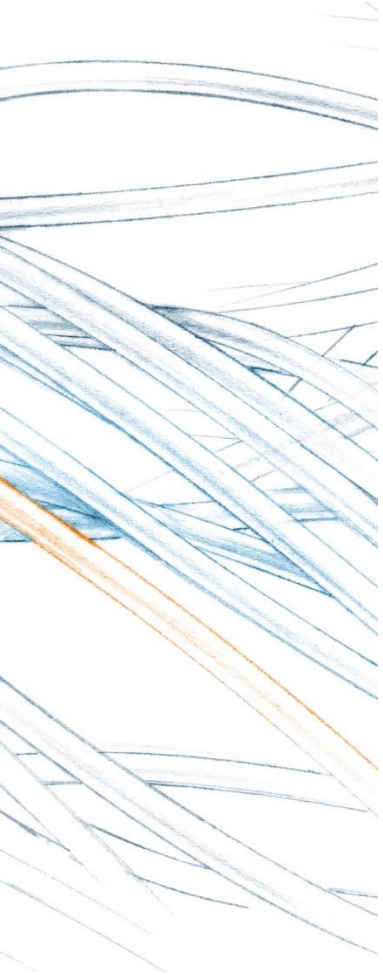
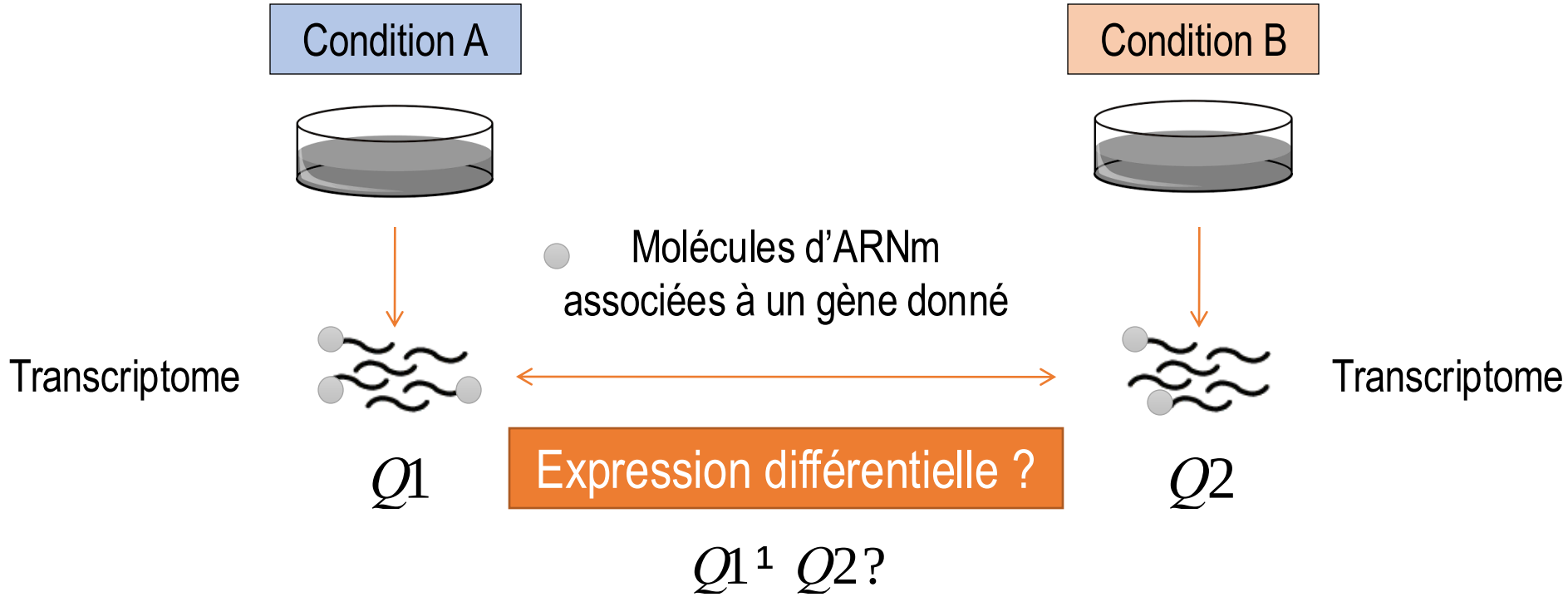


L'analyse différentielle,

taille d'effet, dispersion et force d'expression

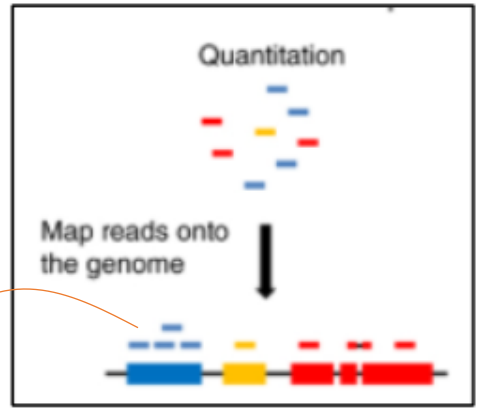
Gaëlle LELANDAIS

Quelle est la question ? (analyse différentielle)



➢ Tous les gènes sont étudiés successivement
(répétition de la question plusieurs dizaines de milliers de fois)

Les données analysées

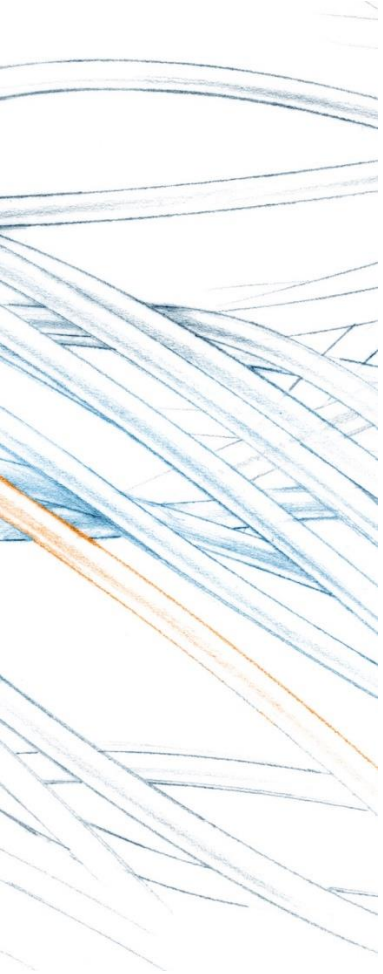


➢ Table de comptages (*counts*) :

	Condition A			Condition 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 des données : Yang et al., Mol Cell Biol (2016)

Problématique de normalisation (1/2)



➢ Le nombre de sequences (*reads*) depend de la longueur des transcrits …

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

Transcrit du gène A

8 *reads* associés

Transcrit du gène B

4 *reads* associés

Problématique de normalisation (2/2)

➢ Le nombre de sequences (*reads*) dépend de la somme totale* (par colonne) …

	Somme 1	Somme 2
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

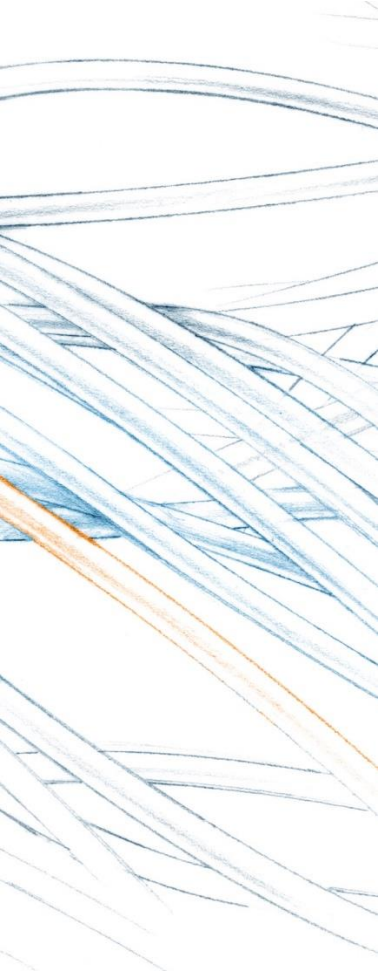
Biais systématique ?

(si par exemple Somme 1 = 40 000 et Somme 2 = 30 000)

Table de comptage normalisée

* La somme des comptages (*counts*) par expérience est souvent nommé *library size* dans les articles.

Quantifier l'expression différentielle



➢ Ratio

$$R = \frac{Q_A}{Q_B} = \frac{Valeur(s)\ de\ comptage\ Condition\ A}{Valeur(s)\ de\ comptage\ Condition\ B}$$

➢ Fold Change

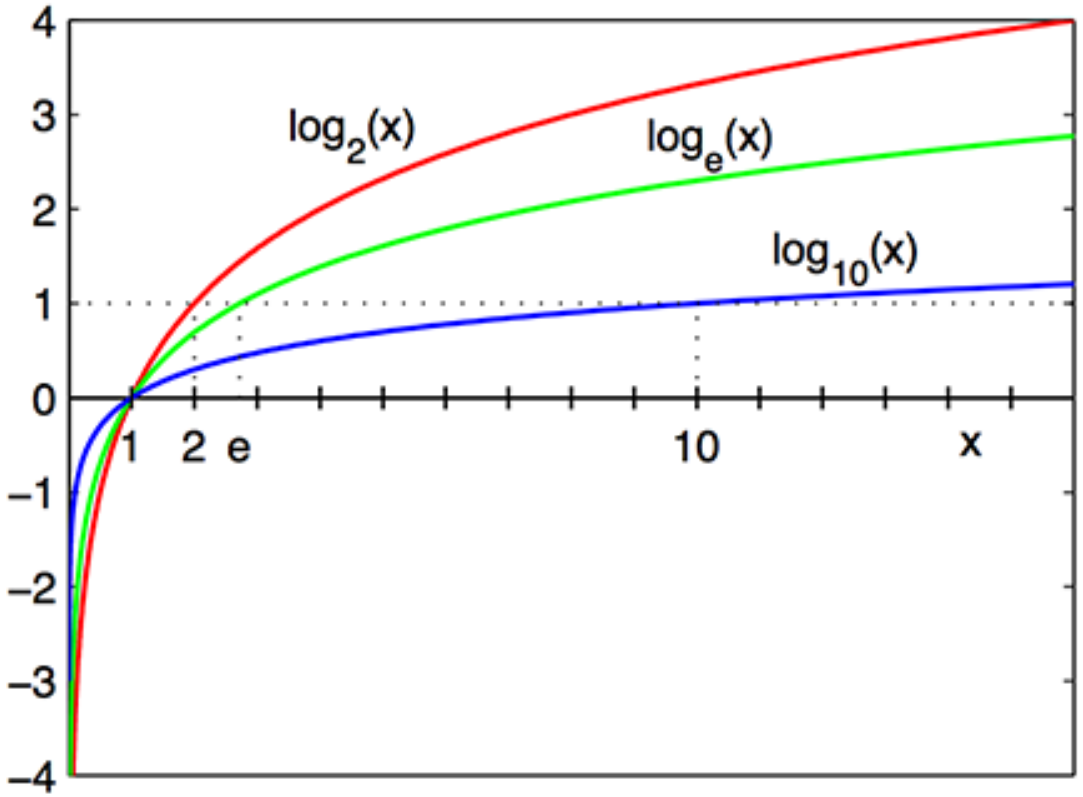
$$FC = \begin{cases} R\ si\ R \geq 1 \\ \dfrac{-1}{R}\ si\ R < 1 \end{cases}$$

➢ Log fold change

$$logFC = log2(R) = log2(Q_A) - log2(Q_B)$$

Interprétation du logFC (taille d'effet)

- Les valeurs du logFC peuvent être positives et negatives. L'utilisation de la base 2 permet de traduire un doublement par une unité de variation (+/-).



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

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

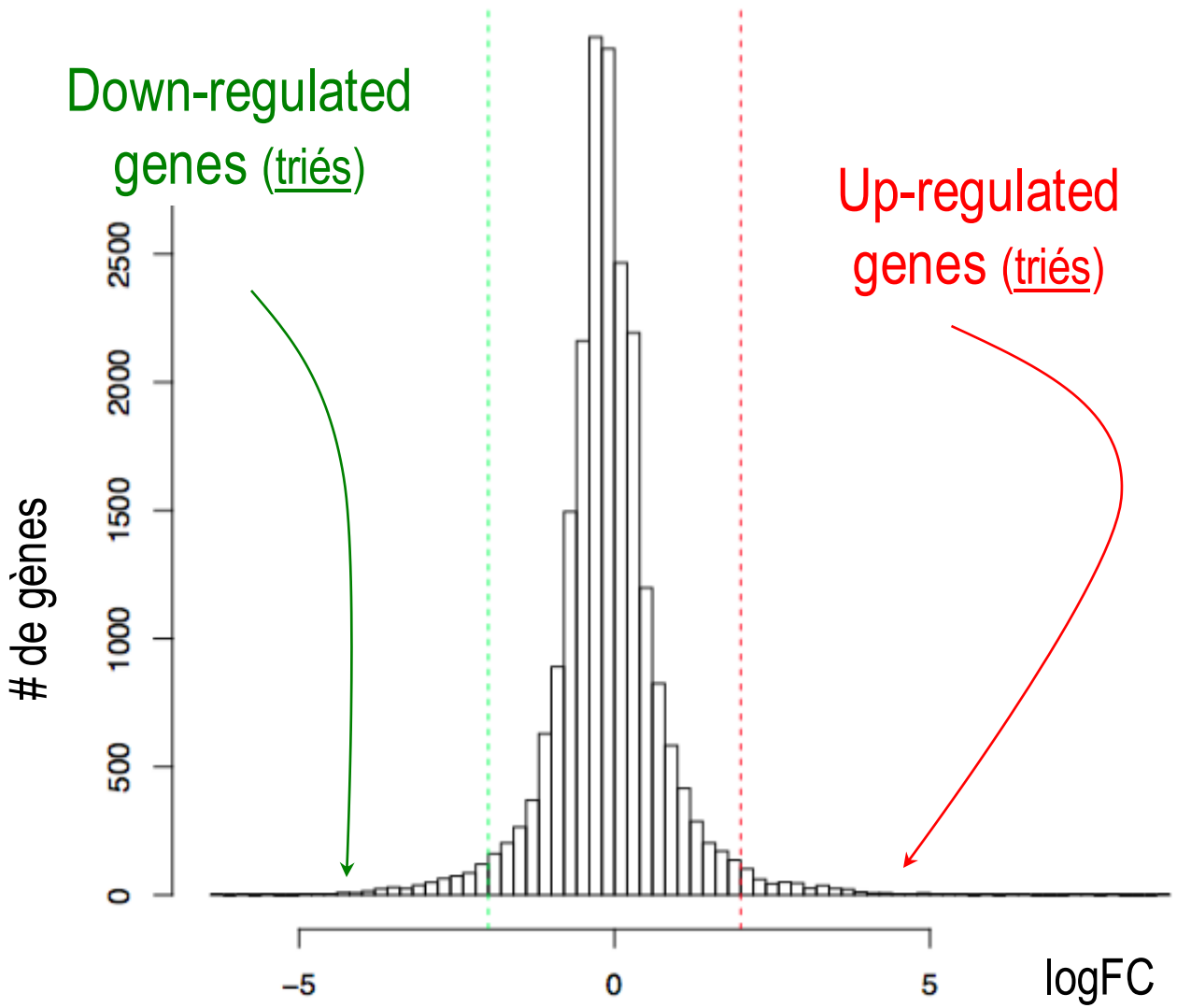
↳ "Up-regulated gene"

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

↳ "Down-regulated gene"

Notion de taille d'effet

Utiliser la taille d'effet



Source des données : Yang et al., Mol Cell Biol (2016)

Liste de gènes, candidats pour être différentiellement exprimés

Tenir compte de la reproductibilité

(dispersion)

- Dans la pratique, les expériences sont répétées. La taille d'effet mesurée est fondée sur plusieurs observations.

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

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

Meilleure cohérence entre les observations

Notion de dispersion

Quantifier la dispersion*

➢ Variance (estimateur)

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

➢ Ecart type

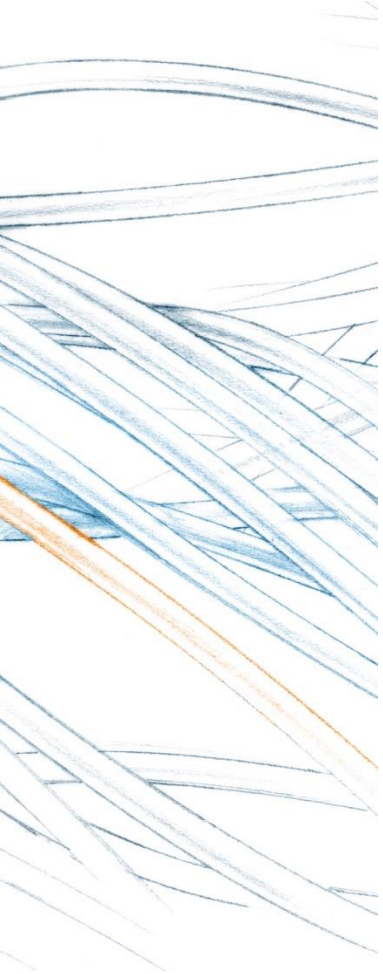
$$SD = \sqrt{Var(X)}$$

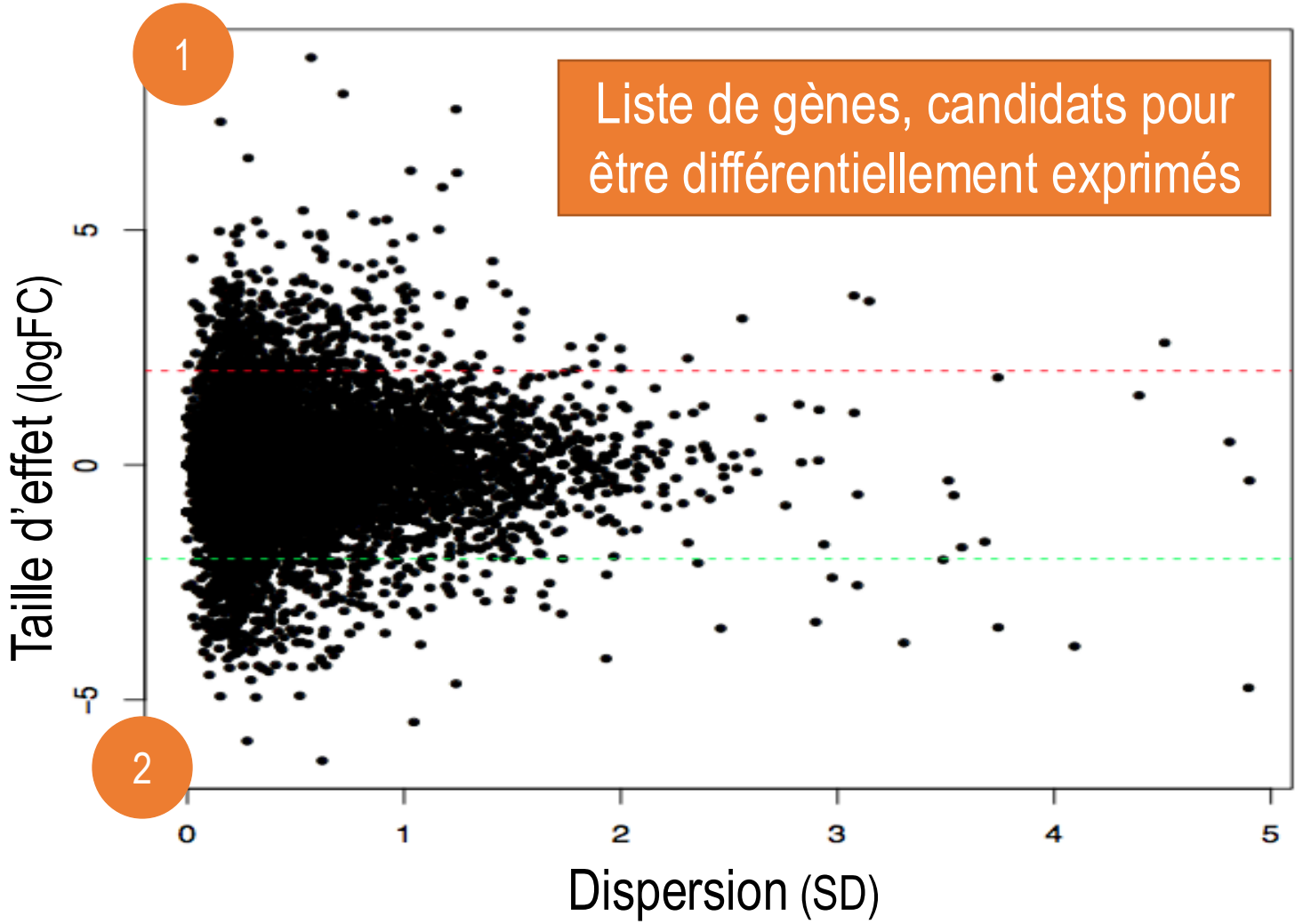
➢ Exemple :

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

* Mais aussi le coefficient de variation biologique, l'erreur standard, etc.

Combiner taille d'effet et dispersion



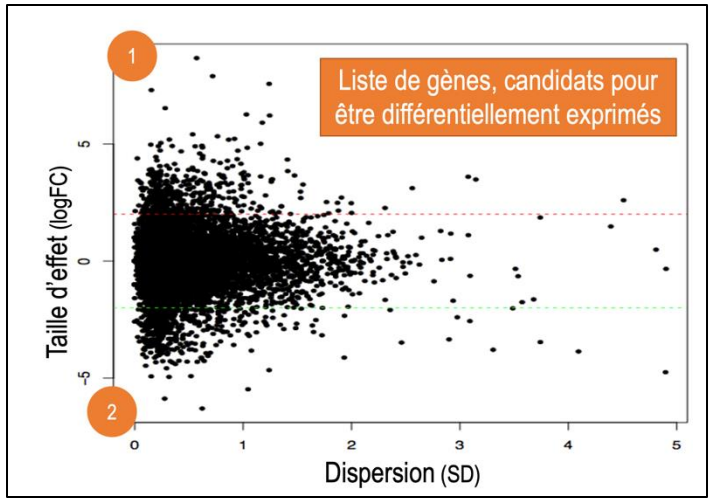


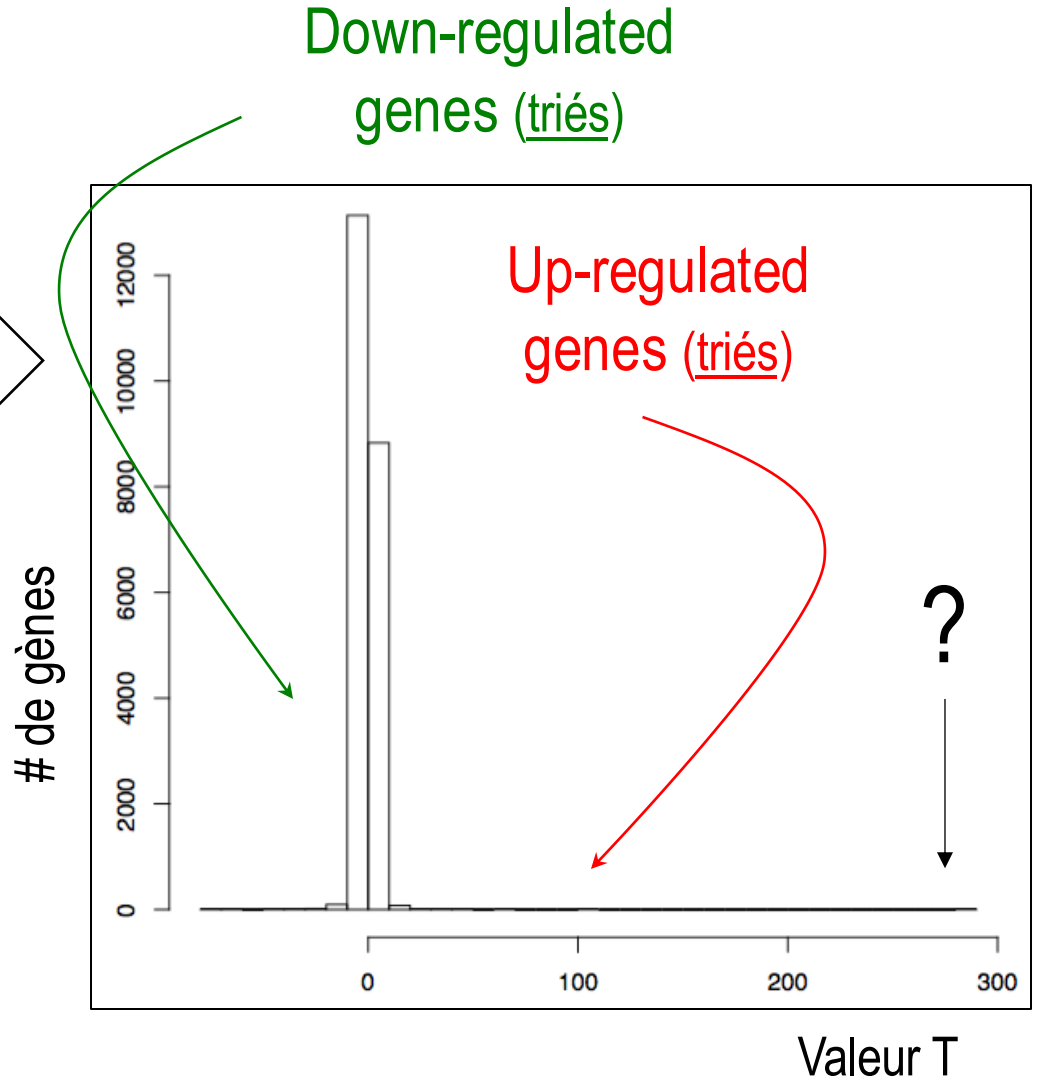
Source des données : Yang et al., Mol Cell Biol (2016)

Statistique de Student

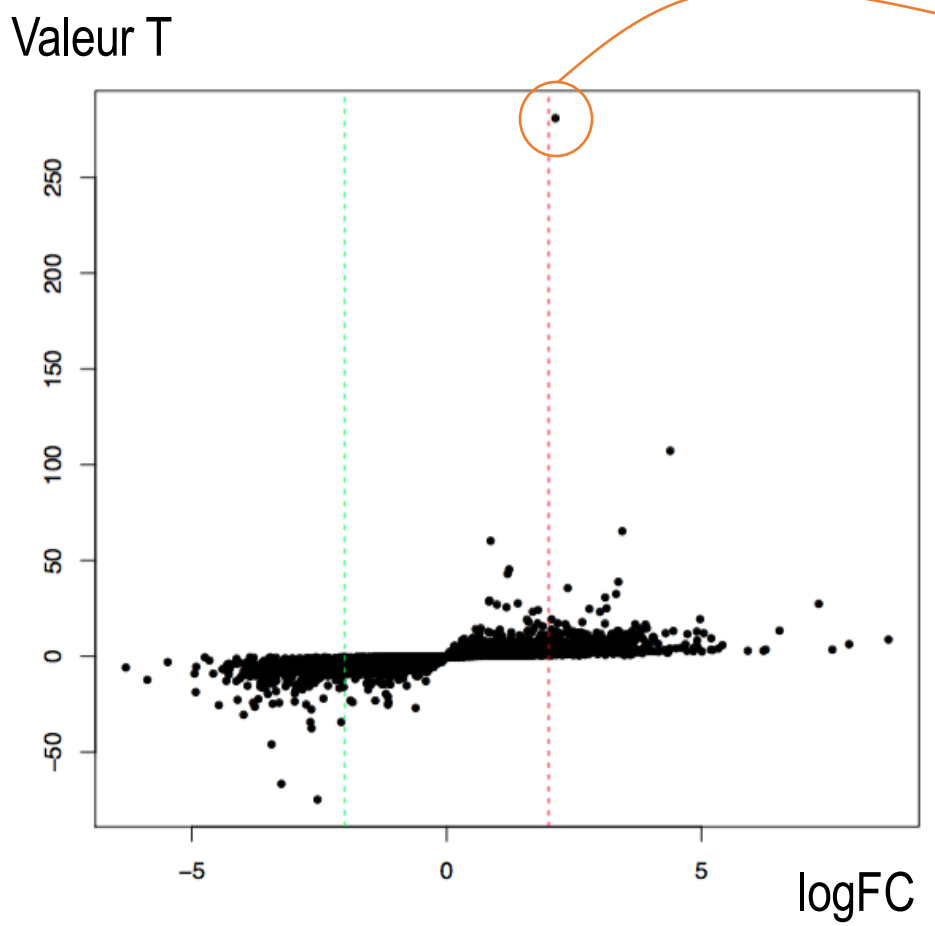
- Taille d'effet, dispersion, nombre d'observations :

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





… malheureusement non !



Importance trop grande donnée au paramètre de dispersion (n = 3)

- Estimer correctement la variabilité associée aux observations de comptage d'un gène est un défi statistique.

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 des données : Yang et al., Mol Cell Biol (2016)

Tenir compte du nombre de séquences alignées (force d'expression)

➢ Une même taille d'effet peut être obtenues avec des données de comptages différentes.

Gene 1:

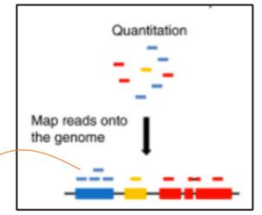
$$\log FC_{G1} = \log 2(8000/1000) = 3$$

Gene 2:

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

Variations aléatoires ?

Les données analysées



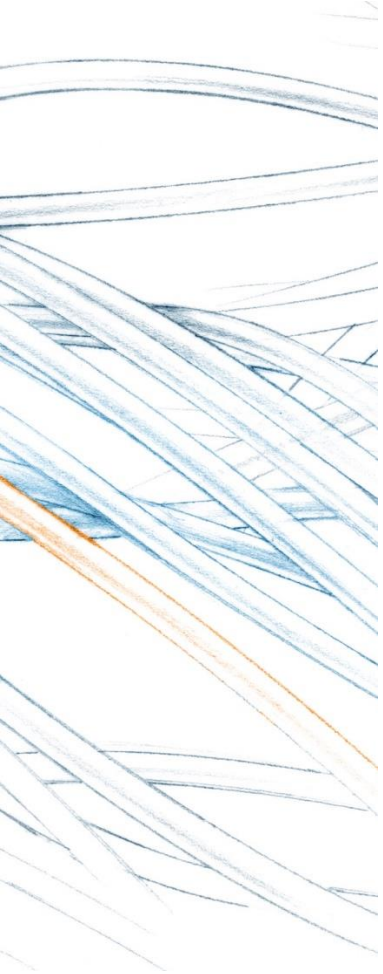
➢ Table de comptages (*counts*) :

	Condition A			Condition 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

La notion de "force d'expression" est intéressante à prendre en compte pour établir une liste de gène candidats, avec une meilleure fiabilité.

* Cette information est nommée *expression strength* dans les articles statistiques.

Bilan



- L'analyse différentielle est un problème difficile qui nécessite l'utilisation de méthodologies d'analyses adaptées.

DESeq et DESeq2 →

edgeR →

LIMMA →

TABLE 1. RNA-seq differential gene expression tools and statistical tests

Name	Assumed distribution	Normalization	Description	Version	Citations[d]	Reference
t-test	Normal	DEseq[a]	Two-sample *t*-test for equal variances	–	–	–
log *t*-test	Log-normal	DEseq[a]	Log-ratio *t*-test	–	–	–
Mann-Whitney	None	DEseq[a]	Mann-Whitney test	–	–	Mann and Whitney (1947)
Permutation	None	DEseq[a]	Permutation test	–	–	Efron and Tibshirani (1993a)
Bootstrap	Normal	DEseq[a]	Bootstrap test	–	–	Efron and Tibshirani (1993a)
baySeq[c]	Negative binomial	Internal	Empirical Bayesian estimate of posterior likelihood	2.2.0	159	Hardcastle and Kelly (2010)
Cuffdiff	Negative binomial	Internal	Unknown	2.1.1	918	Trapnell et al. (2012)
DEGseq[c]	Binomial	None	Random sampling model using Fisher's exact test and the likelihood ratio test	1.22.0	325	Wang et al. (2010)
DESeq[c]	Negative binomial	DEseq[a]	Shrinkage variance	1.20.0	1889	Anders and Huber (2010)
DESeq2[c]	Negative binomial	DEseq[a]	Shrinkage variance with variance based and Cook's distance pre-filtering	1.8.2	197	Love et al. (2014)
EBSeq[c]	Negative binomial	DEseq[a] (median)	Empirical Bayesian estimate of posterior likelihood	1.8.0	80	Leng et al. (2013)
edgeR[c]	Negative binomial	TMM[b]	Empirical Bayes estimation and either an exact test analogous to Fisher's exact test but adapted to over-dispersed data or a generalized linear model	3.10.5	1483	Robinson et al. (2010)
Limma[c]	Log-normal	TMM[b]	Generalized linear model	3.24.15	97	Law et al. (2014)
NOISeq[c]	None	RPKM	Nonparametric test based on signal-to-noise ratio	2.14.0	177	Tarazona et al. (2011)
PoissonSeq[c]	Poisson log-linear model	Internal	Score statistic	1.1.2	37	Li et al. (2012)
SAMSeq[c]	None	Internal	Mann-Whitney test with Poisson resampling	2.0	54	Li and Tibshirani (2013)

[a]See Anders and Huber (2010).
[b]See Robinson and Oshlack (2010).
[c]R (v3.2.2) and bioconductor (v3.1).
[d]As reported by PubMed Central articles that reference the listed reference (December 21, 2015).

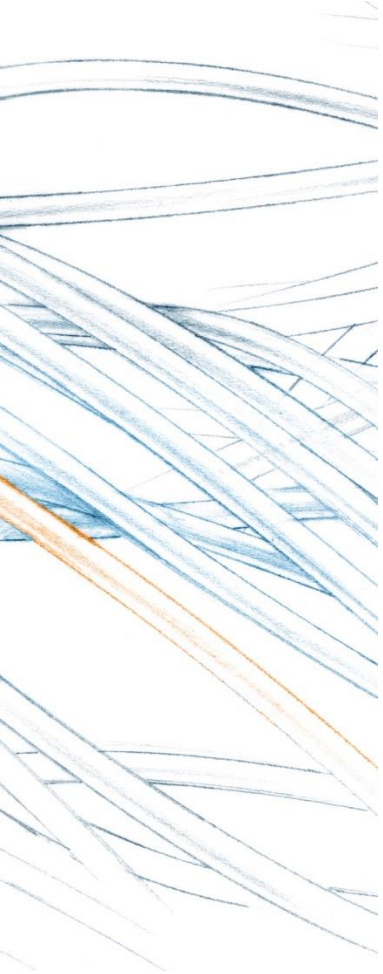
Taille d'effet

Dispersion

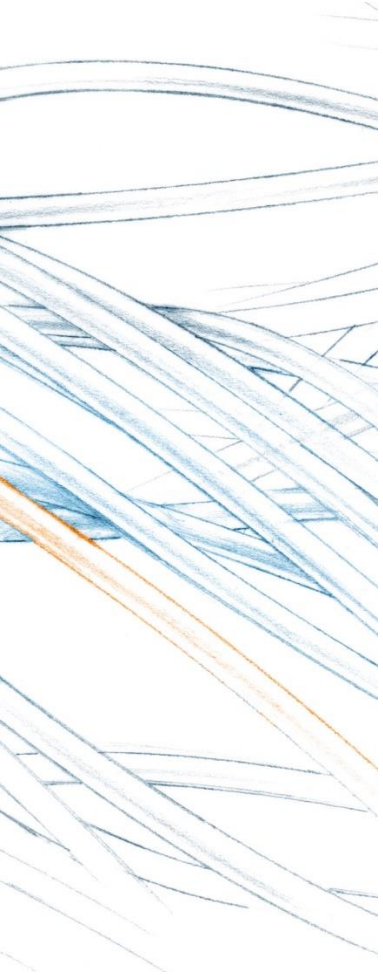
Force expression

http://www.rnajournal.org/cgi/doi/10.1261/rna.053959.115

Ressource(s) complémentaire(s)



- Commentaires enregistrés du cours (16 minutes) :
  - Lien Youtube : https://youtu.be/eztUy86l4ds



Gaëlle LELANDAIS

Version du document : 11/02/2025