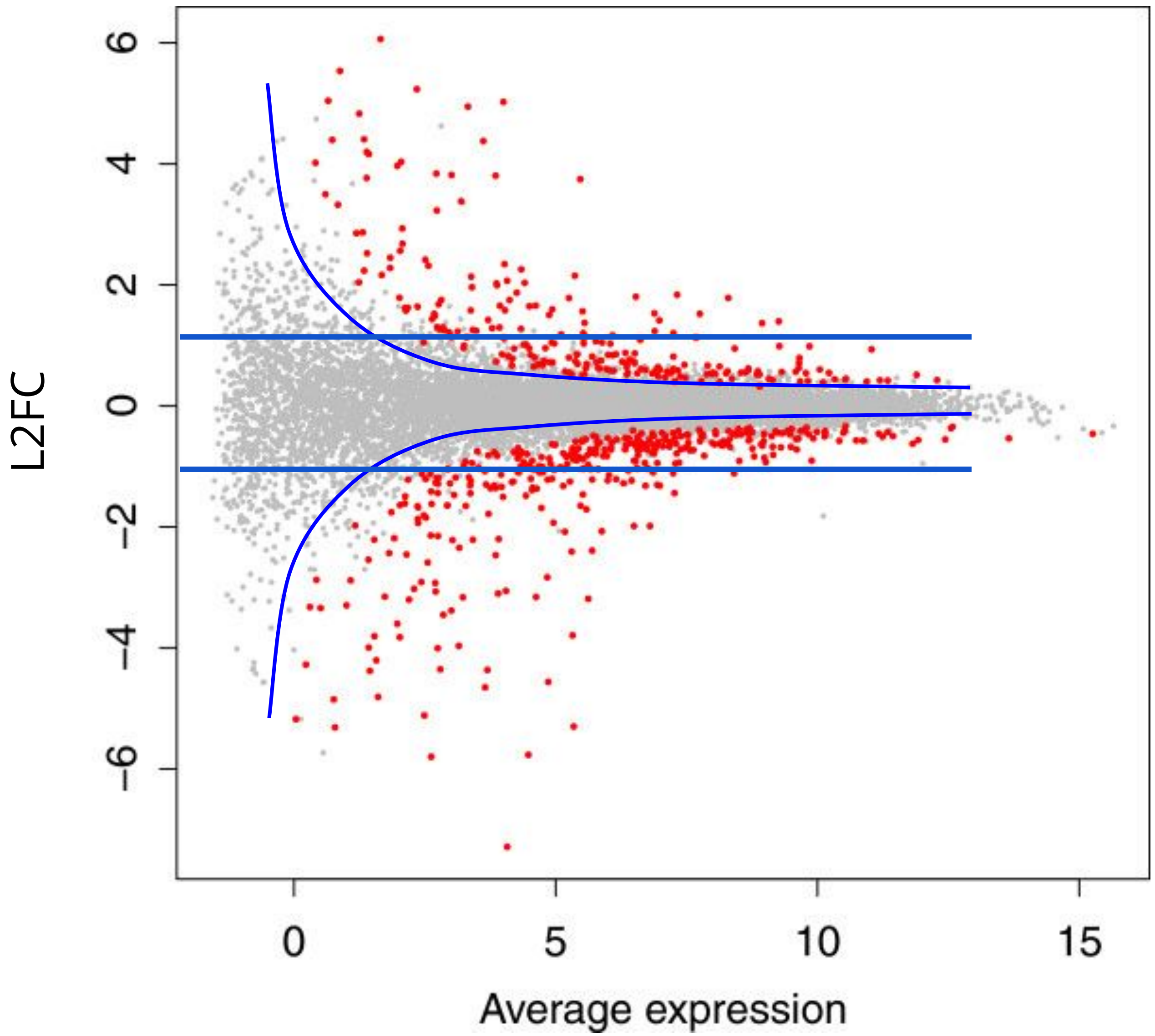
We can choose an arbitrary cutoff, e.g. 1

Lowly-expressed genes usually display higher variability in expression

This can be seen in a **MA-plot**

Can be used as a sanity-check

Or to choose dynamic L2FC cutoffs



Correcting for Multiple Testing

Recall the meaning of a p-value…

Since we are performing multiple tests, we must correct (adjust) p-values

The simple and stringent way - Bonferroni correction

$$p_{adj} = p \times [\# \; of \;\; genes]$$

The common way - False Discovery Rate (FDR - BH procedure):

1. Order p-values from smallest to largest

   $p_1, p_2, \dots, p_k, \dots p_m$

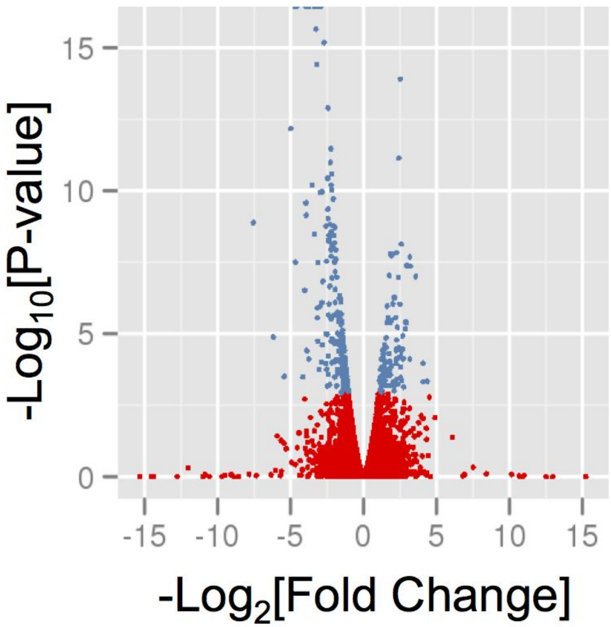
2. $$p_{adj}^{k} = \frac{p^{k} \times [\# \; of \;\; genes]}{k}$$

Himes et al. - DGE Analysis Results

A total of 316 DE genes between treated and untreated samples

Top 5:

Gene	Dex RPKM	Untreated RPKM	Ln[Fold Change]	Test Statistic	Adj. P-Value
C7	38.41	3.76	-3.35	8.74	0
CCDC69	47.39	6.24	-2.92	8.61	0
DUSP1	144.96	18.26	-2.99	8.99	0
FKBP5	53.05	3.43	-3.95	10.52	0
GPX3	613.37	45.18	-3.76	9.19	0


$-\text{Log}_{10}$[P-value]
$-\text{Log}_2$[Fold Change]