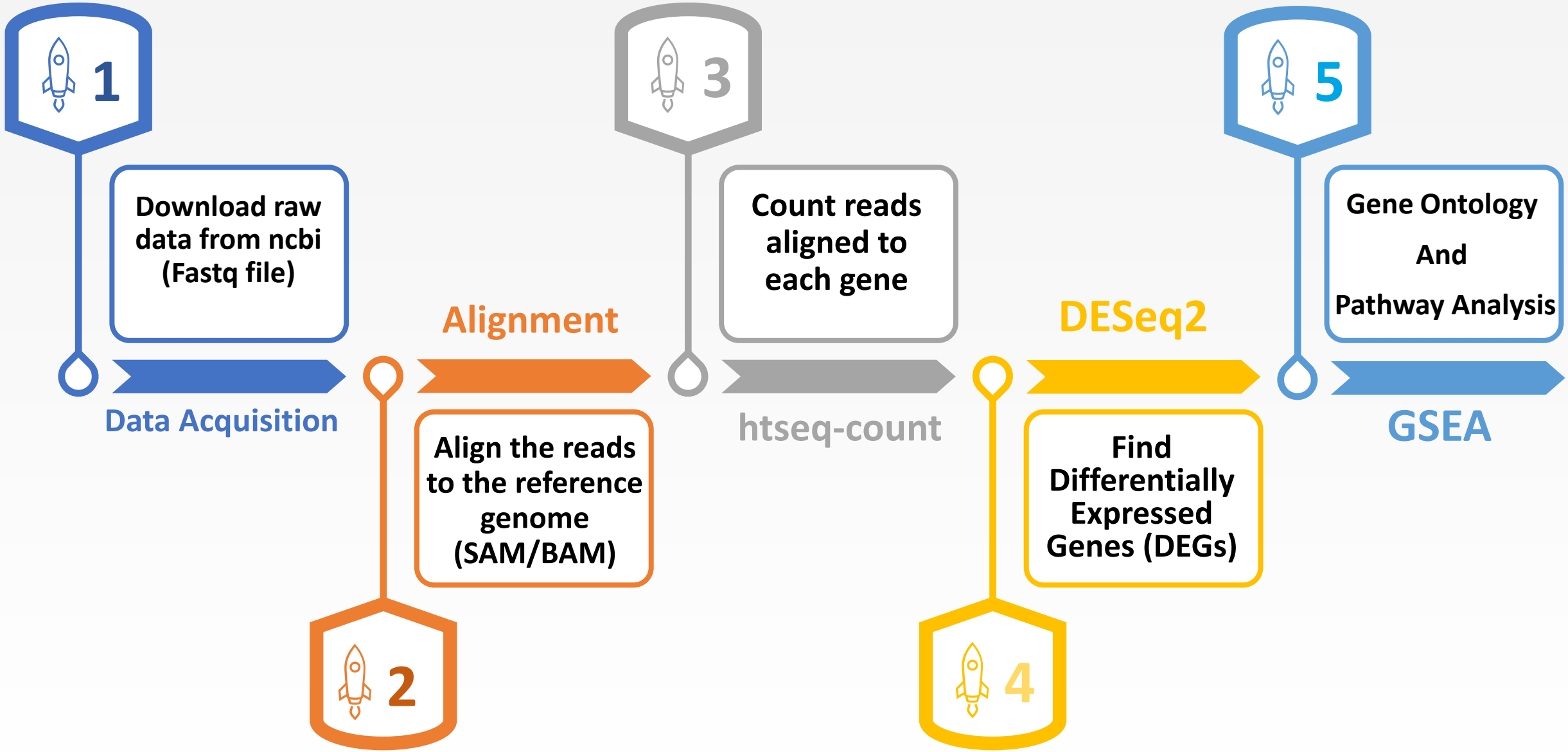
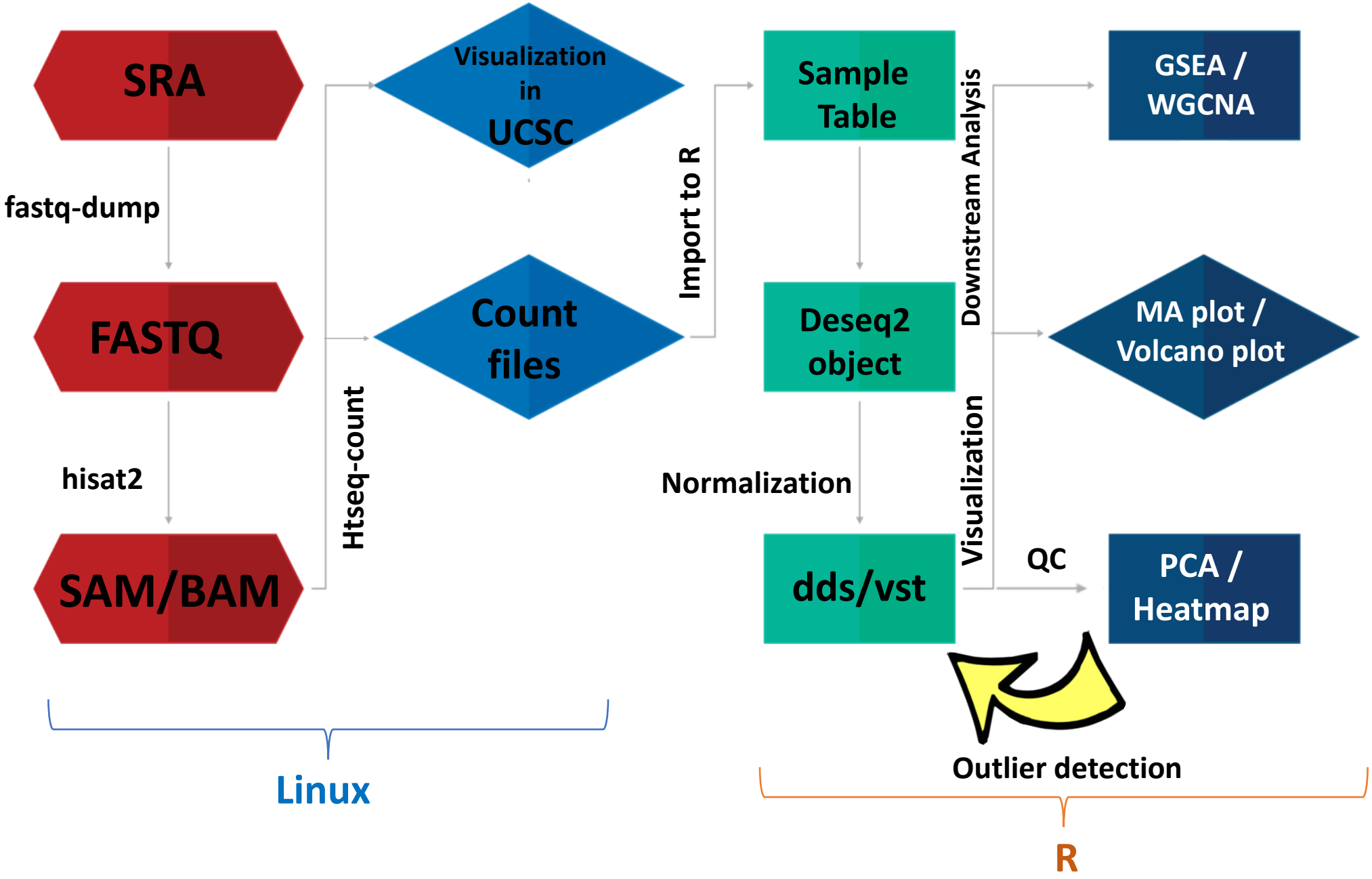
INTRODUCTION TO

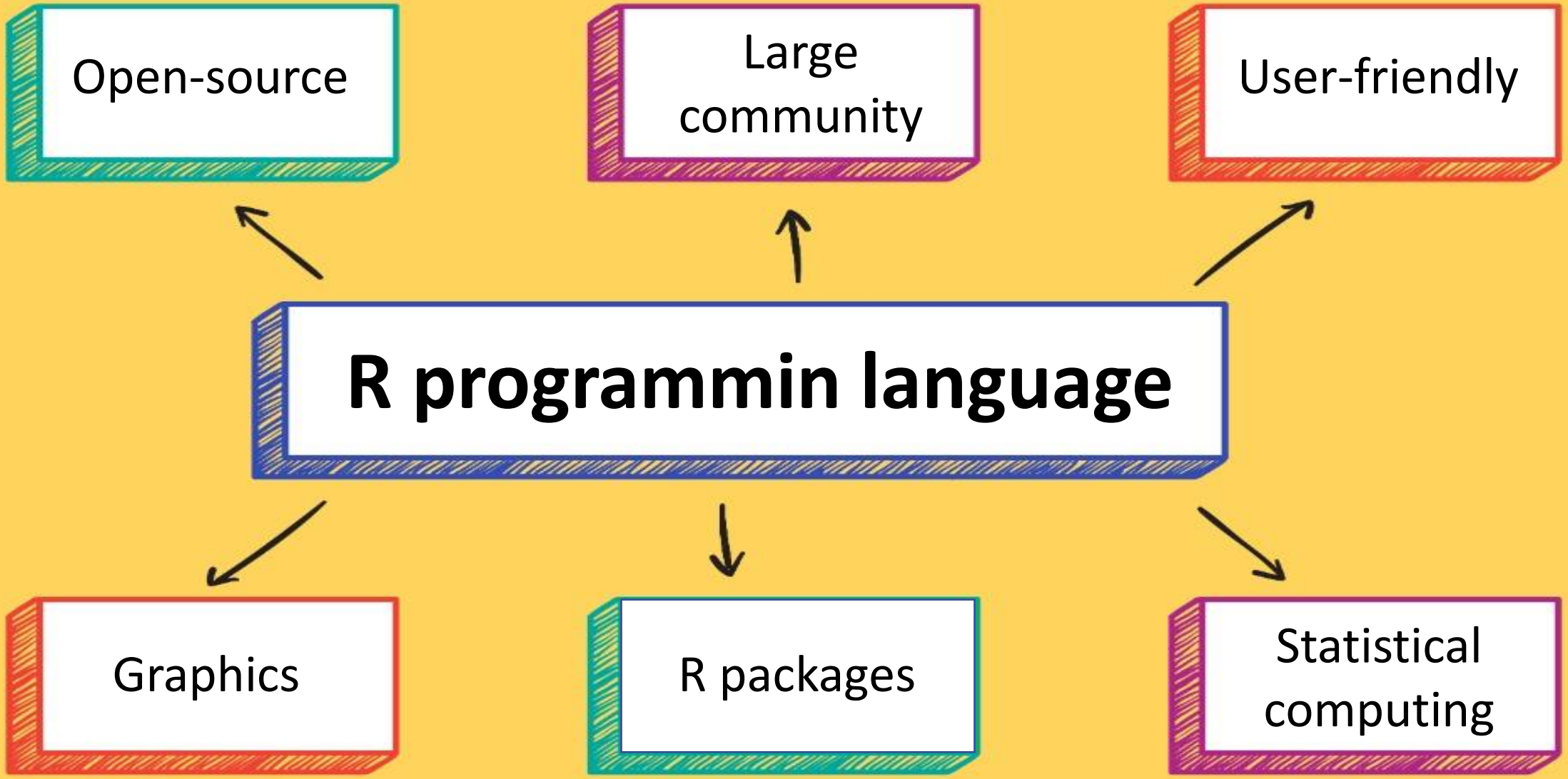
R Programming

RNA-seq Workshop
ACECR, Mashhad
October 2021

RNA-seq Pipeline




SRA
fastq-dump
FASTQ
hisat2
SAM/BAM
Visualization in UCSC
Count files
Htseq-count
Import to R
Sample Table
Deseq2 object
Normalization
dds/vst
Downstream Analysis
Visualization
QC
GSEA / WGCNA
MA plot / Volcano plot
PCA / Heatmap
Outlier detection
Linux
R


Open-source
Large community
User-friendly
R programmin language
Graphics
R packages
Statistical computing

R vs. R Studio

R: Engine

RStudio: Dashboard

How R thinks?

Two slogan of R:

1) everything that exist in an "Object"

2) everything that happen is a "Function" call

Three parts of a function :

- command
- parathesis
- arguments inside paranthesis

(these are not always present)

Statistics in R

var(x)
sd(x)
mean(x)
weighted.mean(x)
geometric.mean(x)
median(x)
t.test()
anova()
….

Data types in R

- ✓Numeric
- ✓Character
- ✓Logical

Vector

Central component of R

✓Vectors are the **most basic data structure in R**.

✓These structures allow to **concatenate data**

of the same type.

A list of numbers, such as (1,2,3,4,5)

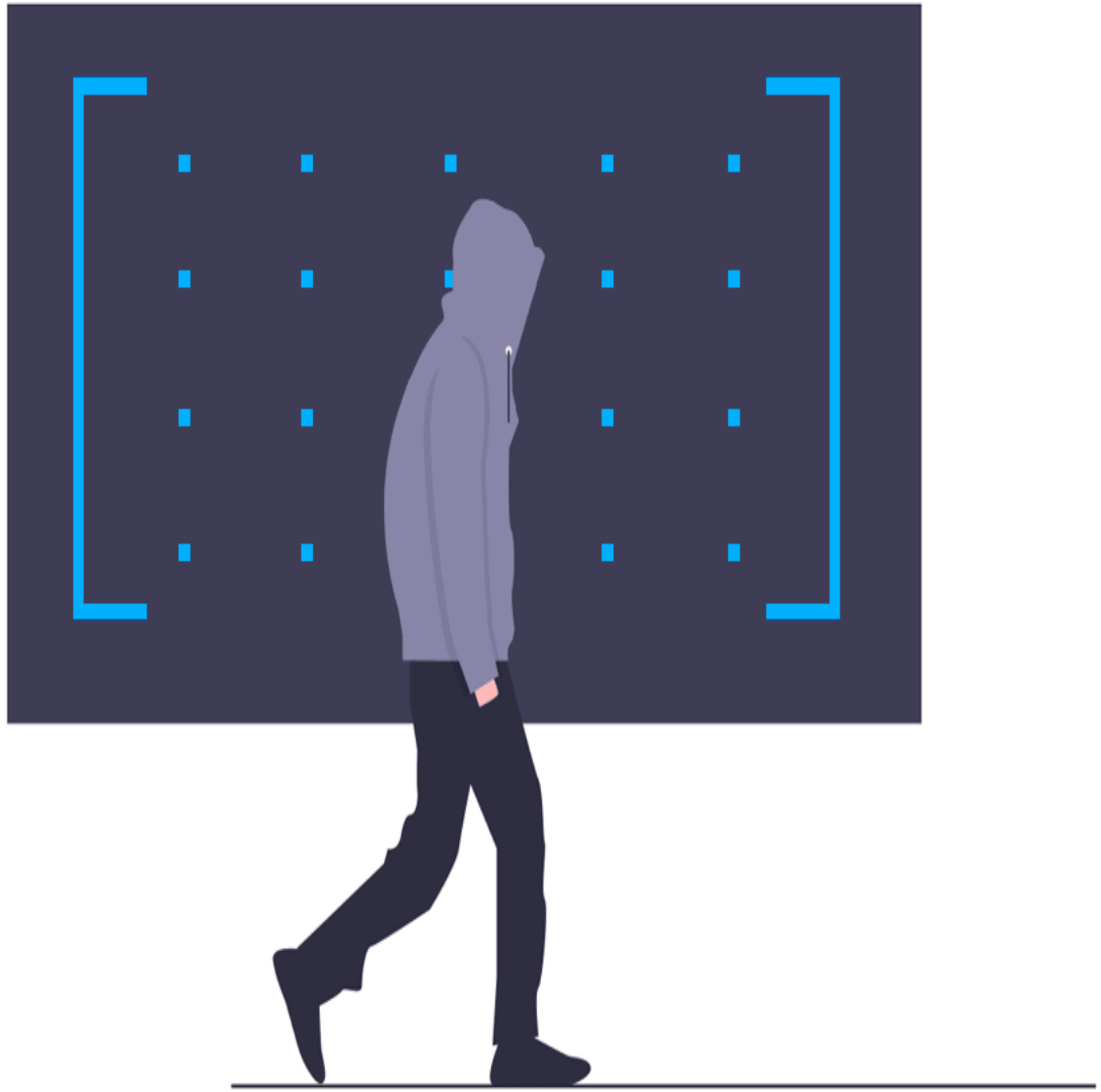
```
> a <- c(1,2,3,4,5)
```

- Command c creates a vector that is assigned to object a

Matrix

- ✓ A vector with an additional attribute
- ✓ A data structure for **storing objects of the same type**
- ✓ The matrix function allows creating a matrix data structure in R, passing a numeric, character or logical vector.

```
m<- matrix(data=5,nrow=2,ncol=2)
```



Data Frame

A table where columns can contain numeric and string values

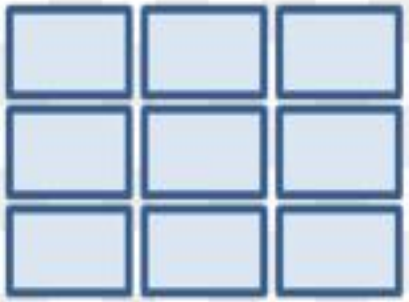
> df <- data.frame(a, b)

Data Structure in R

Vector

- 1 column or row of data
- 1 type (numeric or text)

Matrix



- multiple columns and/or rows of data
- 1 type (numeric or text)

Data Frame

- multiple columns and/or rows of data
- multiple types

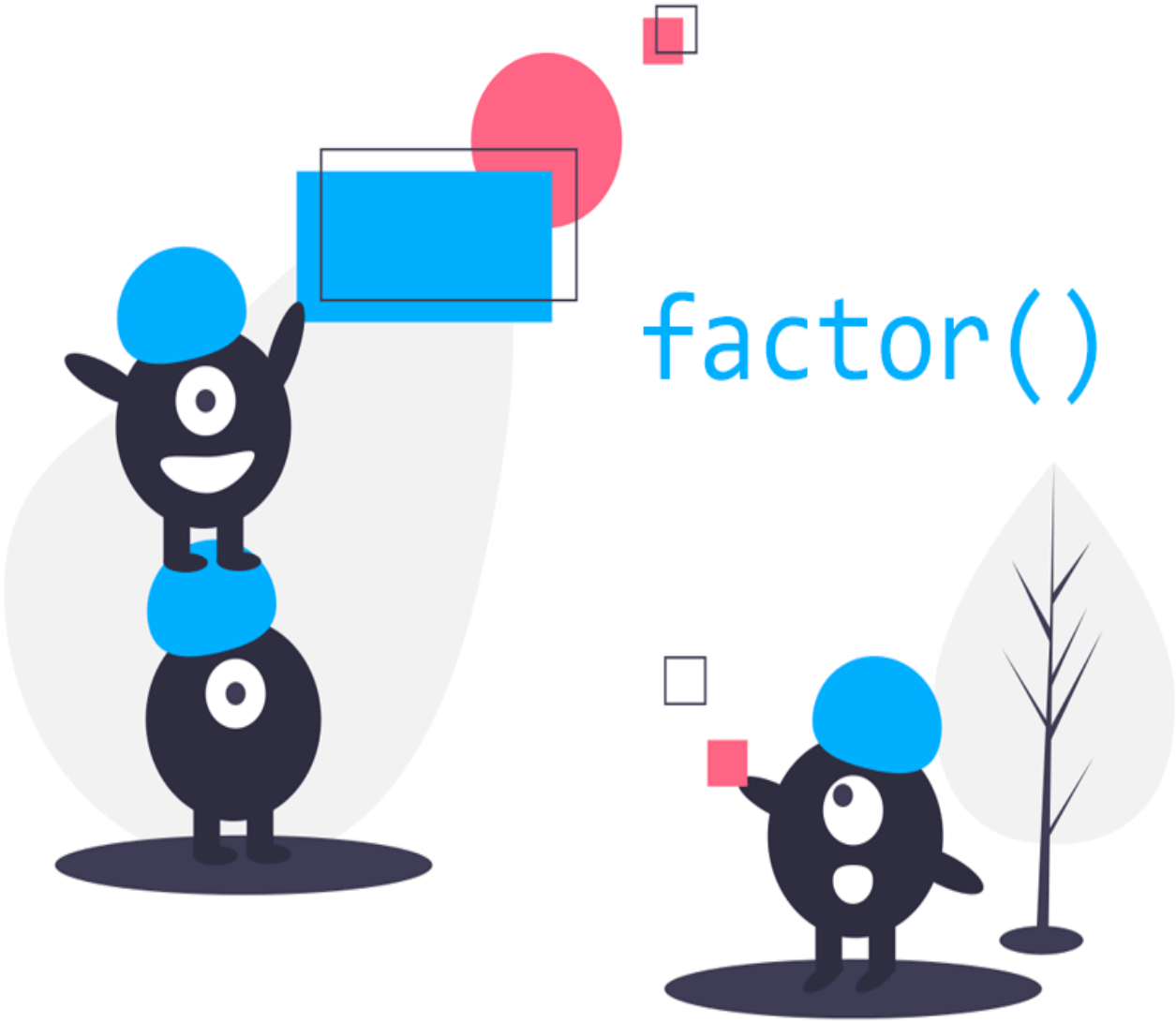
Factor

- **represent categorical data**
  - A list of levels, either numeric or string
  - > b<-factor(1:3)
  - > c<-factor(”Male”,”Female”)

- Possible conversion:

➢ Characters to factors

➢ Numeric to factors



R packages

CRAN is the official R repository!

R package= **a library of functions** (developed to cover some needs or specific scientific methods that are not implemented in base R)

R vs. R packages

R: A new phone	R Packages: Apps you can download
	

Input and Output

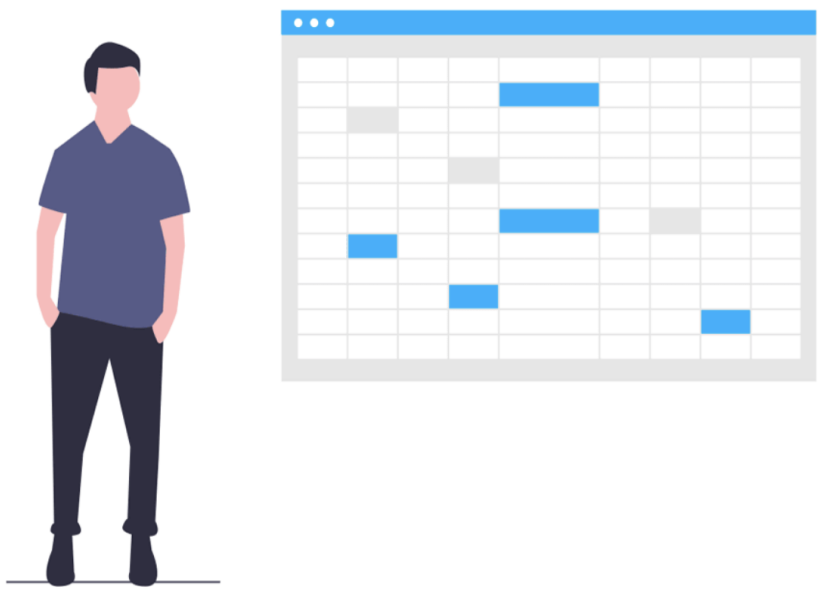
- Spreadsheet (Excel)
- Binary files
- Databases
- URL

Tabular data files

- Every line is a record
- In every record we have items separated by delimiter
- Every record have the same column

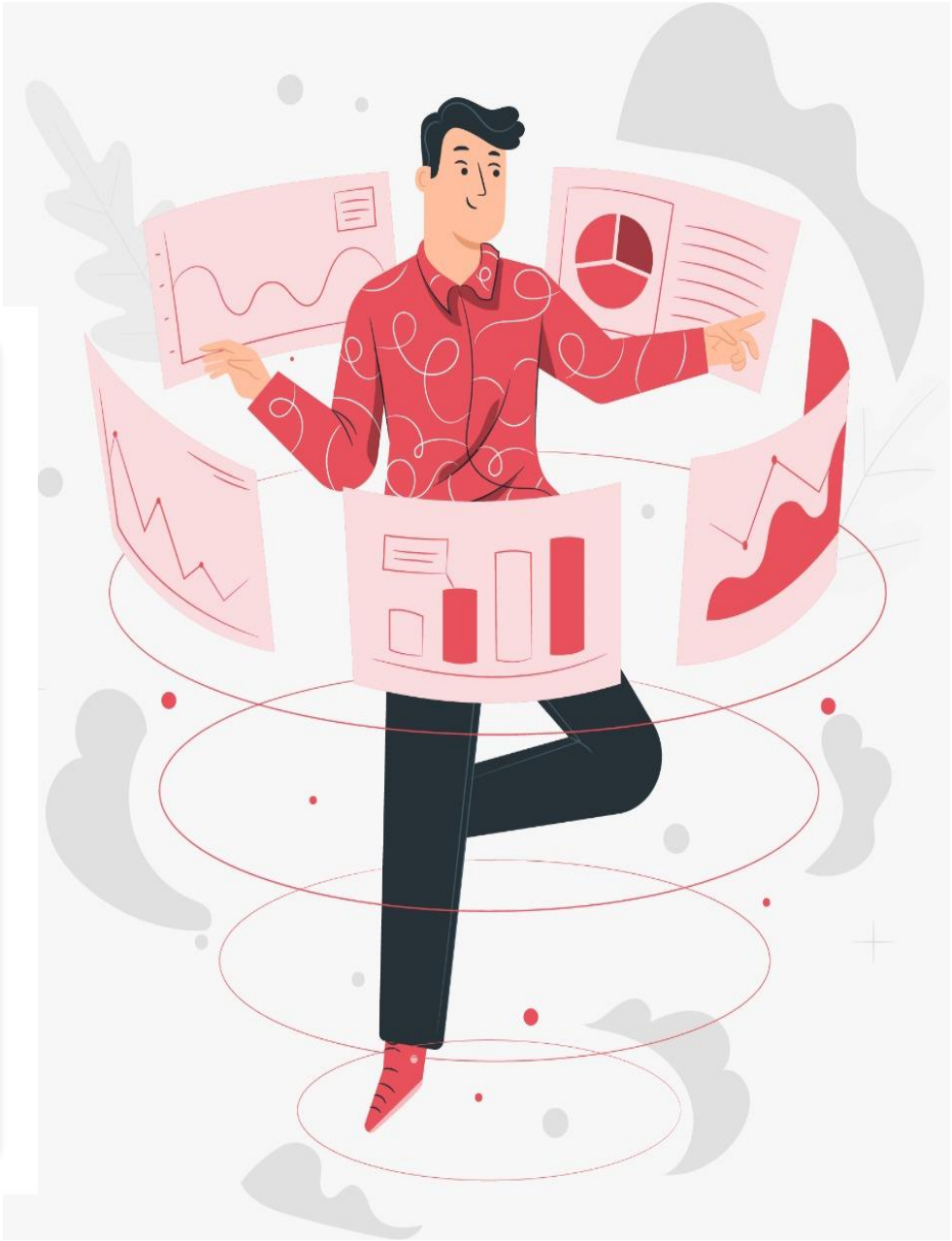
	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amount
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80

Function	Header	Sep	Dec
read.csv	TRUE	","	"."
read.csv2	TRUE	";"	","



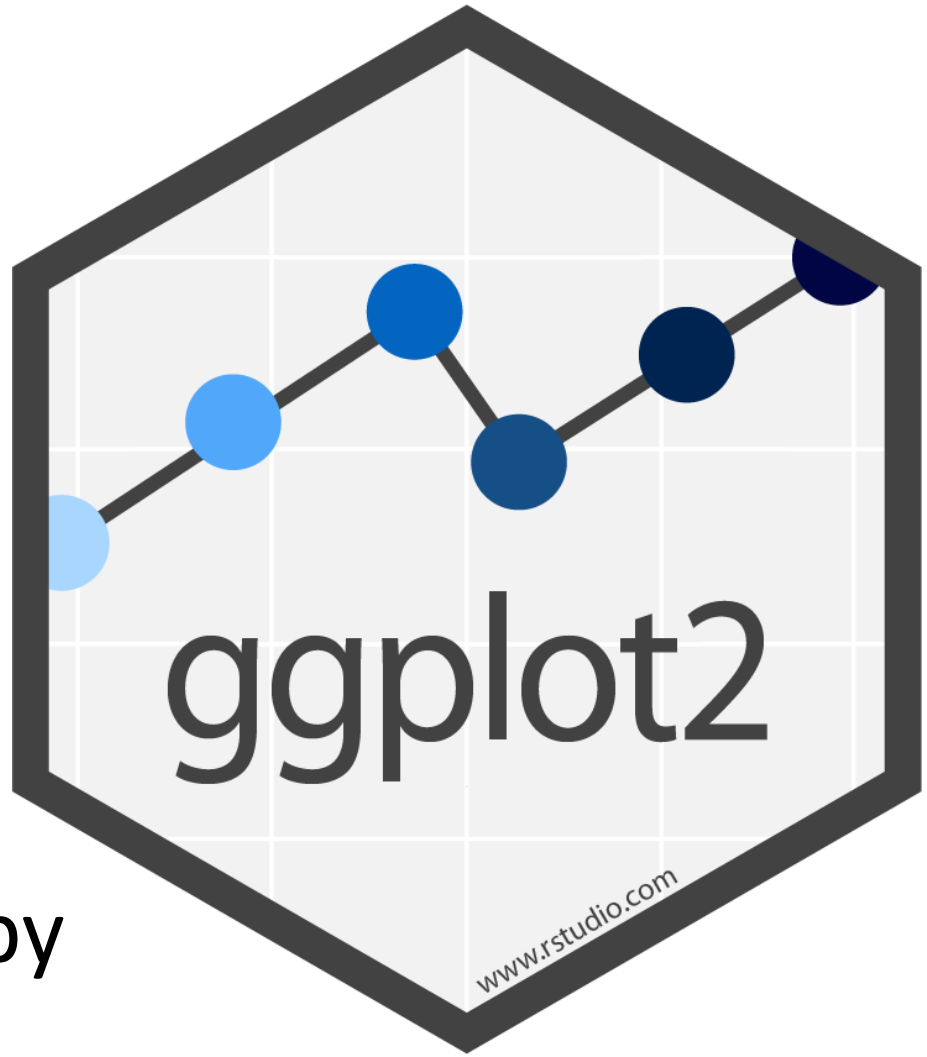
Data Visualization

Function and arguments	Output plot
plot(x, y)	Scatterplot of x and y numeric vectors
plot(factor)	Barplot of the factor
plot(factor, y)	Boxplot of the numeric vector and the levels of the factor
plot(time_series)	Time series plot
plot(data_frame)	Correlation plot of all dataframe columns (more than two columns)
plot(date, y)	Plots a date-based vector
plot(function, lower, upper)	Plot of the function between the lower and maximum value specified



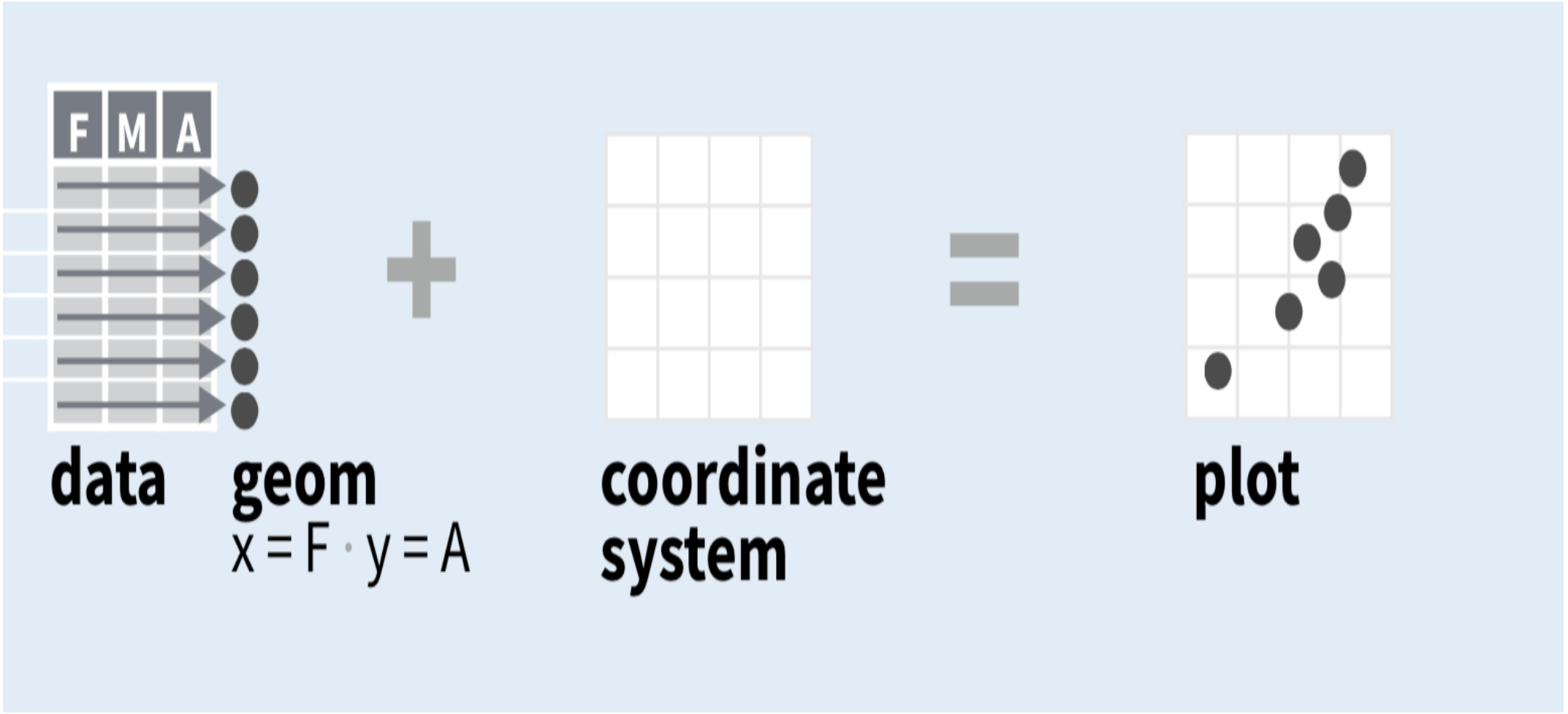
ggplot2 library

(grammer of graphics)
ggplot() function



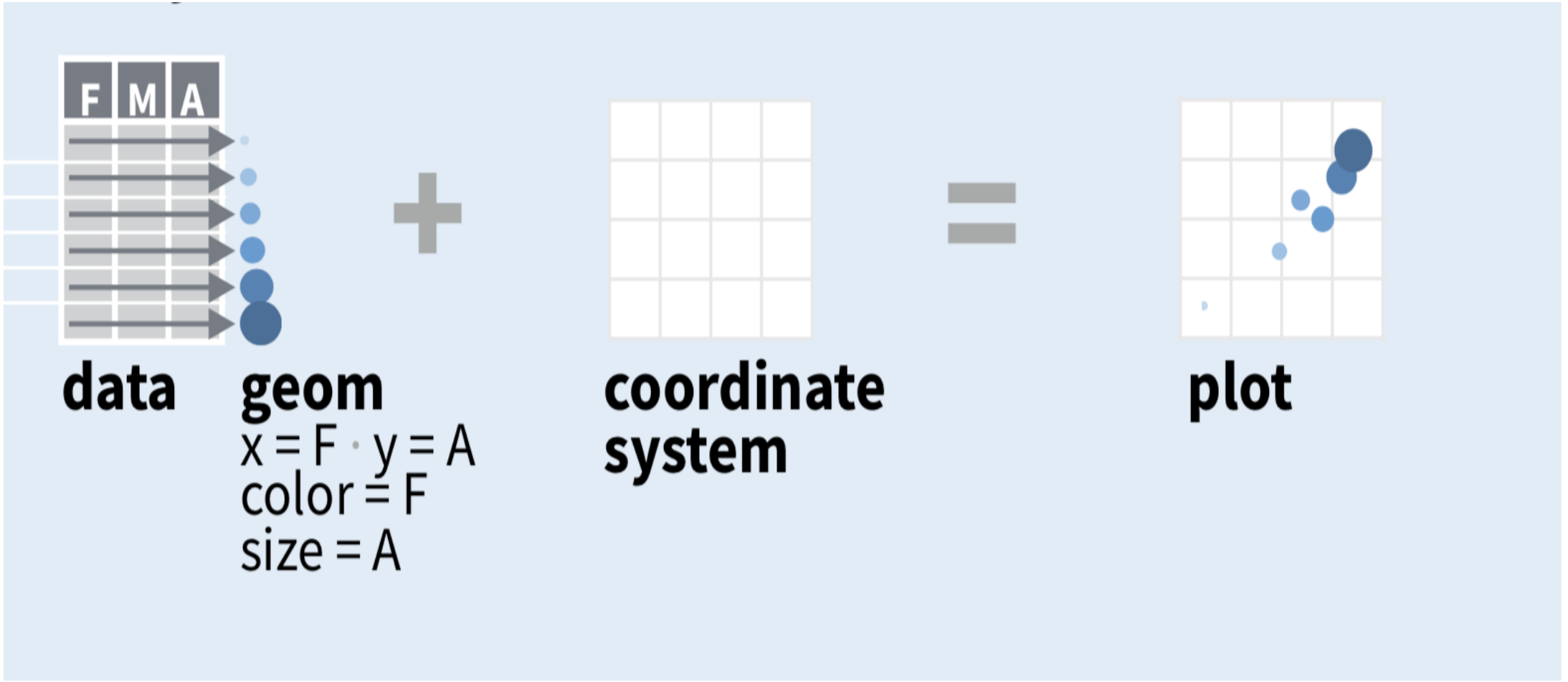
* the ability to be incremented line by line, adjusting parameters by parameters slowly *

How does ggplot2 work?



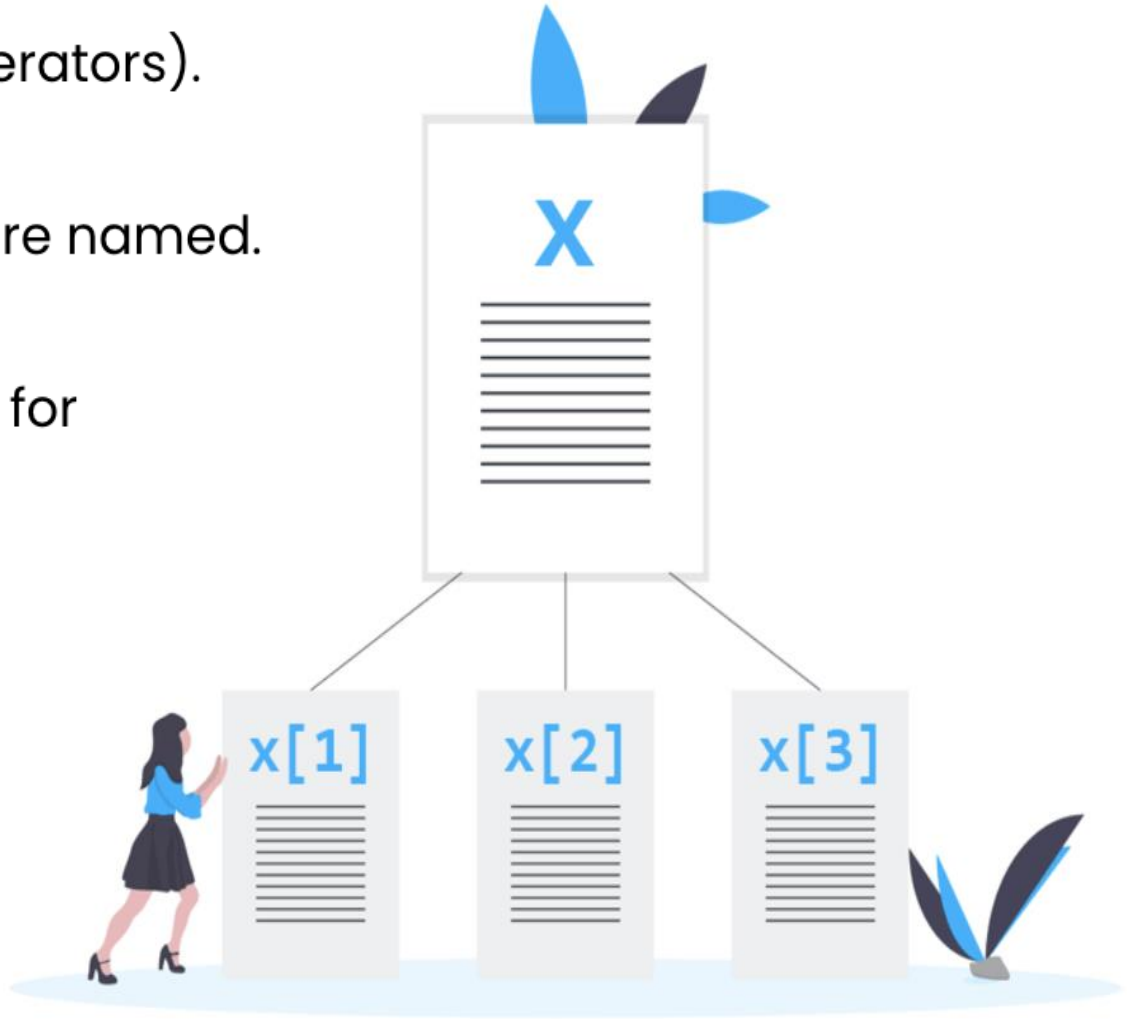
Idea: building graphics up from a set of primitives (i.e verbs,nouns)

Components of a plot



Subset data in R

1. **Using square brackets** (`[]` and `[[]]` operators).
2. Using the **dollar sign** (`$`) if the elements are named.
3. **With functions**, like the `subset` command for conditional or logical subsets.



Subset vector in R

1. Selecting the **indices** you want to display. If more than one, select them using the `c` function.
2. Using **boolean indices** to indicate if a value must be selected (`TRUE`) or not (`FALSE`).
3. Using **logical operators** with the `subset` function.
4. If you want to select all the values except one or some, make a subset indicating the index with negative sign.

Indexing System

How to index?	Based on what?
[]	position
-	exclude
c(…..) A vector of index	multiple elements
logical vector	condition
names	names

Differential Expression Analysis

✓ **Aim:**

Finding Differentially Expressed Genes(DEGs) in:

Cancer vs. Normal

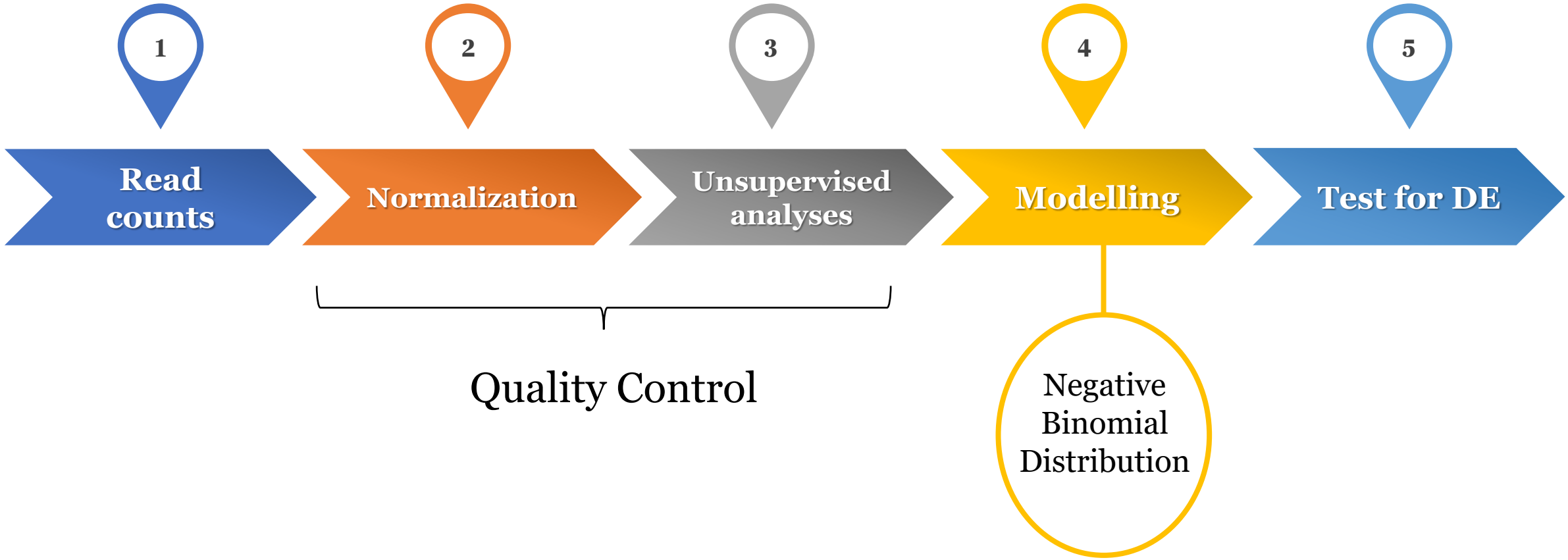
Treated vs. Control

Cell line A vs. Cell line B

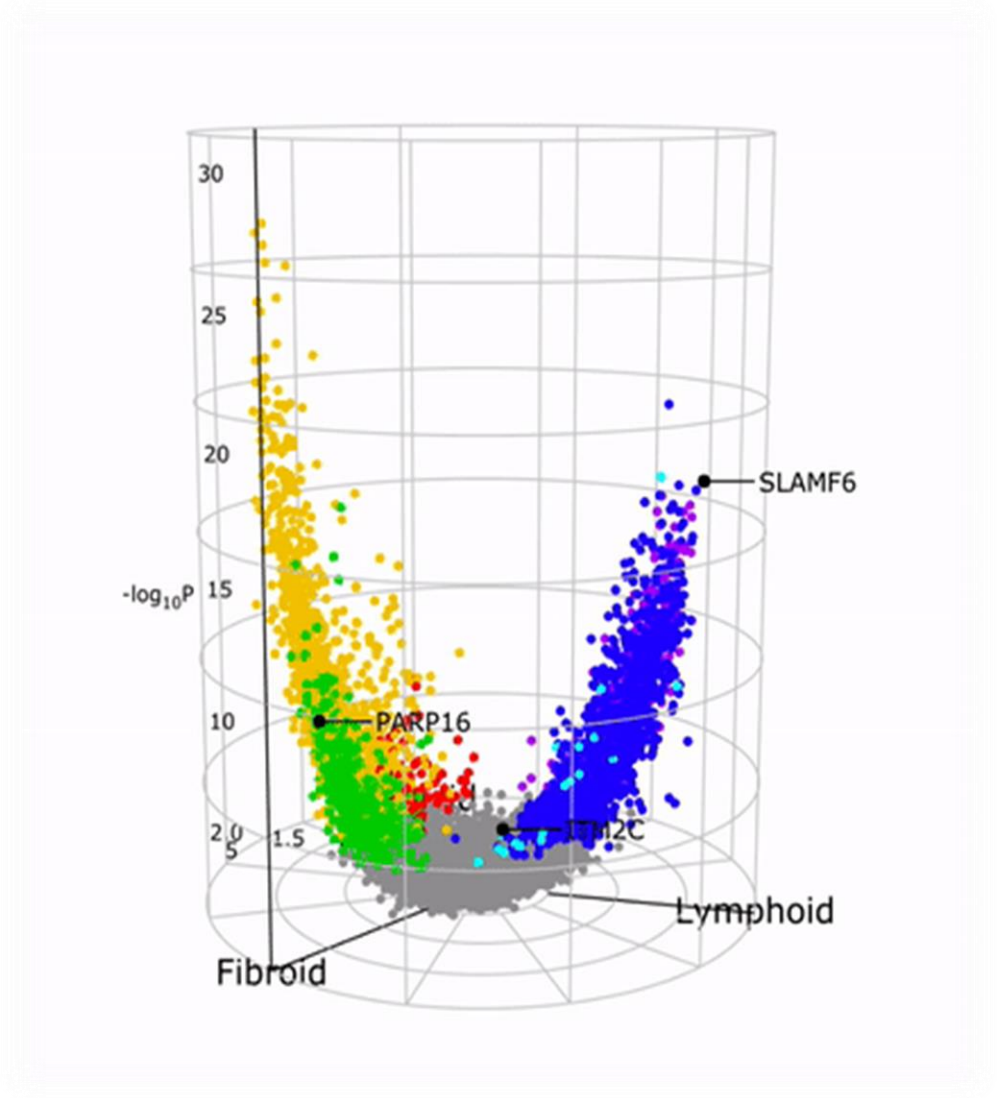
✓ **R packages:**

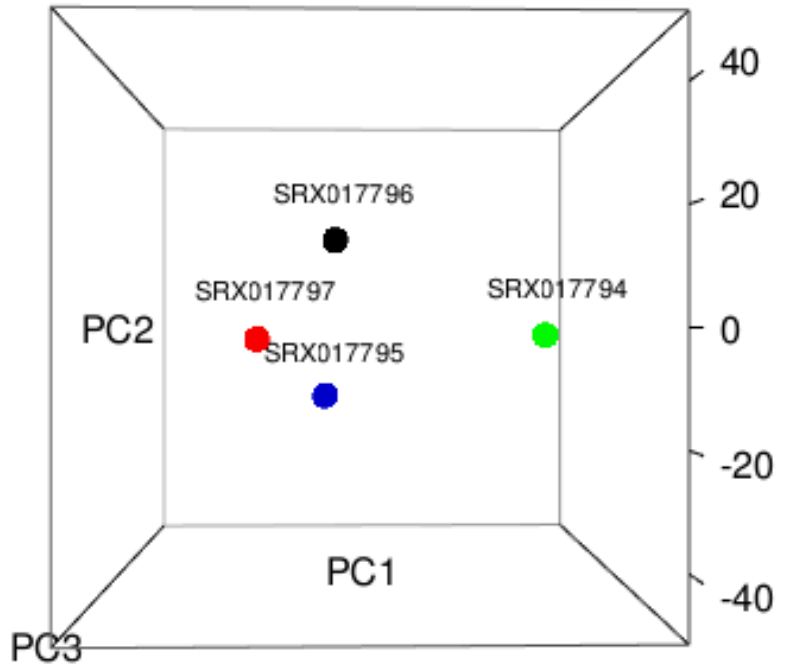
DESeq, DESeq2, edgeR, limma, …

DESeq2 steps

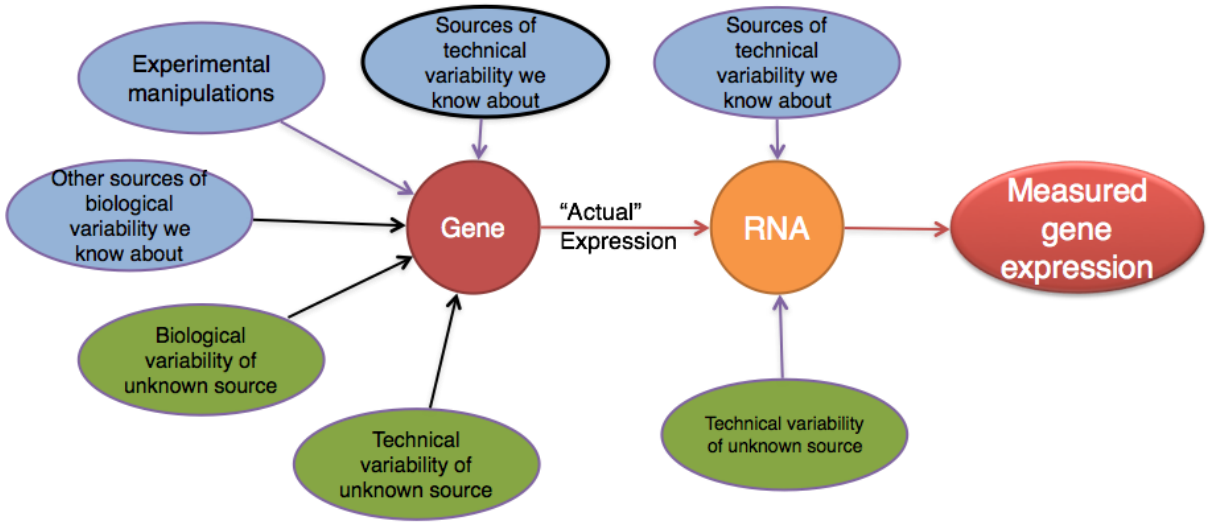


Differential Expression Analysis

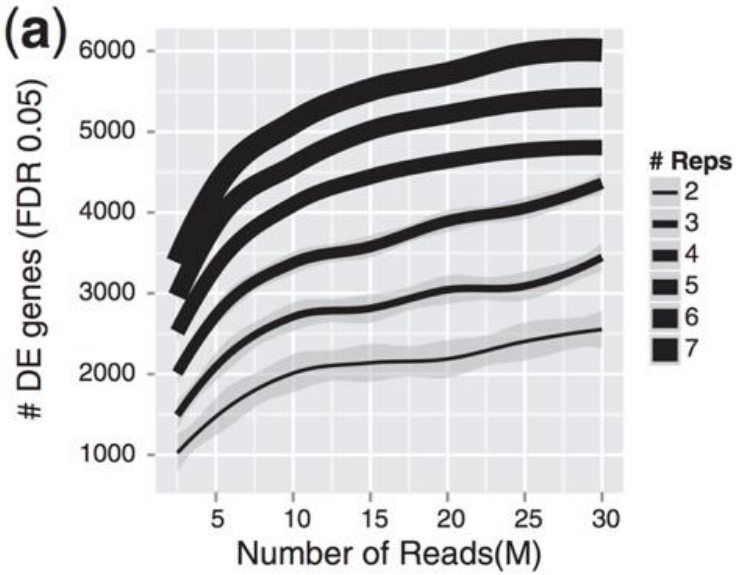




Read counts



Courtesy of Paul Pavlidis, UBC



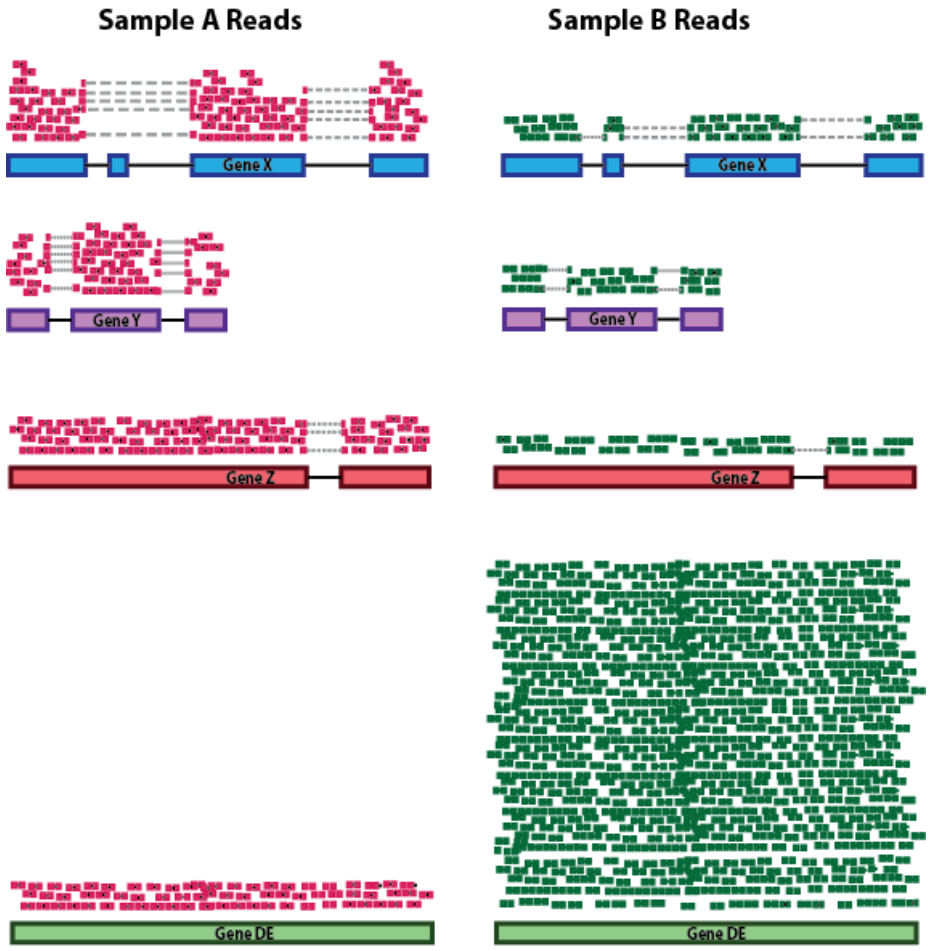
samples: want to see if differences across condition are significant
(w.r.t. biological and technical variation)

features (e.g. genes)

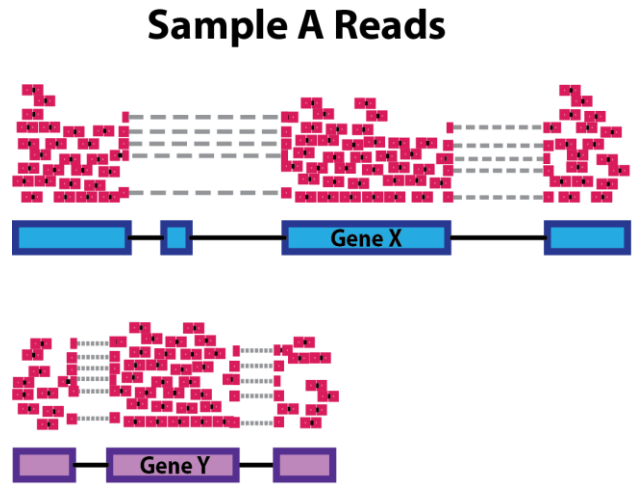
	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

Normalization

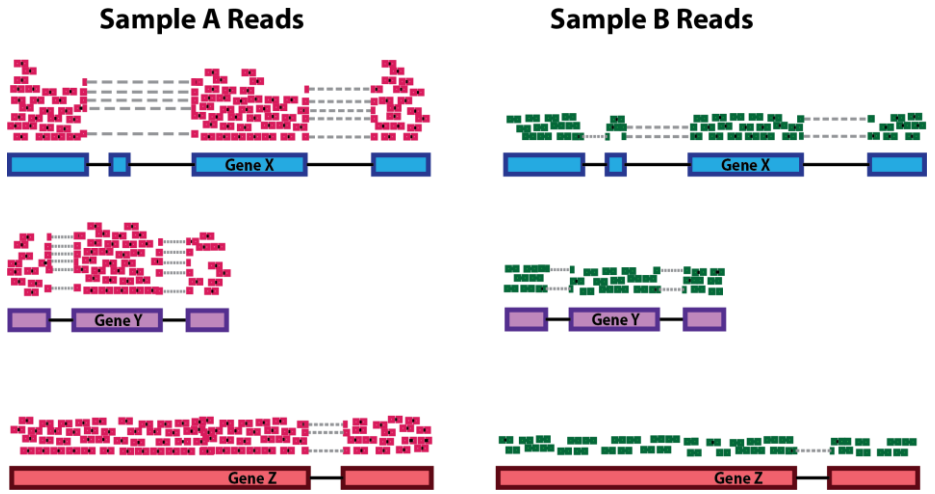
RNA composition



Gene length



Sequencing depth



Unsupervised analyses

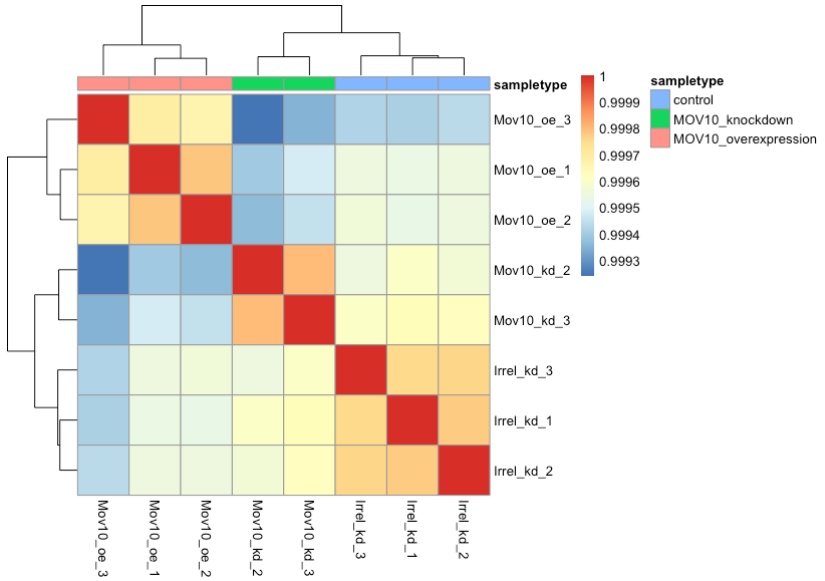
Principal Component Analysis (PCA) and hierarchical clustering

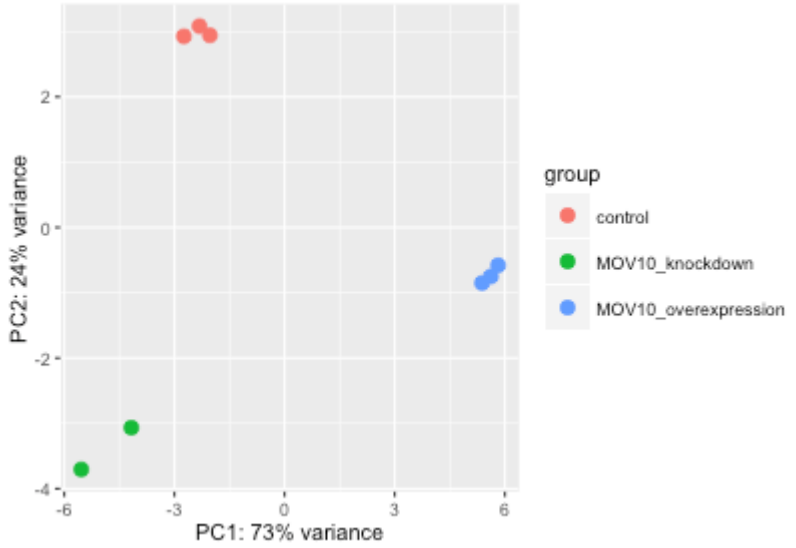
Read counts associated with genes

Normalization

Unsupervised clustering analyses

Quality control





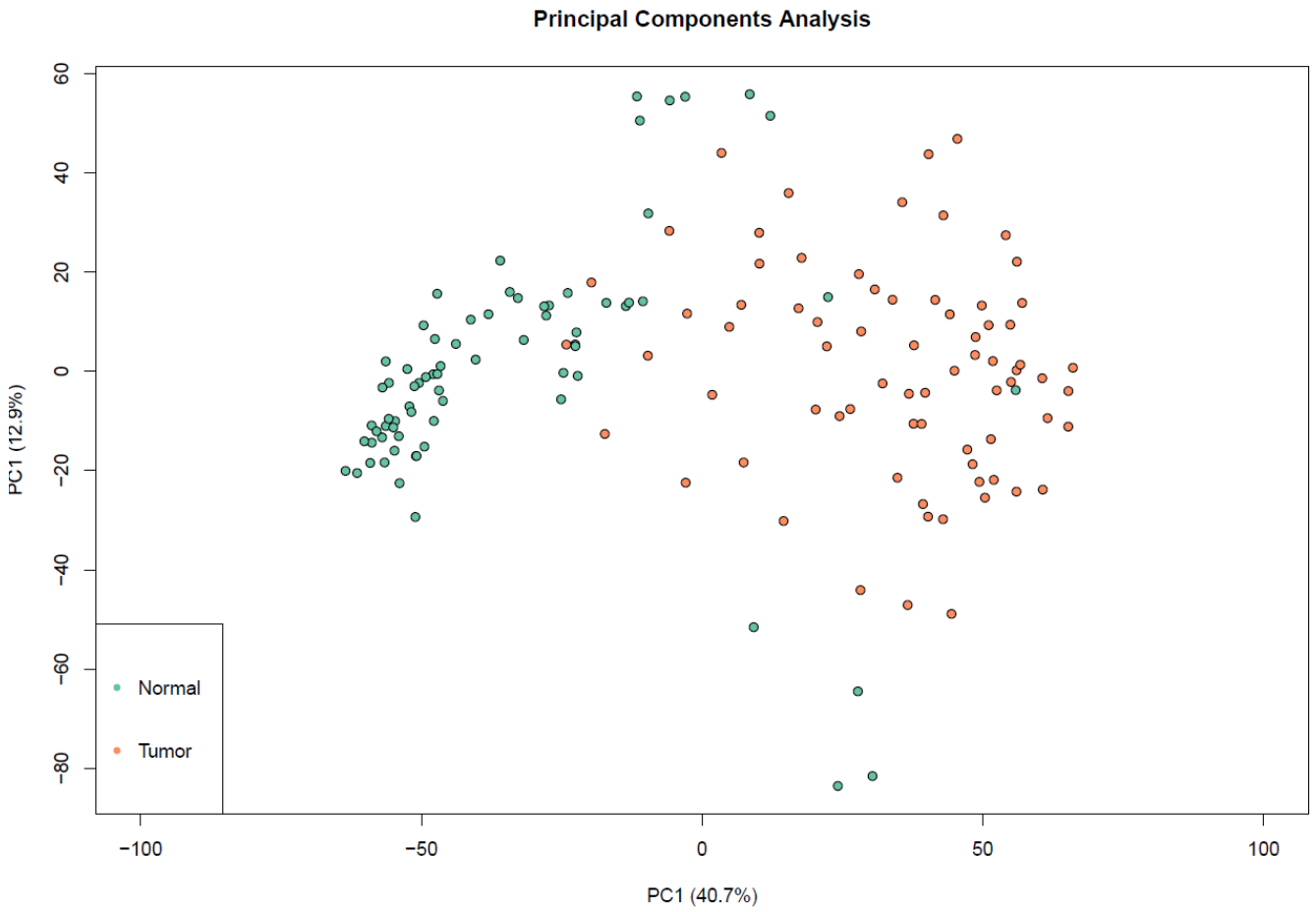
Find the outliers

Read counts associated with genes

Normalization

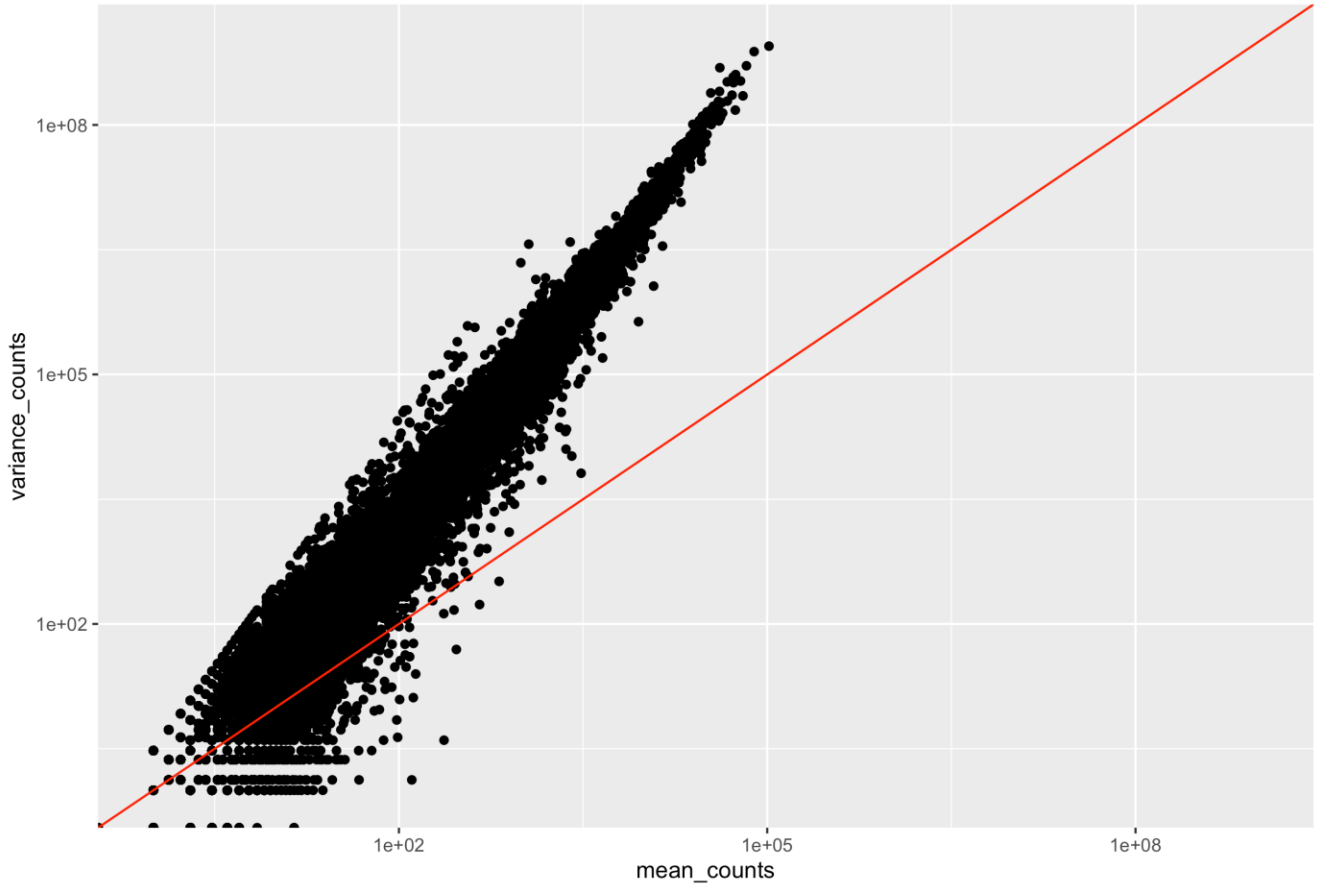
Unsupervised clustering analyses

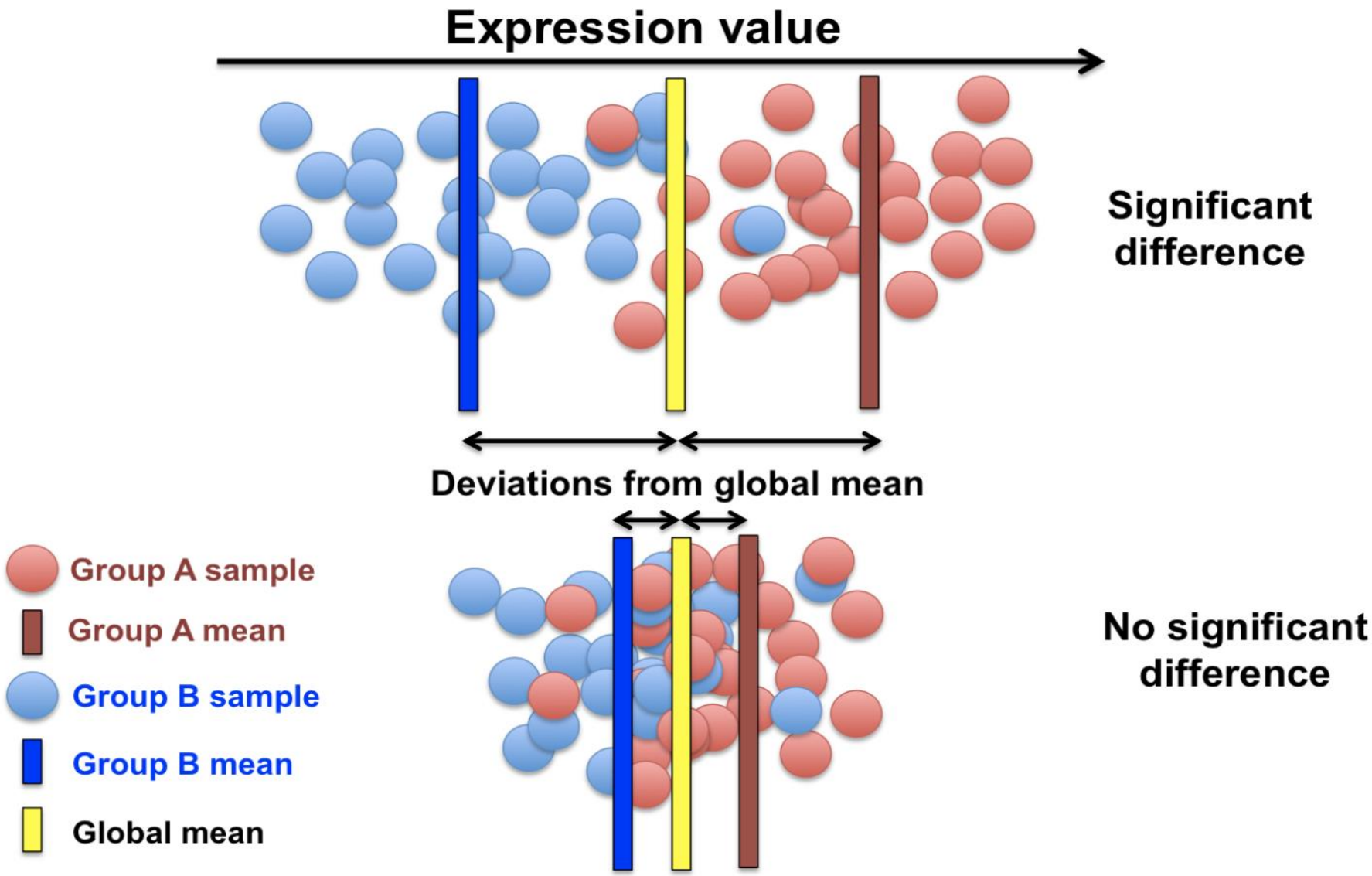
Quality control



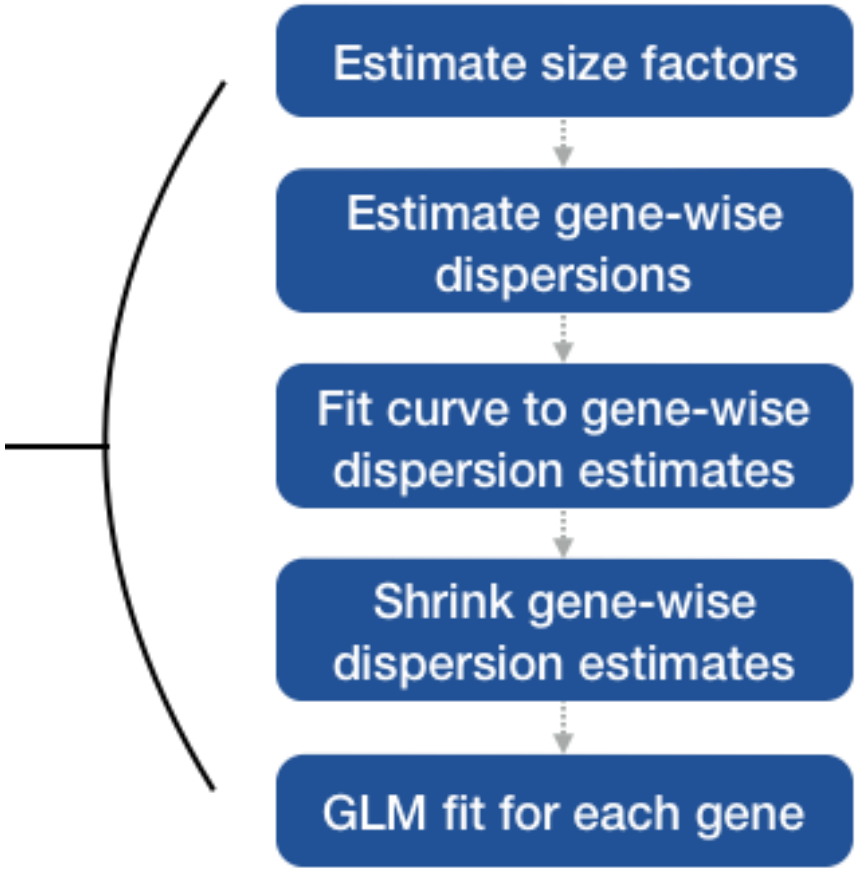
Poisson distribution Vs. Binomial distribution

So what do we use for RNA-seq count data?




Expression value
Significant difference
Deviations from global mean
Group A sample
Group A mean
Group B sample
Group B mean
Global mean
No significant difference

Test for DE



Results()

- `baseMean`
- `log2FoldChange`
- `lfcSE`
- `Stat`
- `Pvalue`
- `Padj`

baseMean

mean of normalized counts for all samples

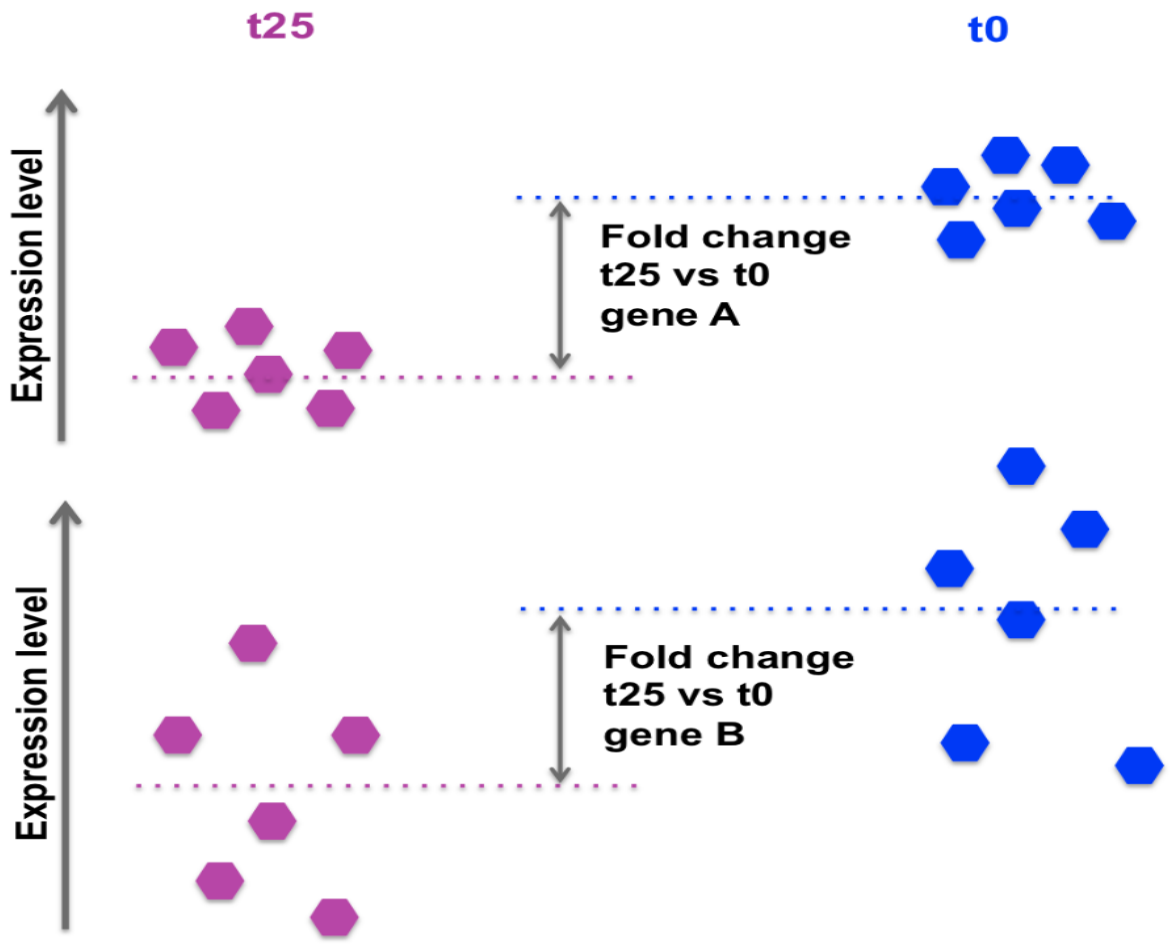
log2 foldchanges

log2 (normalized_counts_interested/normalized_counts_control)

- Fold Change > 0 for gene A means that gene A is **more expressed (= up-regulated)** in interested condition compared to control.

- Fold Change < 0 for gene A means that gene A is **less expressed (= down-regulated)** in interested condition compared to control.

Why the selection shouldn’t be based on fold changes only?



The fold change for gene A and gene B is **the same**!!!

P-value

Results

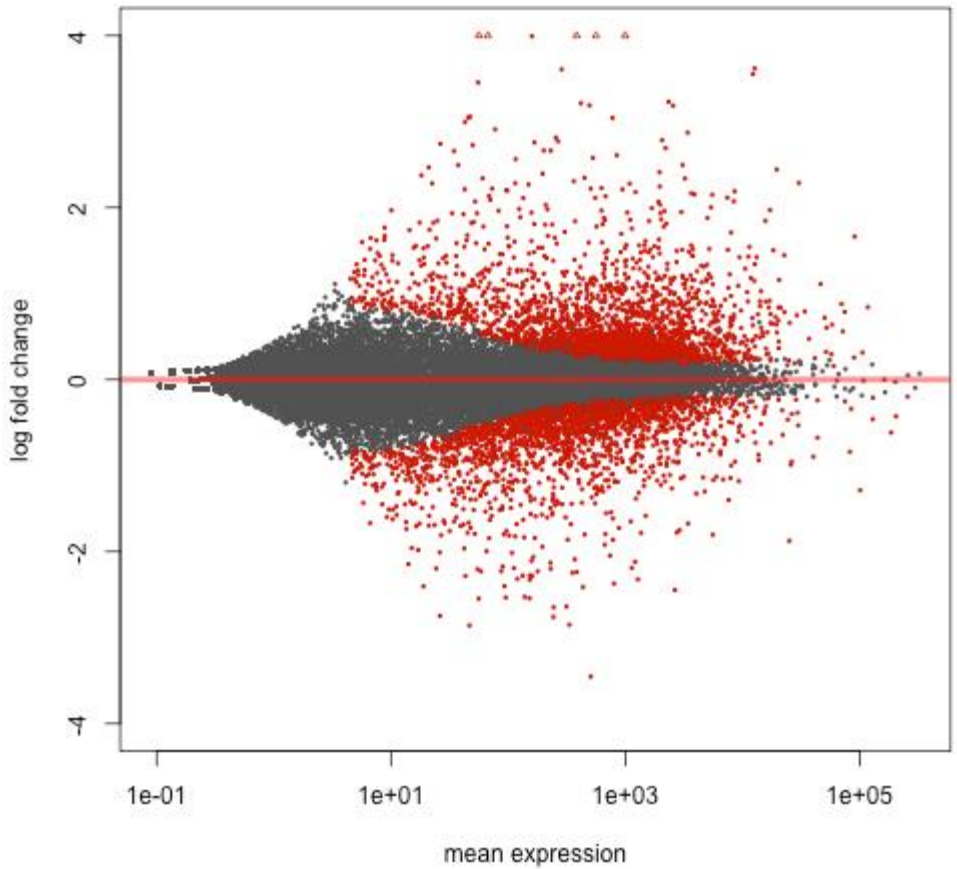
- Gene-level filtered Table

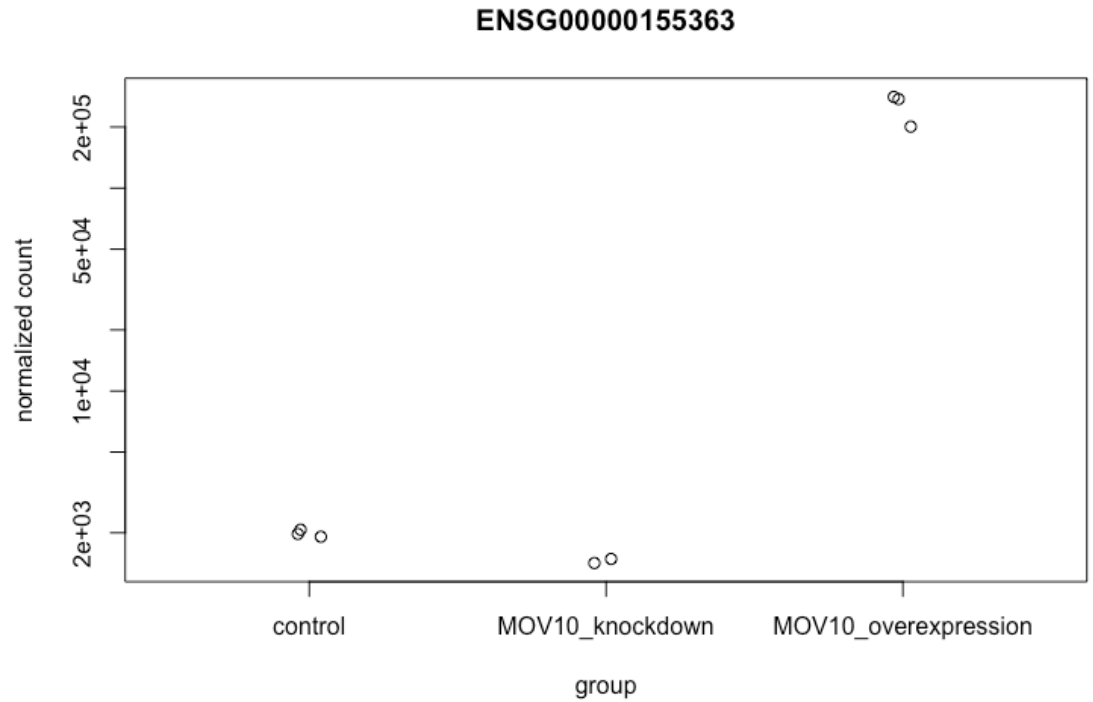
- Plot counts

plot expression of a single gene

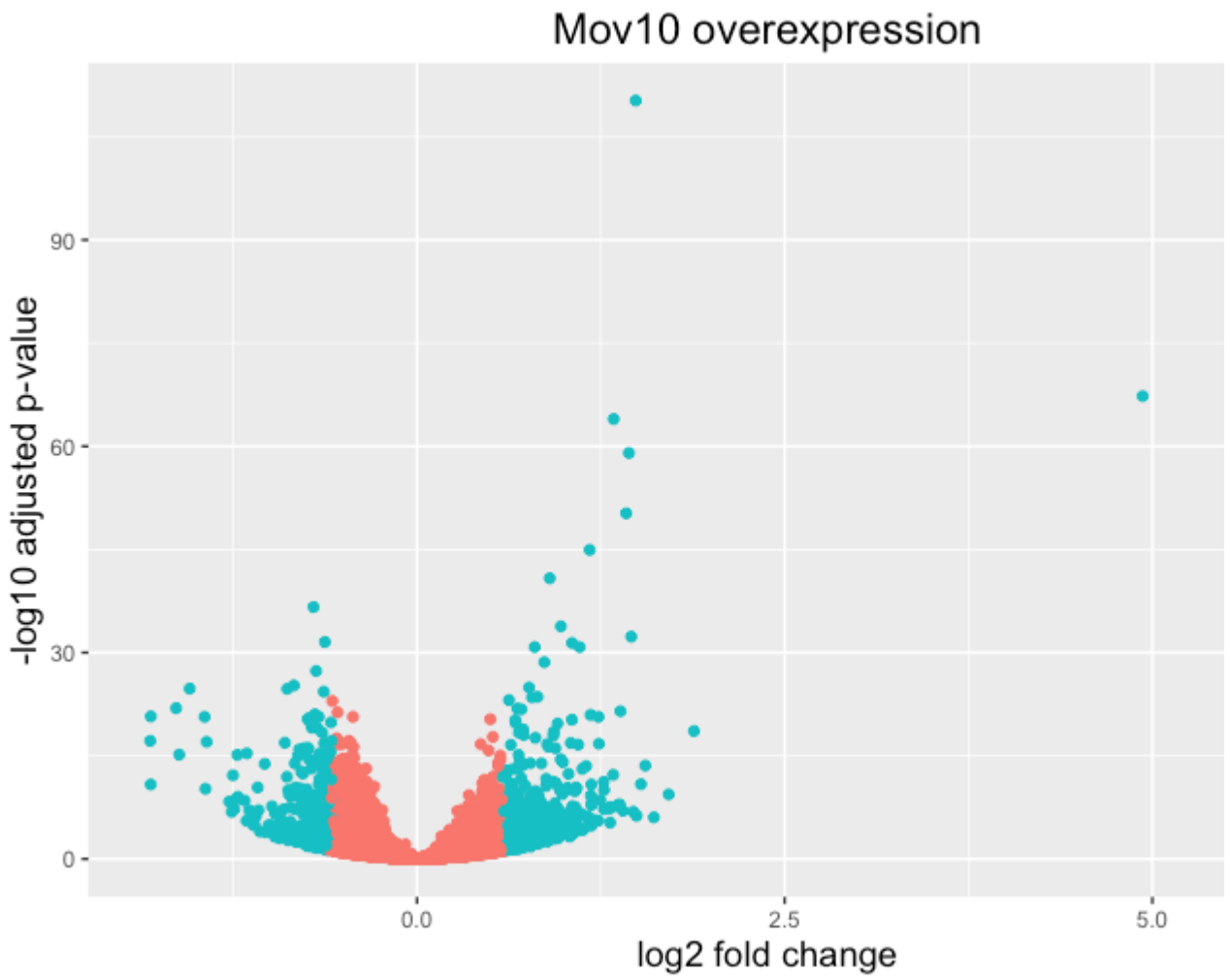
- MA plot

mean of the normalized counts versus the log2 foldchanges for all genes tested


log fold change
4
2
0
-2
-4
1e-01
1e+01
1e+03
1e+05
mean expression


ENSG00000155363
normalized count
2e+05
5e+04
1e+04
2e+03
control
MOV10_knockdown
MOV10_overexpression
group

- Volcano


Mov10 overexpression
-log10 adjusted p-value
90
60
30
0
0.0
2.5
5.0
log2 fold change

- Heatmap