Université Paris Sud – Paris Saclay

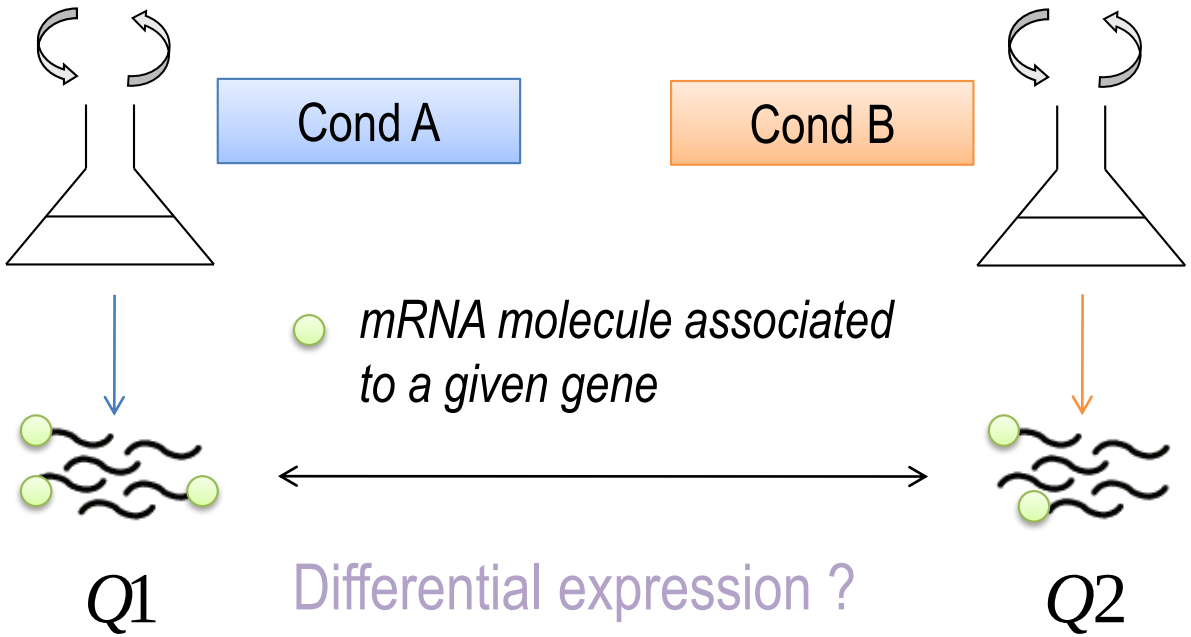
2018 - 2019

Differential expression analyses:
How to (better) understand your results ?

Gaëlle LELANDAIS

gaelle.lelandais@u-psud.fr

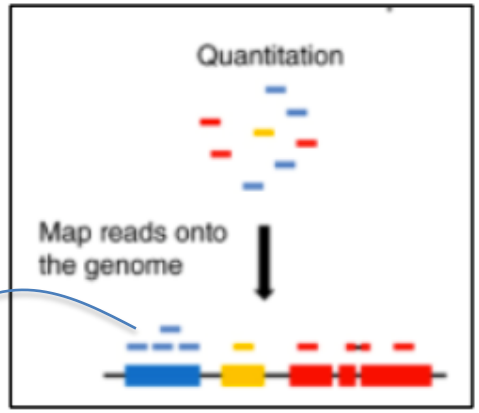
What is the question ?



➢ All genes are successively analyzed in a differential analysis

Experimental dataset

(number of mapped reads)



	Cond A			Cond B		
	R1	R2	R3	R1	R2	R3
A1BG	4	6	2	7	6	7
A1CF	41	33	42	32	42	32
A2M	1	3	1	4	3	7
A2ML1	3	2	2	6	7	3
A2MP1	3	2	2	1	1	0
A3GALT2	1	4	4	3	2	1
A4GALT	420	344	291	327	360	371
A4GNT	1	1	2	1	3	3
AA06	0	0	0	0	0	0
AAAS	2452	2192	1977	2054	2134	2100
AACS	3234	2804	2609	1678	1670	1742
AACSP1	1544	1369	1300	1926	2015	1963

Source of dataset: Yang et al., Mol Cell Biol (2016)

Data normalisation : Why ?

Number of mapped reads is related to gene length

A1BG	4
A1CF	41
A2M	1
A2ML1	3
A2MP1	3
A3GALT2	1
A4GALT	420
A4GNT	1
AA06	0
AAAS	2452
AACS	3234
AACSP1	1544

Gene A

8 mapped reads

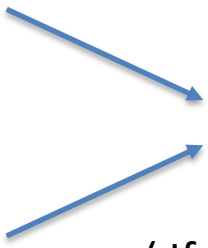
Gene B

4 mapped reads

Data normalisation : Why ?

Number of mapped reads is related to library size

A1BG	4	7
A1CF	41	32
A2M	1	4
A2ML1	3	6
A2MP1	3	1
A3GALT2	1	3
A4GALT	420	327
A4GNT	1	1
AA06	0	0
AAAS	2452	2054
AACS	3234	1678
AACSP1	1544	1926
	Sum 1	Sum 2



Systematic bias ?

(If Sum 1 = 40.000 and Sum 2 = 30.000)

Section 1

GENERAL PRINCIPLE

Working hypothesis for normalization

Most of the genes are not differentially expressed

$\text{LogFC} \approx 0$

LogFC (Fold Change) parameter

- Average (mean) value of log2(C_A/C_B) in 3 replicates

$$\log FC_g = \frac{1}{3}\left(\log 2\left(\frac{Q1}{Q2}\right)_{g,R1} + \log 2\left(\frac{Q1}{Q2}\right)_{g,R2} + \log 2\left(\frac{Q1}{Q2}\right)_{g,R3} \right)$$

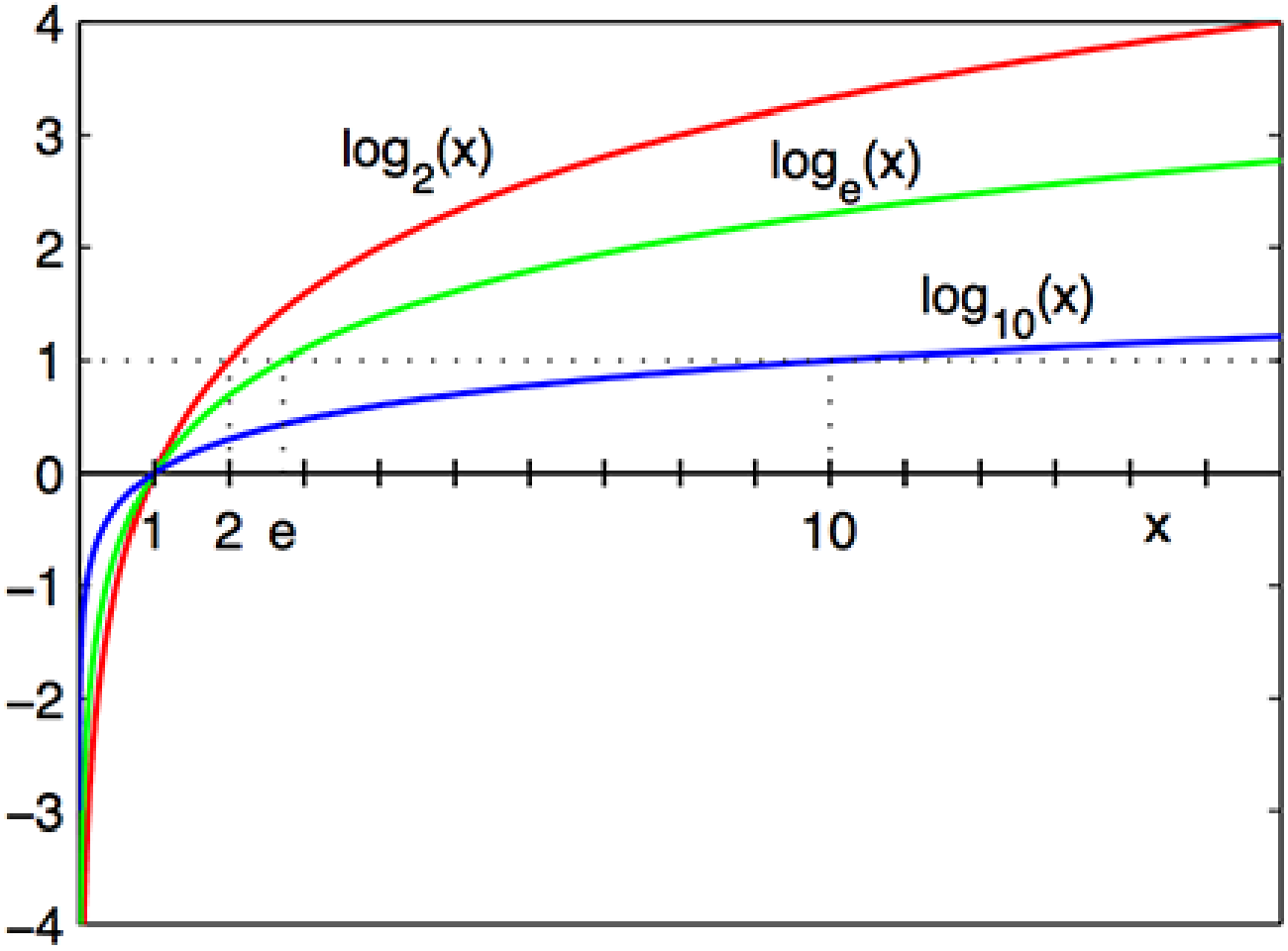
$$\log FC_g > 0 \Leftrightarrow Q1_g > Q2_g$$

↳ Up-regulated gene

$$\log FC_g < 0 \Leftrightarrow Q1_g < Q2_g$$

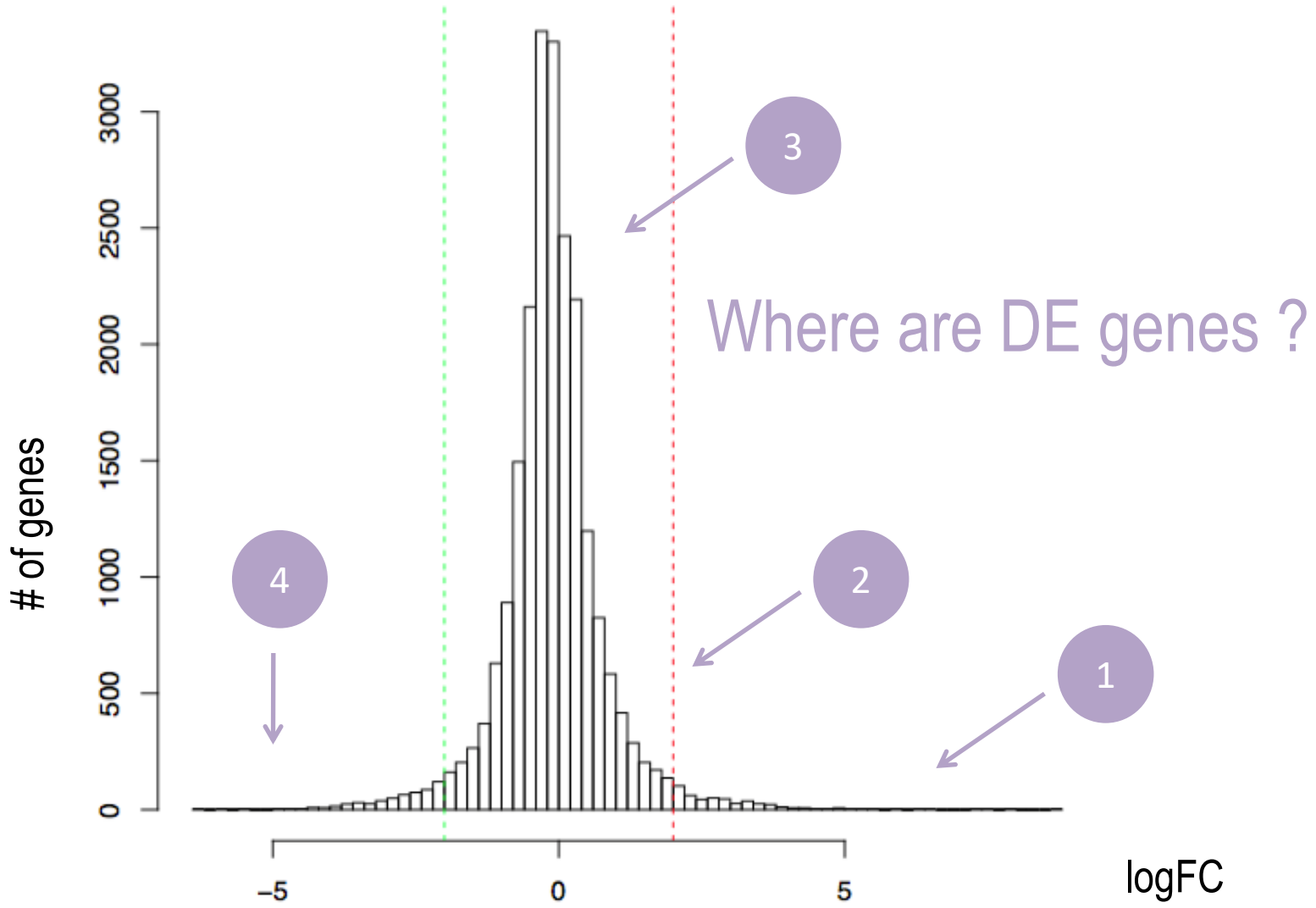
↳ Down-regulated gene

Why using logFC instead of FC ?



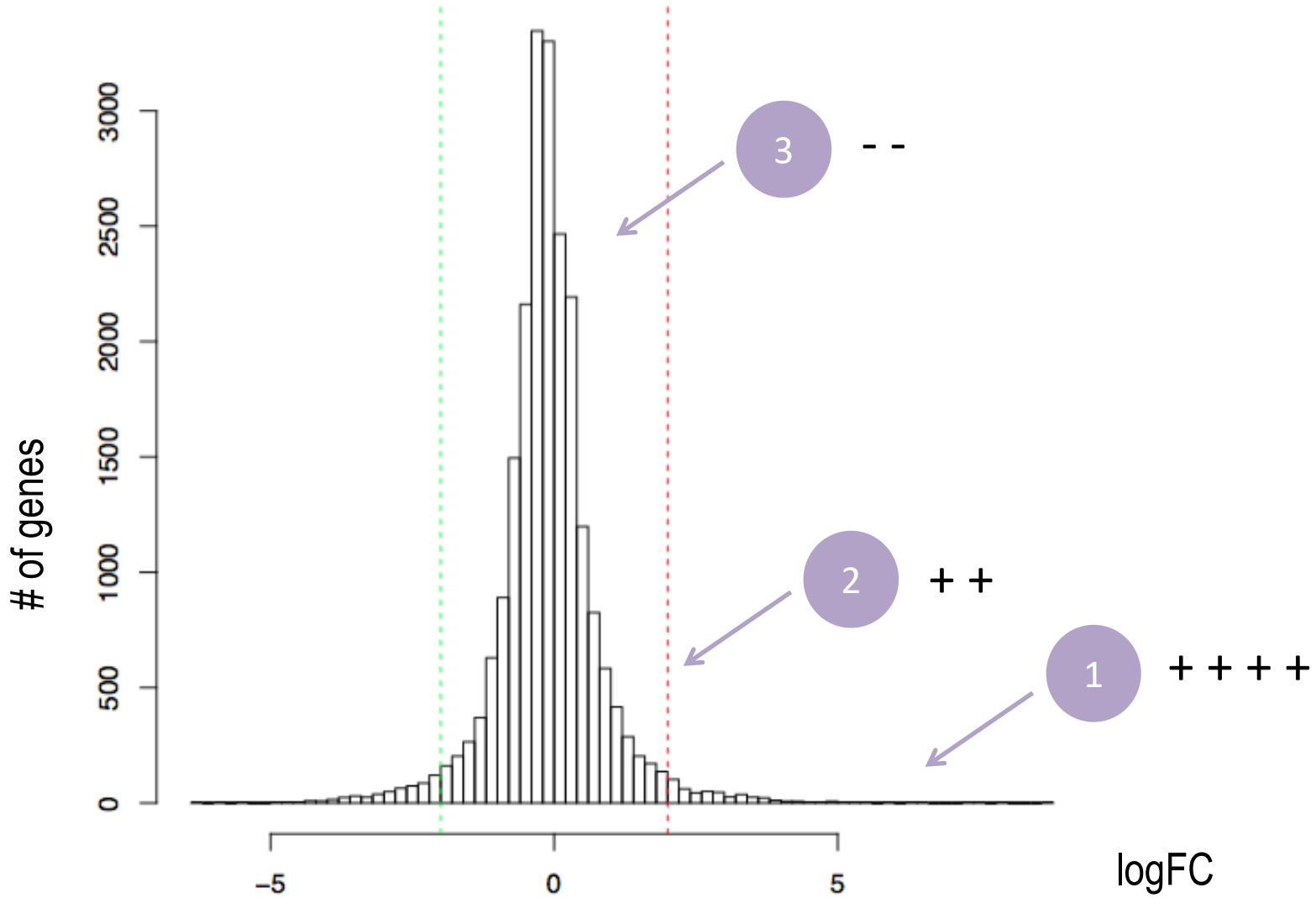
https://fr.wikipedia.org/wiki/Logarithme

Distribution of logFC values



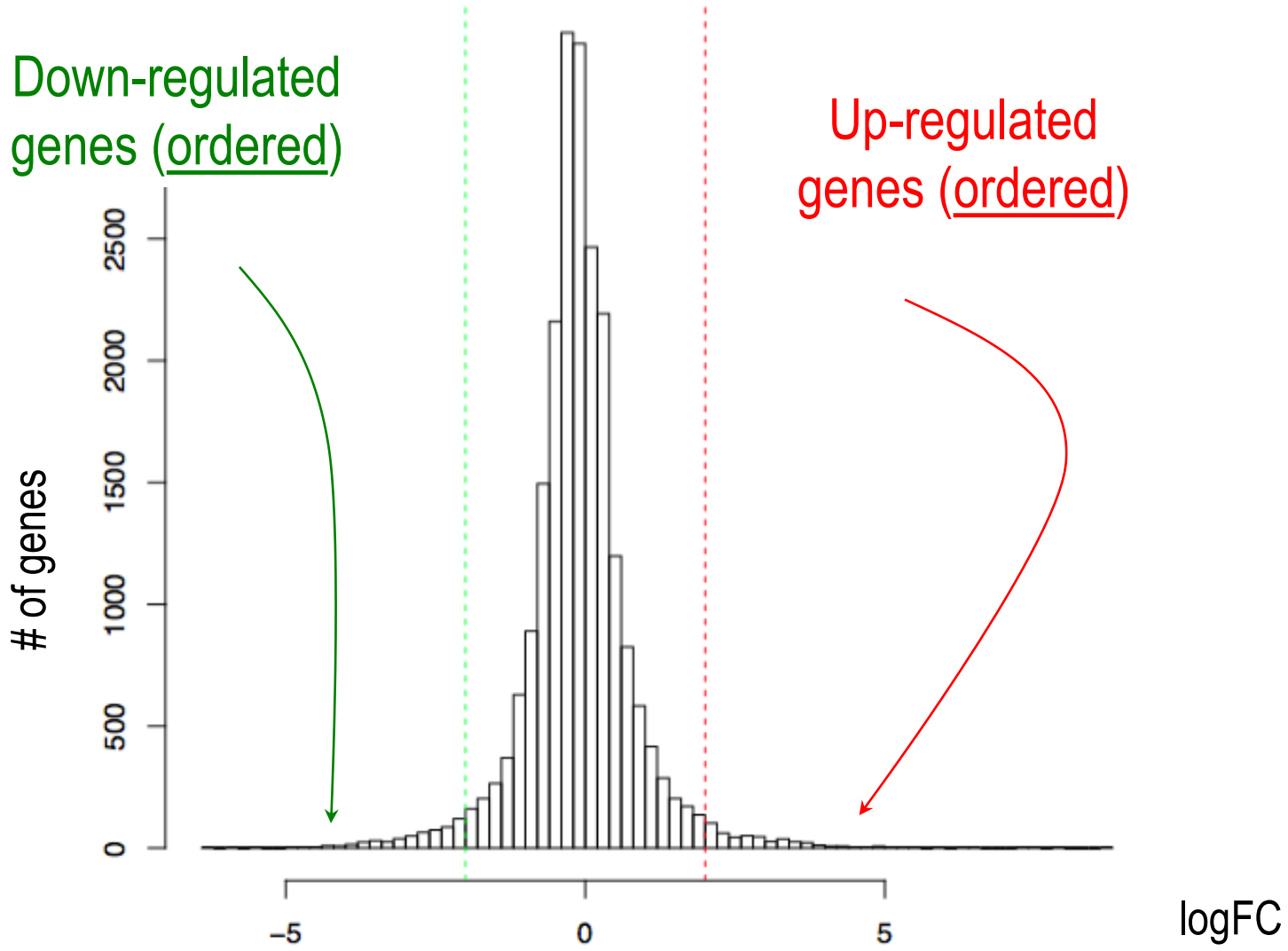
Source of dataset: Yang et al., Mol Cell Biol (2016)

It's all about "confidence level"



Source of dataset: Yang et al., Mol Cell Biol (2016)

DE analysis based on logFC only



Source of dataset: Yang et al., Mol Cell Biol (2016)

In addition to logFC …

- Two genes can have identical logFC, but different individual $\log_2(C_A/C_B)$ values

++ Gene 1: $\log FC_{G1} = \frac{1}{3}(0.5 + 2.5 + 6) = 3$

++++ Gene 2: $\log FC_{G2} = \frac{1}{3}(3 + 2.8 + 3.2) = 3$

Data variability

Same level of confidence ?

Quantification of the “data variability”

- Variance estimation

$$Var(X) = \frac{1}{n-1} \sum (x_i - m)^2$$

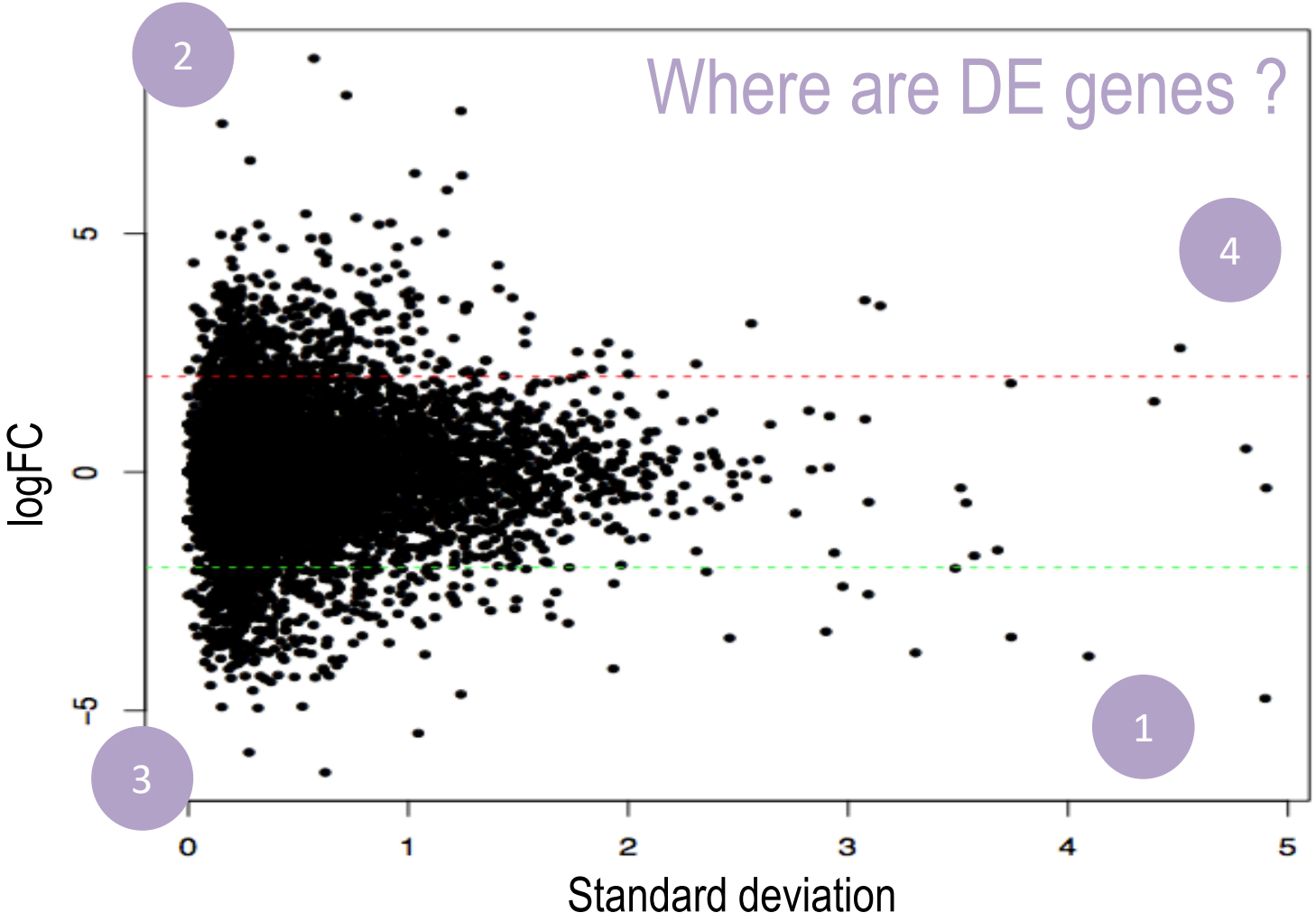
- Standard deviation

$$SD = \sqrt{Var(X)}$$

Application to logFC :

	logFC	Variance	SD
Gene 1	3	7.75	2.78
Gene 2	3	0.04	0.20

DE analysis based on logFC and SD



Source of dataset: Yang et al., Mol Cell Biol (2016)

Solution to combine differential expression (logFC) and data variability (SD)

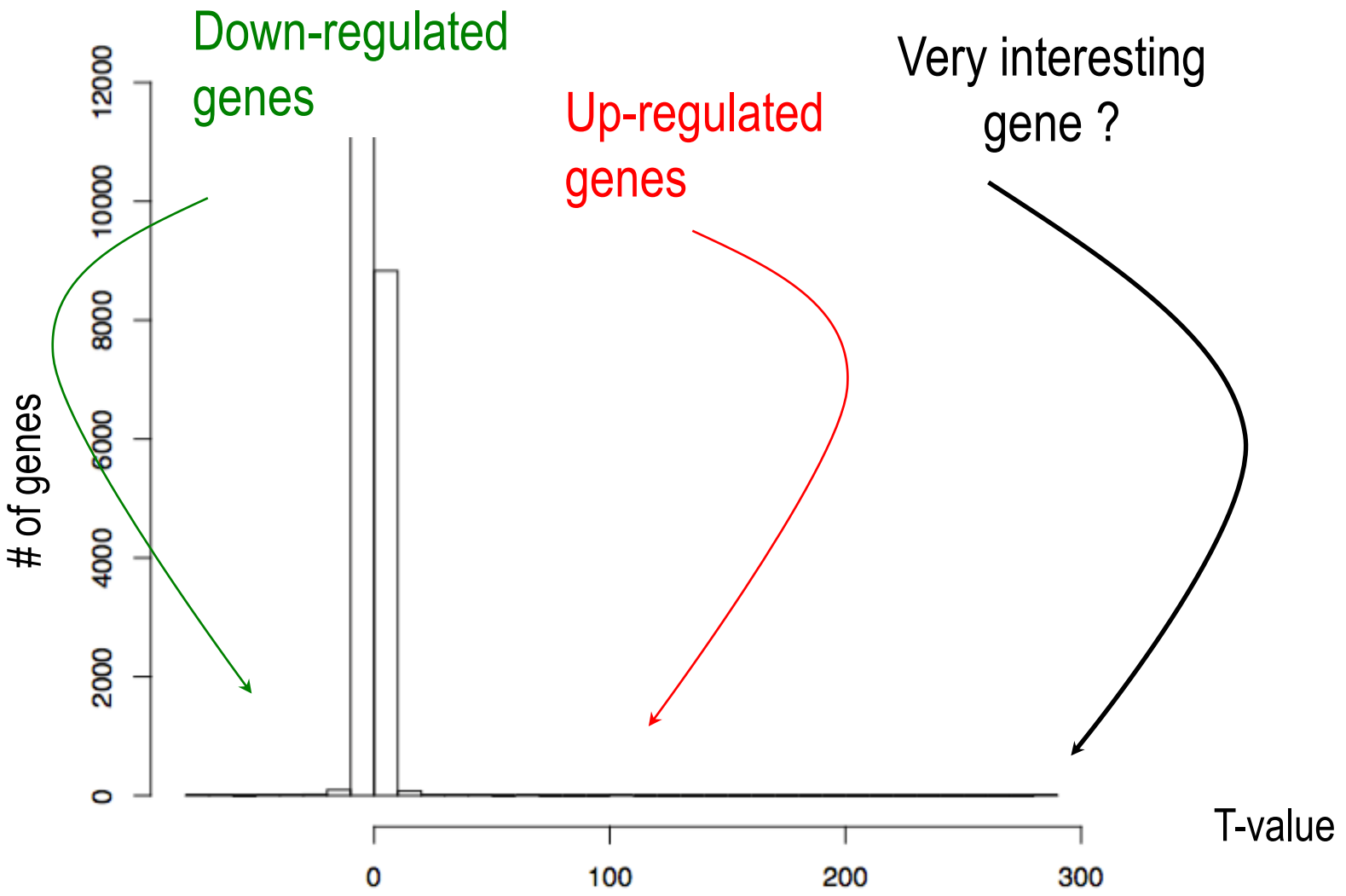
- Student parameter t (used in classical statistics)

$$t_g = \frac{\log FC_g}{\frac{SD_g}{\sqrt{n}}}$$

How to obtain high/low values of “t” ?

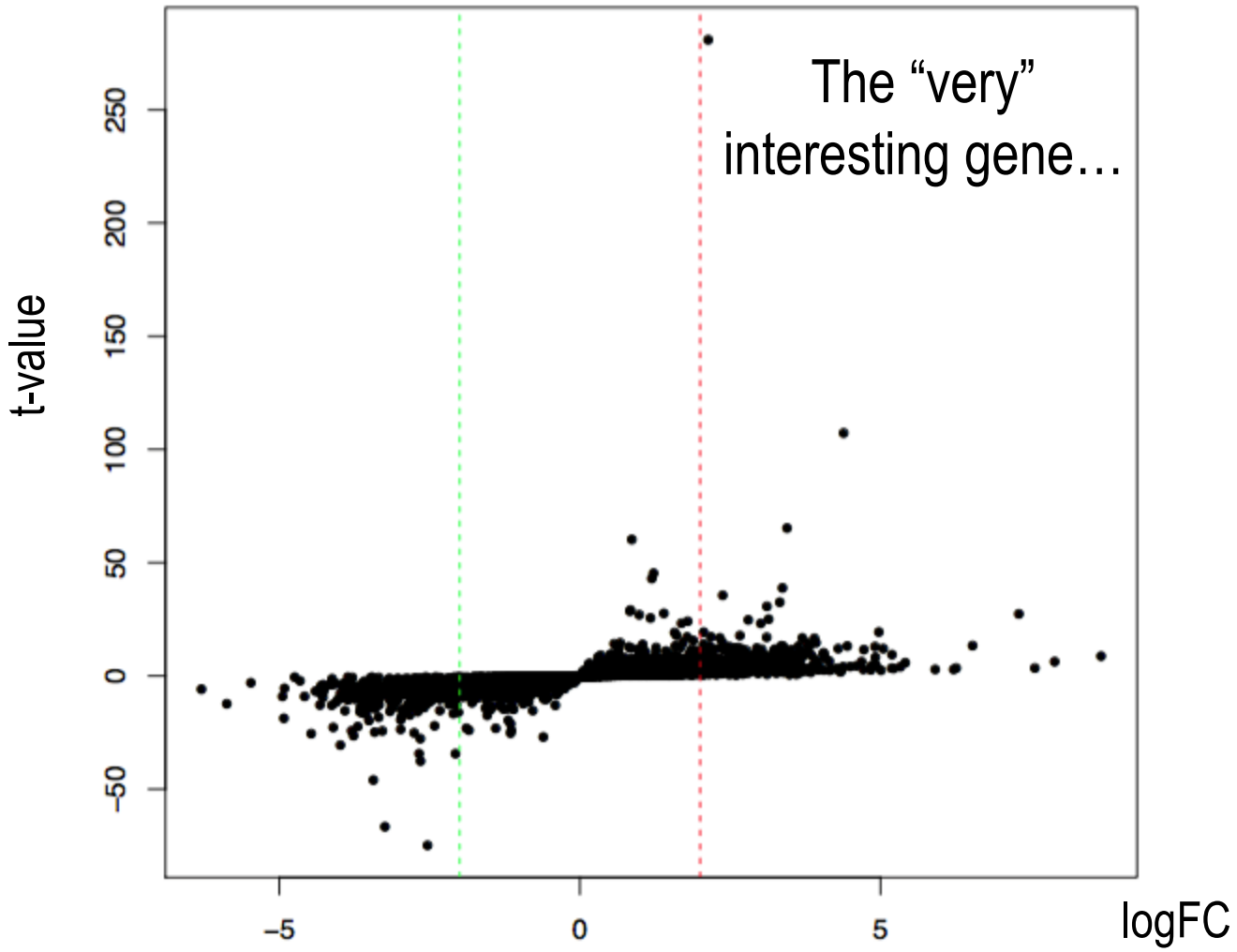
Does it work correctly ?

Distribution of t-values (logFC)



Source of dataset: Yang et al., Mol Cell Biol (2016)

Drawback associated to the classical t-value



Source of dataset: Yang et al., Mol Cell Biol (2016)

Also, p-value can't be calculated because …

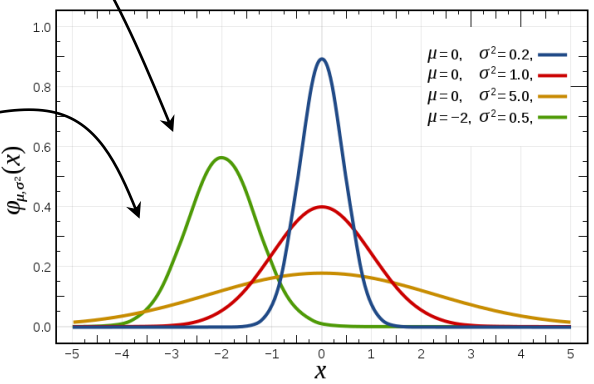
- The random variable must be normally distributed :

Gene 1: $\log FC_{G1} = \frac{1}{3}(0.5 + 2.5 + 6) = 3$

Gene 2: $\log FC_{G2} = \frac{1}{3}(3 + 2.8 + 3.2) = 3$

?

?



This is impossible to test, because of the limited number of replicates

R script to be used/modified

Search_DEgenes.R

search_DEgenes_M2Bioinfo.R

search_DEgenes_M2Bioinfo.R 〉 No Selection

```
#####
# Gaelle LELANDAIS <gaelle.lelandais@u-psud.fr>
# Melina GALLOPIN <melina.gallopin@u-psud.fr>
#####

####
# Ce script R est destine aux etudiants de Master. Il a pour objectif de les aider dans la
# comprehension des methodes de selection des genes differentiellement exprimes.
####

####
# Librairies R necessaires : LIMMA, MASS, DESeq2, edgeR
####

#-------------------------------------------------------------------------------------------
#-------------------------------------------------------------------------------------------

pdf("graphics.pdf")

# Lecture des données
countData = read.table("count_dataFile.txt", header = T, row.names = 1)
# --> Ces donnees sont en relation avec l'article :
#
# "Determination of a Comprehensive Alternative Splicing Regulatory Network and
# Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition"
#
# Yang et al. (2016, Molecular and Cellular Biology)
#
# --> Deux points de temps ont ete etudies par RNAseq : "Day0" et "Day7", 3 replicats sont
# disponibles pour chaque point de temps.

#-------------------------------------------------------------------------------------------
# Pretraitement des donnees entre les experiences (filtrage des faibles et normalisation)
#-------------------------------------------------------------------------------------------

## Suppression des genes sans valeurs (0 pour toutes les experiences)
length(which(rowSums(countData)==0))  ## 2957
countData <- countData[-which(rowSums(countData) == 0),]

# Distribution des valeurs de comptages dans les différents échantillons
boxplot(log(countData + 1), ylab = "#reads (log scale)",
        main = "Boxplot of read counts", col = c(rep("grey", 3), rep("blue", 3)),
        names = c(paste("Day0_", 1:3, sep = ""), paste("Day0_", 1:3, sep = "")))
# --> Les distributions des donnees de comptage sont tres proches entre les differentes
# conditions.
```

SECTION 2

DEDICATED METHODS TO IDENTIFY DIFFERENTIALLY EXPRESSED GENES

R package DESeq2

- Read counts

DEseq2 →

- Statistical parameters

RNAseq_counts_parapsilosis.txt

	Normoxia_1	Normoxia_2	Normoxia_3	Normoxia_4	Normoxia_5	Normoxia_6
	Hypoxia_1	Hypoxia_2	Hypoxia_3	Hypoxia_4		
CPAR2_404750	589	498	630	612	627	642	462	611	319	619
CPAR2_212570	559	648	576	689	988	716	438	536	258	493
CPAR2_302130	167	83	230	151	180	136	112	385	103	172
CPAR2_209760	869	1438	972	1250	1588	1329	555	879	313	1206
CPAR2_403260	2549	2313	3115	2801	5834	3529	3373	3290	851	2873
CPAR2_209040	533	313	646	532	335	396	262	879	266	743
CPAR2_804040	44	42	57	44	68	43	41	76	25	71
CPAR2_211500	621	835	736	783	1156	860	529	635	29[illegible]	588
CPAR2_102540	1672	2432	1845	2134	3653	3521	1470	1967	8[illegible]	1970
CPAR2_109950	1772	2769	2175	2424	1629	1441	454	1485	[illegible]12	920
CPAR2_800250	166	144	157	179	181	223	163	84	[illegible]6	179
CPAR2_503900	1023	1435	1320	1414	1894	1664	544	1283	370	1212
CPAR2_100740	238	180	279	224	187	208	67	63	18	114
CPAR2_206160	705	933	831	823	1129	729	430	384	224	557
CPAR2_101420	383	552	448	526	612	388	246	253	100	289
CPAR2_805280	378	515	413	500	850	567	410	525	205	546
CPAR2_202350	1253	2099	1529	1826	2918	2[illegible]				
CPAR2_703020	3668	5607	4518	5199	5891	5[illegible]				
CPAR2_502480	219	291	284	302	233	2[illegible]				
CPAR2_100810	1118	1908	1332	1616	1293	1[illegible]				
CPAR2_105300	5288	7896	6291	7105	11283	9[illegible]				
CPAR2_407740	323	422	422	435	956	628	361	428	189	643
CPAR2_802340	640	848	822	900	1686	1236	488	622	217	543
CPAR2_106320	8875	13180	10310	11985	16549	8906	4887	2543	1266	3225
CPAR2_301210	167	191	218	204	540	451	209	280	124	315
CPAR2_302570	743	1121	856	947	1407	974	632	386	268	716
CPAR2_503570	99	137	135	101	482	213	193	234	86	300

baseMean

DESeq2_statistics.txt

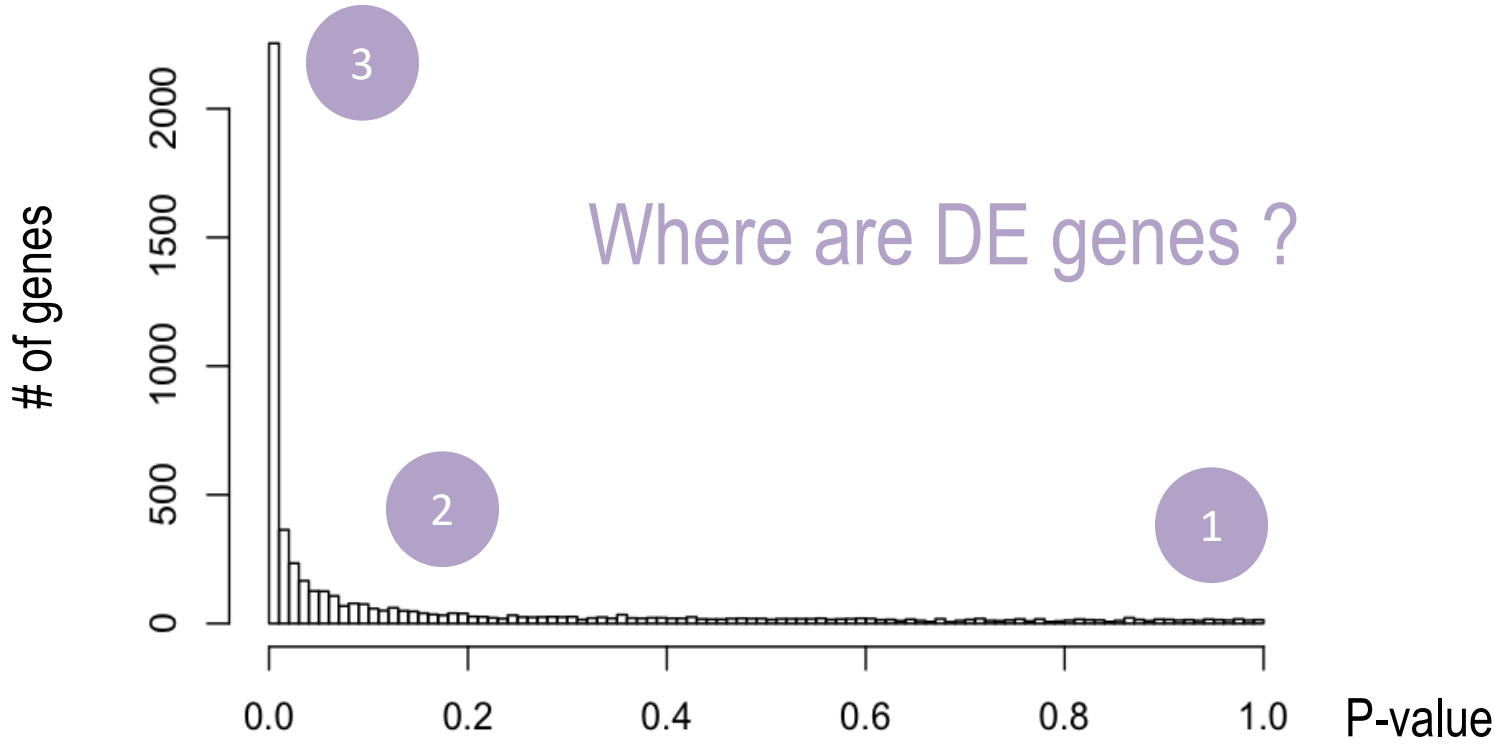
baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
CPAR2_404750	557.1[illegible]6255077326		0.133449464[illegible]5645		0.190487408943552
0.700544708942886		0.48358719428854		0.599[illegible]972093409	
CPAR2_212570	559.37[illegible]382948797		0.028735866112831[illegible]		0.12662336890635
0.226933518350279		0.820475438026274		0.876935520[illegible]8357	
CPAR2_302130	173.841921340983		0.35023424511046		0.320288521371104
1.09349608787466		0.274176045726641		0.389376678031688	
CPAR2_209760	954.765738326101		-0.114035759306079		
0.164726548624648		-0.692273105083535		0.488765836618937	0.604529147169239
CPAR2_403260	2857.44[illegible]847832337		0.358640646962554		0.197[illegible]1989899545
1.81450562245708		0.0695999179174566		0.137646093209841	
CPAR2_209040	491.80[illegible]116175746		0.140998897015593		0.3091[illegible]624371794
0.456130526498074		0.648296119826938		0.744923394087986	
CPAR2_804040	50.1[illegible]59296854689		0.345329251780373		0.261316423135524
1.32149846395716		0.186335214355602		0.291206670418996	
CPAR2_211500	6[illegible].715757128726		0.006496216159173[illegible]		0.119238635927135
0.0544808447088[illegible]		0.956552069628808		0.968184140922596	
CPAR2_102540	1997.88804218508		0.236796415663859		0.148912116775271
1.[illegible]			67	0.1976103	
CP[illegible]			5	-0.913291	
0.[illegible]			0435876423	1	
CP[illegible]			3	0.0920878	
0.[illegible]			1457826686	0	
CP[illegible]			5	-0.076716	
0.[illegible]			84540483756	0	
CP[illegible]			4	-0.918009	
0.[illegible]			8398147556	0	
CPAR2_206160	622.942944989741		-0.399161611153727		

log2Fold Change

pvalue and padj

P-values to quantify confidence levels

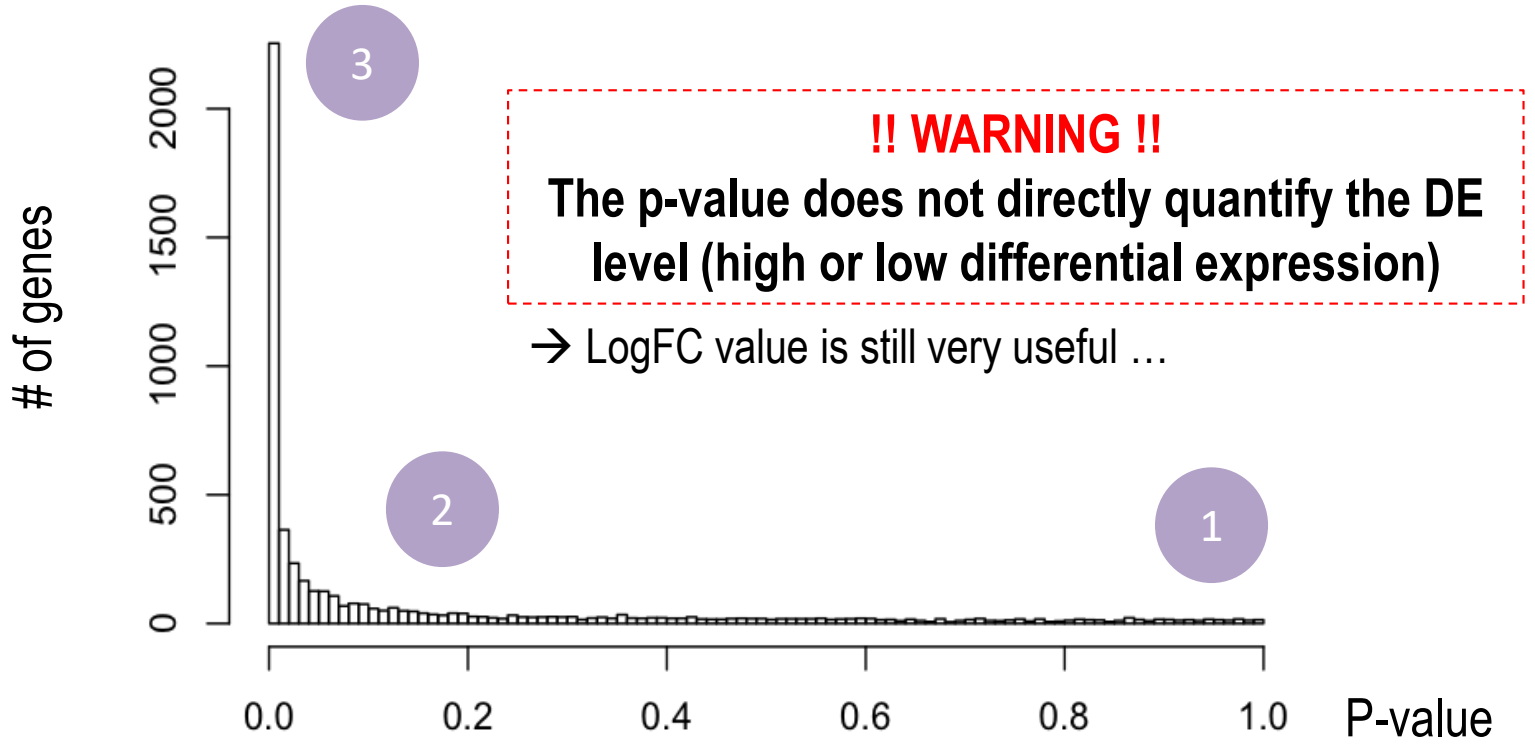
➢ A p-value **quantifies the risk** to select a gene as differentially expressed, whereas it is not



Dataset: Guida et al., BMC Genomics (2011); LIMMA method: Smyth et al., BMC Genomics (2009)

P-values to quantify confidence levels

- A p-value **quantifies the risk** to select a gene as differentially expressed, whereas it is not



!! WARNING !!

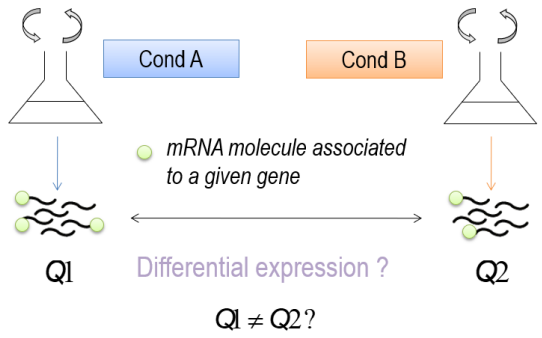
The p-value does not directly quantify the DE level (high or low differential expression)

→ LogFC value is still very useful …

Dataset: Guida et al., BMC Genomics (2011); LIMMA method: Smyth et al., BMC Genomics (2009)

Multiple testing correction (do you remember ?)

What is the question ?



➢ All genes are successively analyzed in a differential analysis

If conditions A and B are identical (no DE gene), since many statistical tests are perform (for instance 10 000) ...

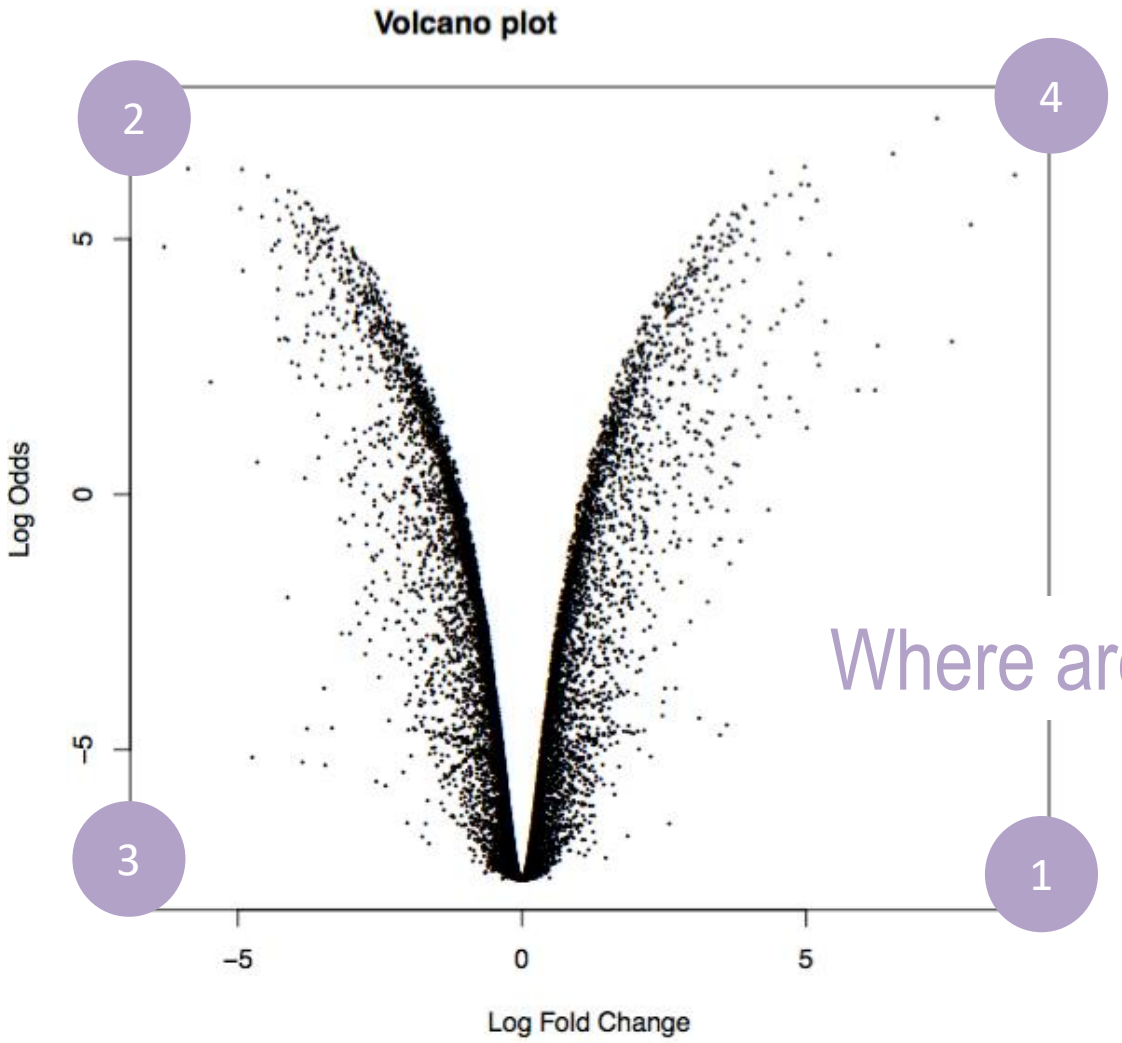
… we can expect **500 false positive** genes*, with a standard p-value cut off of 5%

* Expected value according to a binomial distribution

Combining logFC and adj. P-values …

Low adj. pval

High adj. pval



Where are DE genes ?

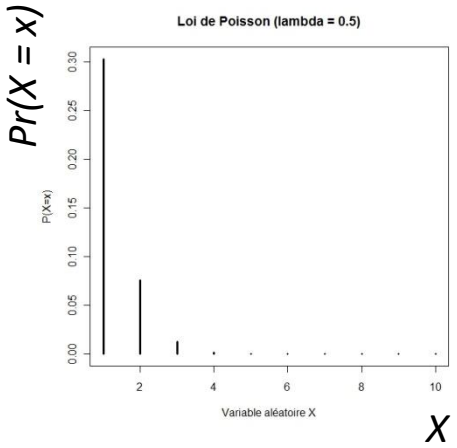
Other strategy to calculate p-values : Poisson Model

$$p(k) = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

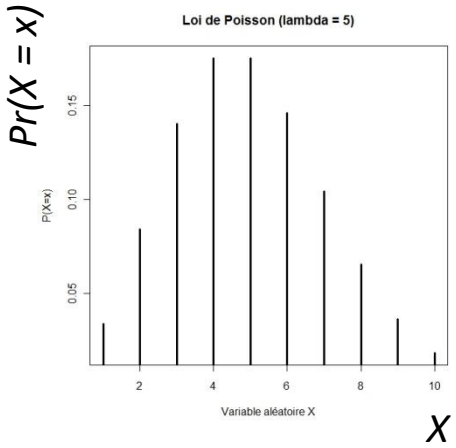
- Do you remember ?
  - Bernoulli trial
  - Binomiale distribution
  - Poisson distribution (rare events)

How to estimate the value of λ ?

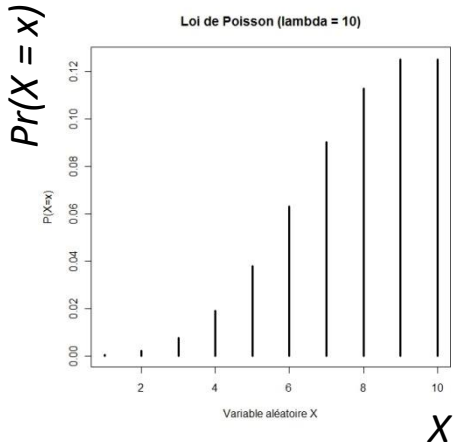
λ = 0.5



λ = 5



λ = 10



Application to HTS data analysis

Read sequences

Gene i

x reads
associated to gene i

Bernoulli trial =
Sequencing of a read

p **1 - p**

Gene i | Other gene

Repetition of Bernoulli trials (for each read)

❖ The number of reads associated to the gene *i* is a random variable that follow a Binomial distribution B(n,p).

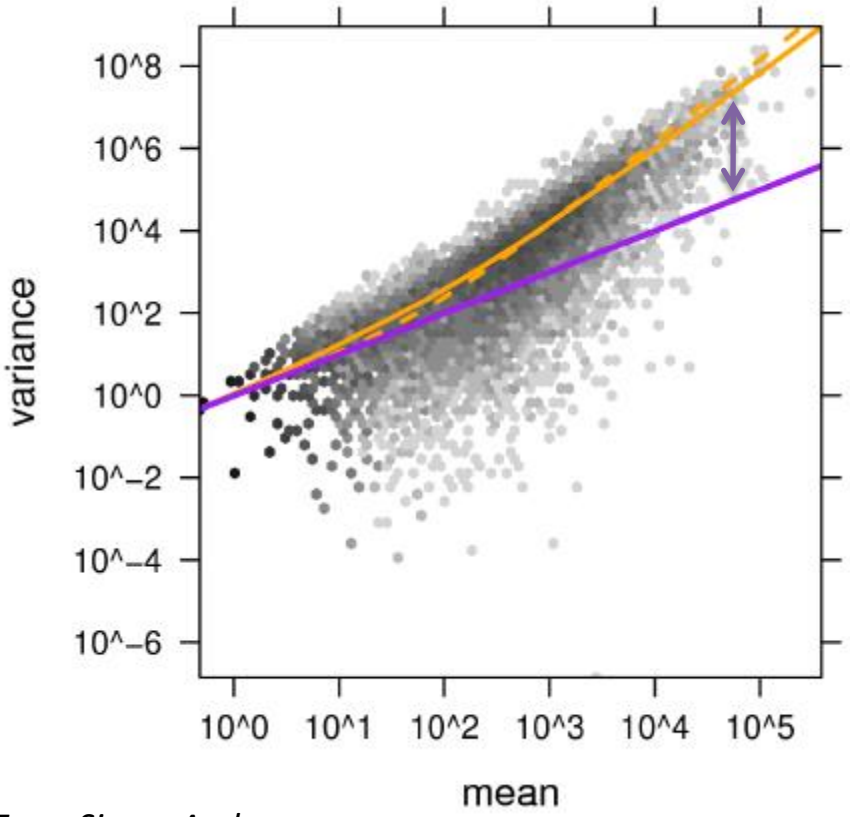
❖ As n is high and p small, Poisson distribution can be used.

P and λ are unknown parameters

Is the Poisson Model Relevant ?

❖ The Poisson model has the major interest to require the estimation of only one parameter (λ).

Mean and variance calculated from two replicates



From Simon Anders

Property:
$E(X) = Var(X) = \lambda$

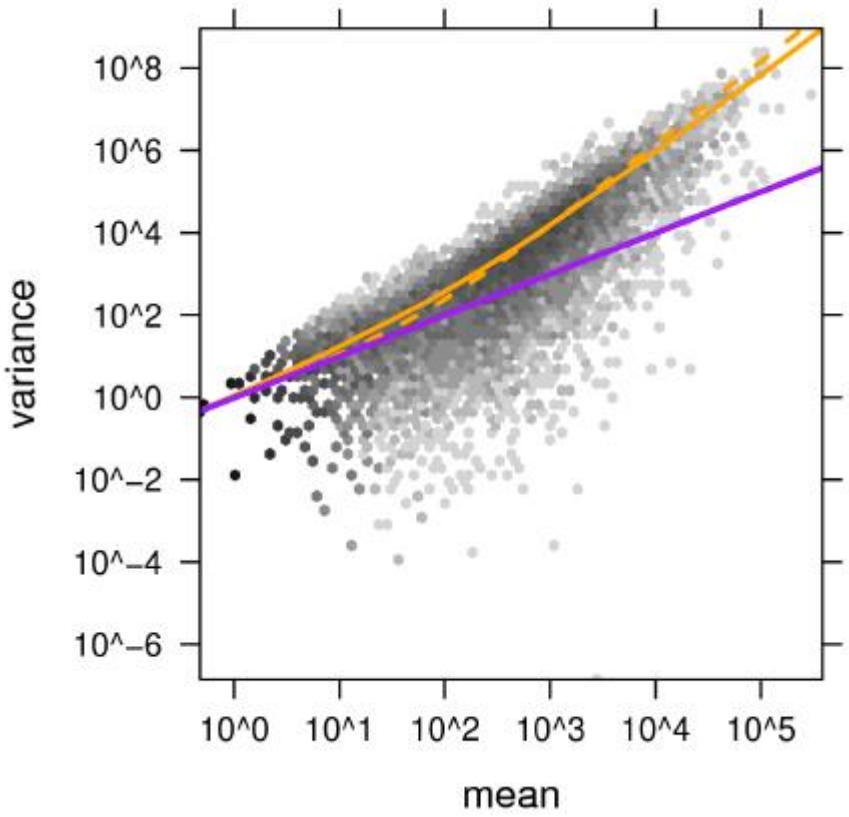
❖ Distribution expected if the Poisson model is true.

Variability associated to the number of reads increases more than the mean value

❖ Poisson model under-estimates the variability for highly expressed genes (Increases false positive rate).

Variance Estimation is “the key”

Mean and variance calculated from two replicates



Variance calculated from comparing two replicates

Poisson	$v = \mu$
Poisson + constant CV	$v = \mu + \alpha \mu^2$
Poisson + local regression	$v = \mu + f(\mu^2)$

❖ ***DESeq*** (Anders et al. 2010):

❖ ***edgeR*** (Robinson et al. 2008)):

From Simon Anders

Gene expression level is another important information

- Two genes can have identical logFC, but different levels of transcriptional activity

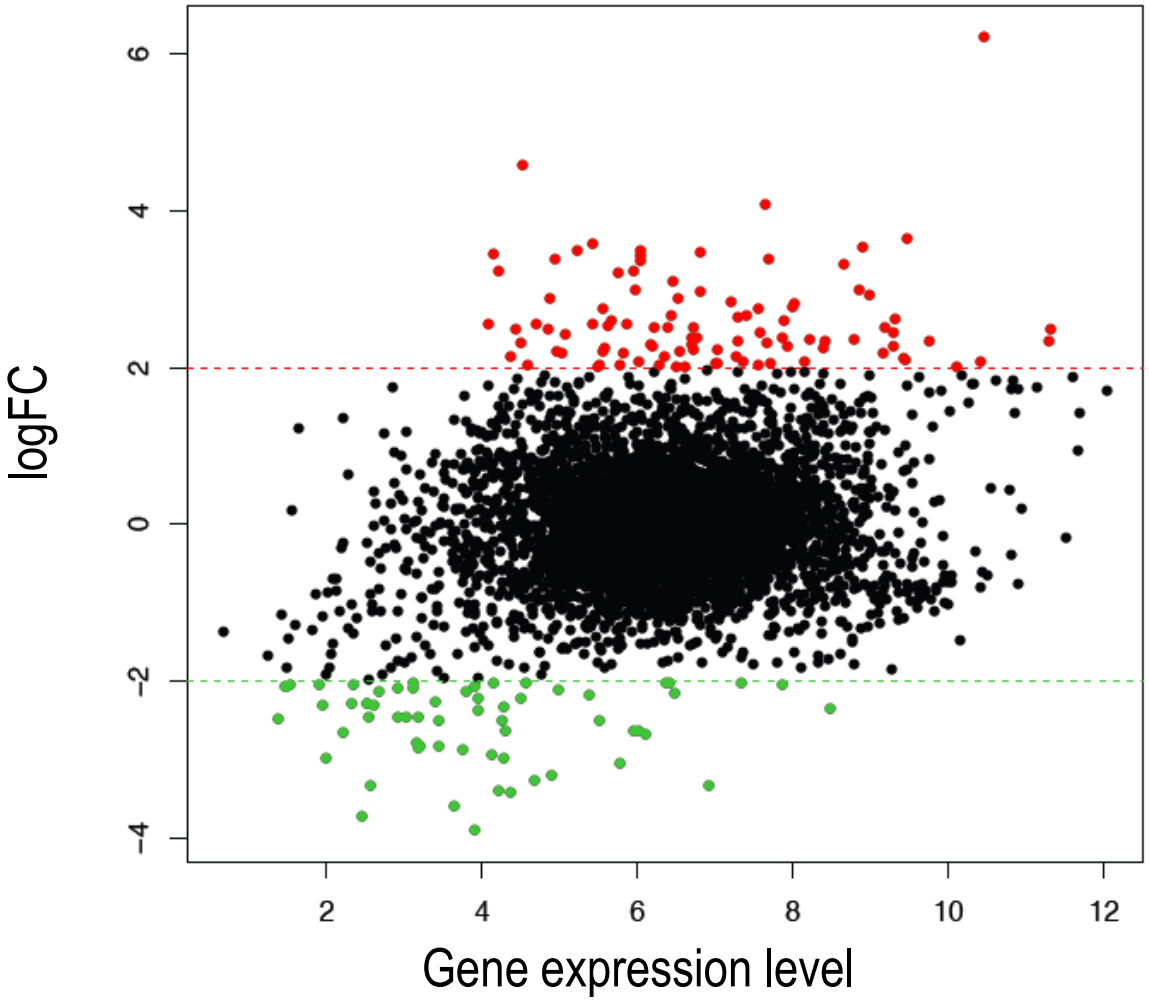
+ + + + Gene 1: $\log FC_{G1} = \log 2(8000/1000) = 3$

Gene 2: $\log FC_{G2} = \log 2(8/1) = 3$

Random fluctuation ?

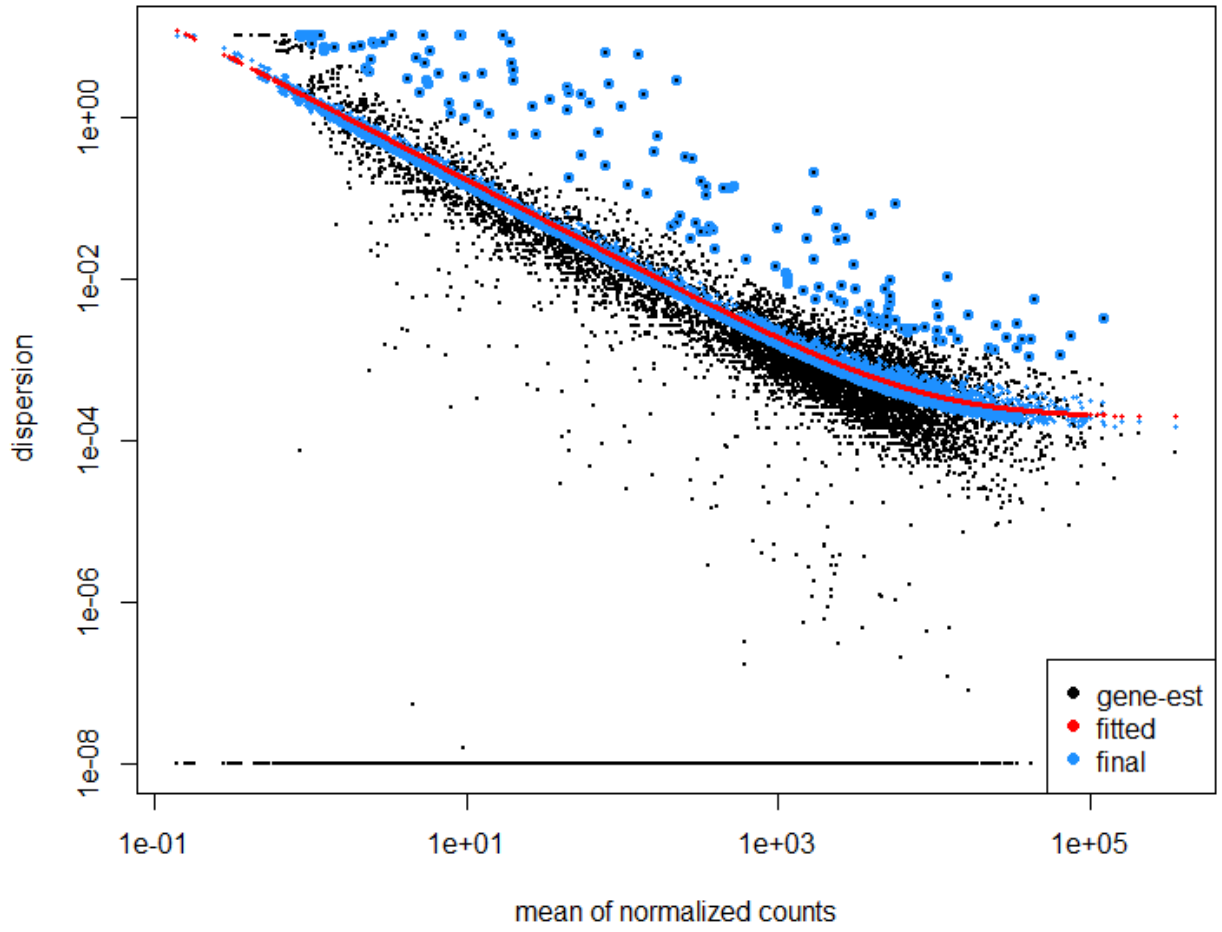
Same level of confidence ?

Gene expression level is another important information

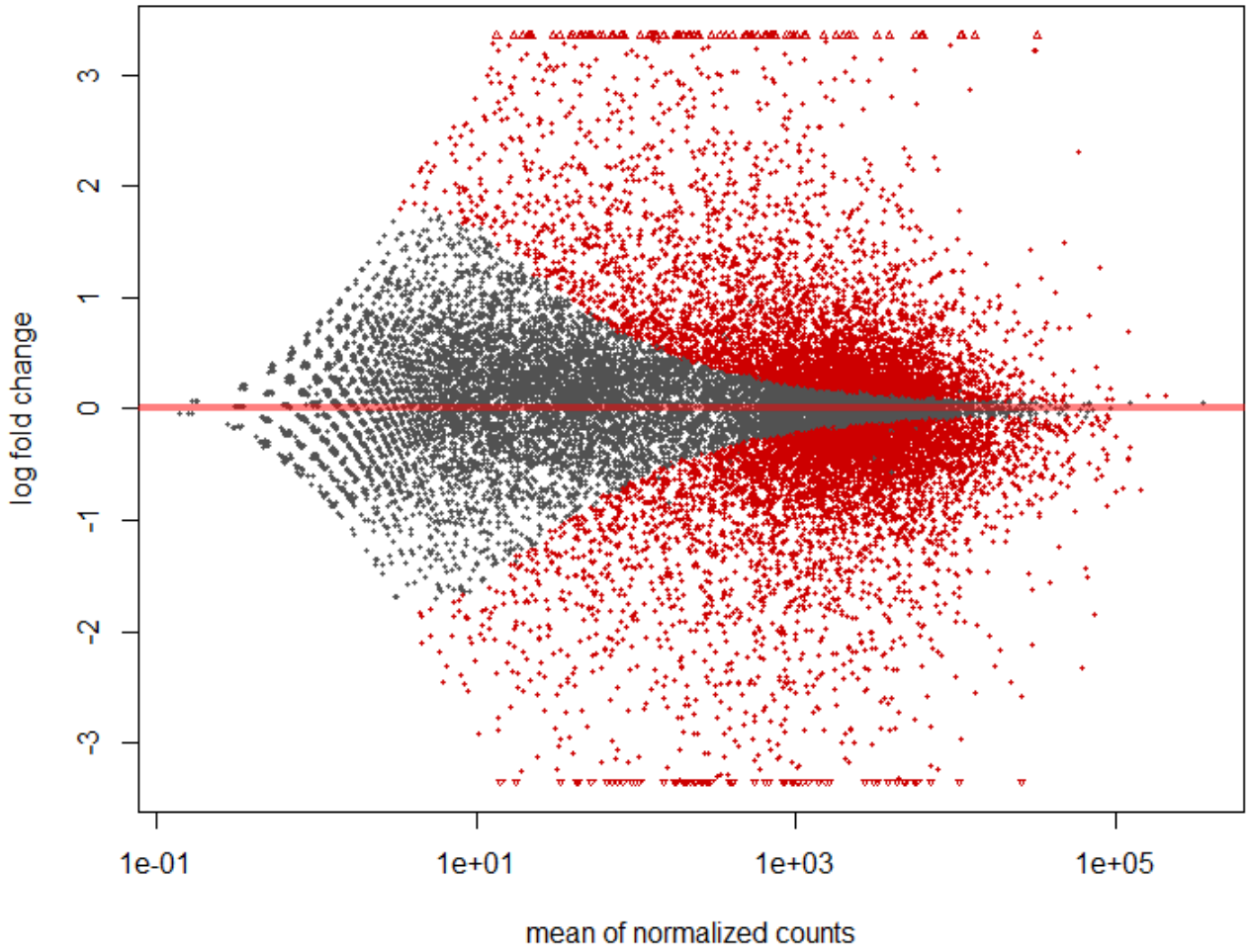


Source of dataset: Guida et al., BMC Genomics (2011)

Results obtained with DESeq2 (1/2)



Results obtained with DESeq2 (2/2)



R script to be used/modified

Search_DEgenes.R

search_DEgenes_M2Bioinfo.R

search_DEgenes_M2Bioinfo.R 〉 No Selection

```
#####
# Gaelle LELANDAIS <gaelle.lelandais@u-psud.fr>
# Melina GALLOPIN <melina.gallopin@u-psud.fr>
#####

####
# Ce script R est destine aux etudiants de Master. Il a pour objectif de les aider dans la
# comprehension des methodes de selection des genes differentiellement exprimes.
####

####
# Librairies R necessaires : LIMMA, MASS, DESeq2, edgeR
####

#-------------------------------------------------------------------------------------------
#-------------------------------------------------------------------------------------------

pdf("graphics.pdf")

# Lecture des données
countData = read.table("count_dataFile.txt", header = T, row.names = 1)
# --> Ces donnees sont en relation avec l'article :
#
# "Determination of a Comprehensive Alternative Splicing Regulatory Network and
# Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition"
#
# Yang et al. (2016, Molecular and Cellular Biology)
#
# --> Deux points de temps ont ete etudies par RNAseq : "Day0" et "Day7", 3 replicats sont
# disponibles pour chaque point de temps.

#-------------------------------------------------------------------------------------------
# Pretraitement des donnees entre les experiences (filtrage des faibles et normalisation)
#-------------------------------------------------------------------------------------------

## Suppression des genes sans valeurs (0 pour toutes les experiences)
length(which(rowSums(countData)==0))  ## 2957
countData <- countData[-which(rowSums(countData) == 0),]

# Distribution des valeurs de comptages dans les différents échantillons
boxplot(log(countData + 1), ylab = "#reads (log scale)",
        main = "Boxplot of read counts", col = c(rep("grey", 3), rep("blue", 3)),
        names = c(paste("Day0_", 1:3, sep = ""), paste("Day0_", 1:3, sep = "")))
# --> Les distributions des donnees de comptage sont tres proches entre les differentes
# conditions.
```

To conclude,

Statistics help you to take a decision (is the gene differentially expressed ?) …

… it should not take the decision for you

- **Important parameters to consider in a DE analysis:**

Differential expression

Confidence level

Gene information in databases, literature …

Gene expression level

Data reproducibility

SECTION 3

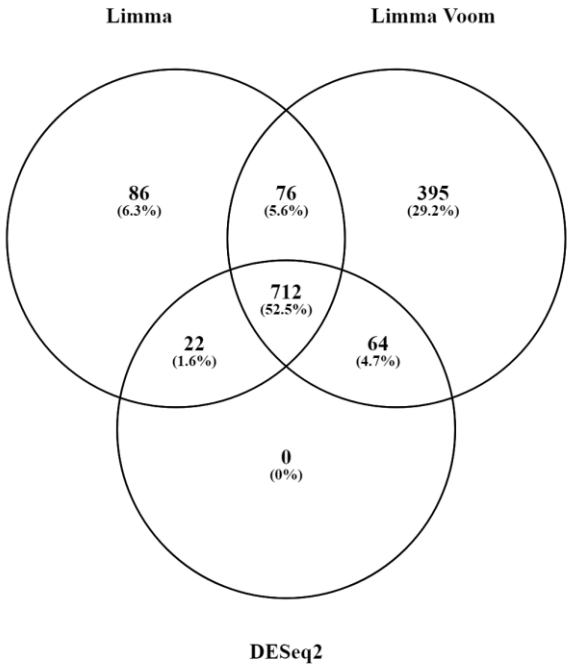
PRACTICAL TRAINING

Your work …

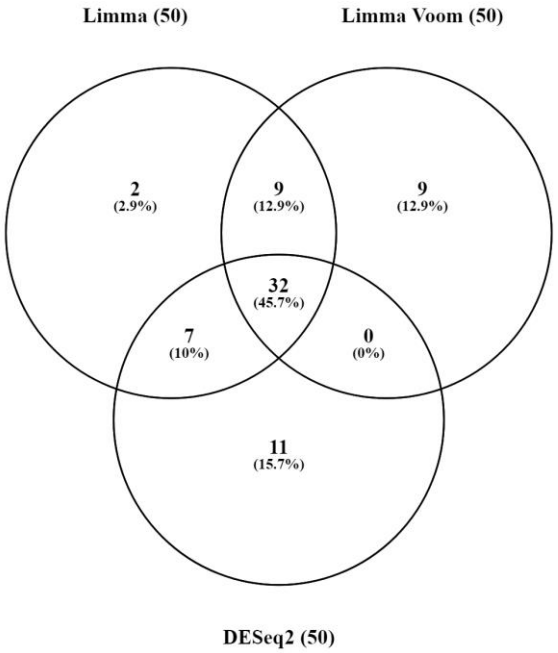
- Compare results obtained with different approaches (logFC threshold, DESeq2, etc.)
- Search for genes located on Chr18 in your lists of DE genes. Check for read coverage with IGV
- Define (or refine) your own list of “favorite genes”

Method comparisons

- Genes with logFC values > 2 ou < -2 AND adj.pal < 0.01 were selected and compared between the three methods :

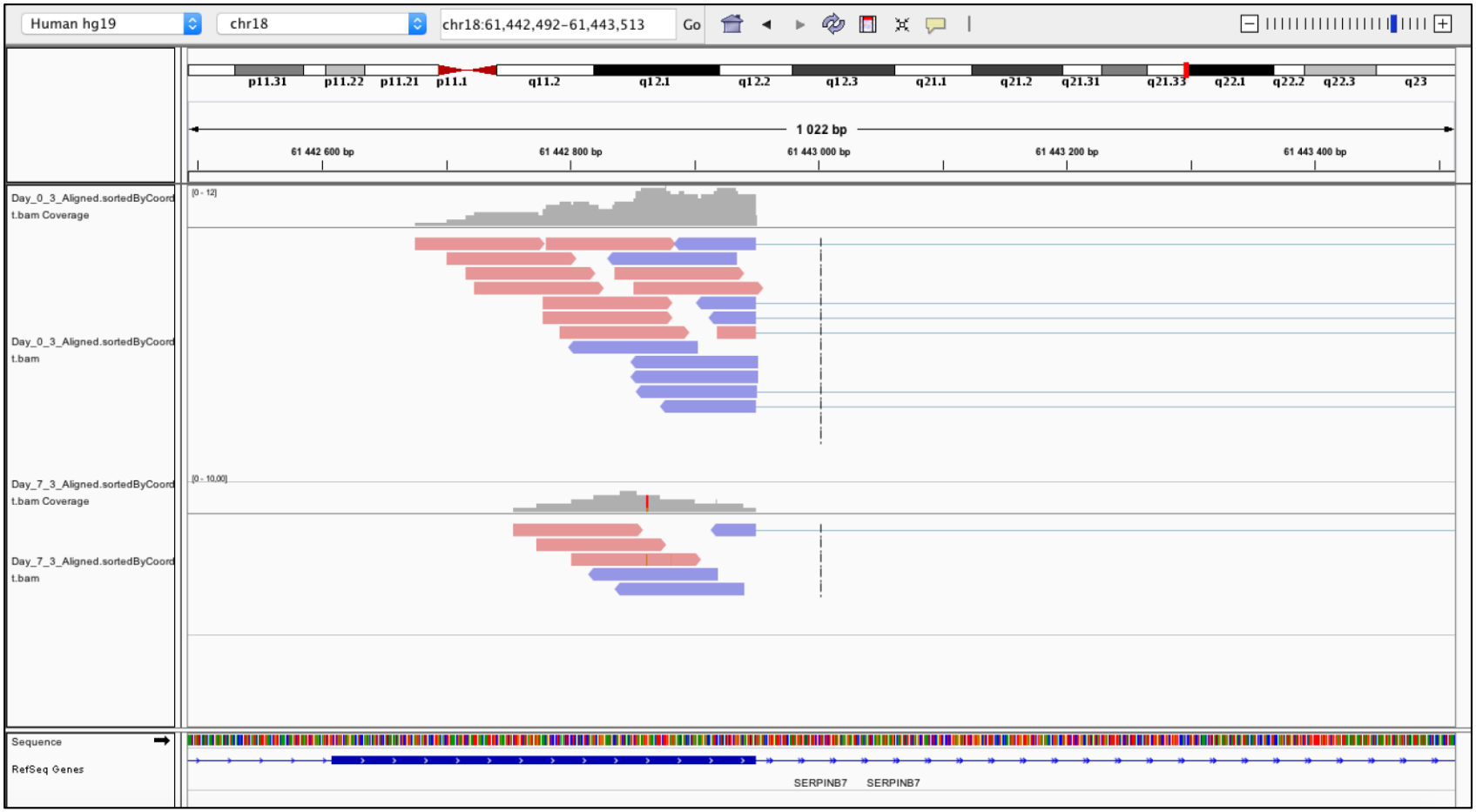


All selected genes



Top 50 genes

DE genes verified with IGV (1/2)



DE genes verified with IGV (2/2)