

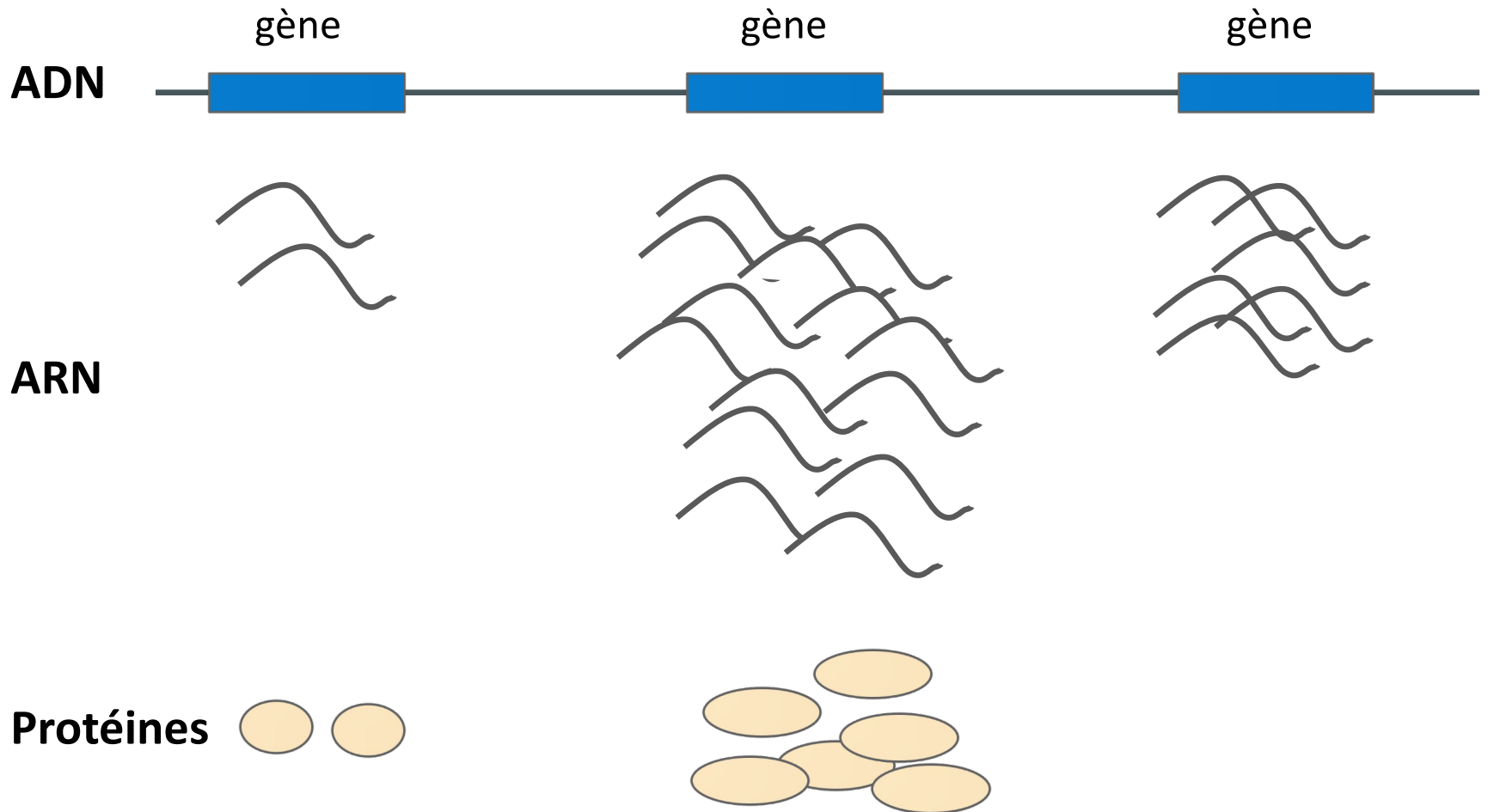
# Quantification de l'expression des gènes et analyse de l'expression différentielle avec RNA-seq

Adil El Filali et Anamaria Necsulea

18/10/2023



# Comment mesurer le niveau d'expression des gènes ?

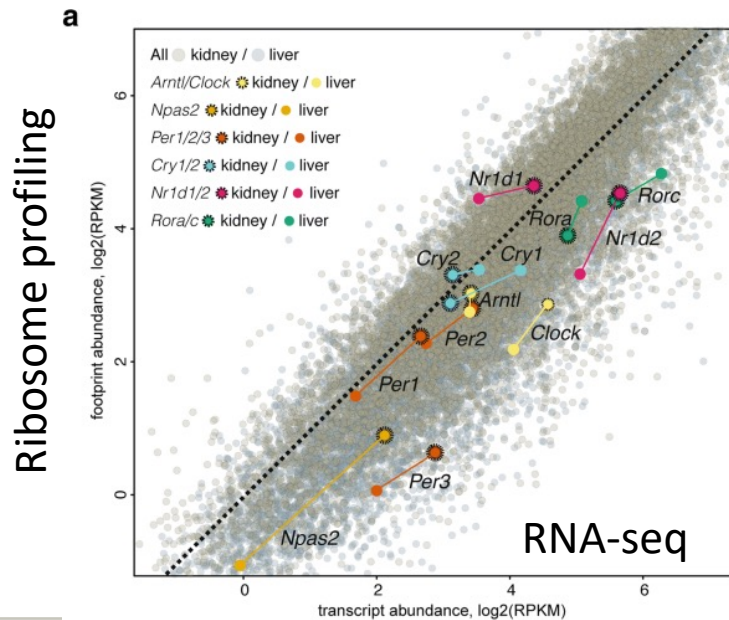


# Comment mesurer le niveau d'expression des gènes ?

- L'activité du gène peut être mesurée à plusieurs niveaux : taux de transcription, quantité d'ARN produit (à l'état stable), quantité de protéines produites.
- Plusieurs niveaux de régulation de l'activité des gènes : régulation transcriptionnelle, post-transcriptionnelle, traductionnelle.
- Une définition usuelle du niveau d'expression du gène : **nombre de molécules d'ARNm produites par le gène**, par cellule, tous isoformes confondus.

# Que peut-on mesurer avec la méthode RNA-seq ?

- RNA-seq classique : on ne mesure pas **le taux de transcription** !
- On a accès à la **quantité d'ARNs à l'état stable** = transcription + (modification post-transcriptionnelle) + dégradation.
- Approximation de l'abondance des protéines ?

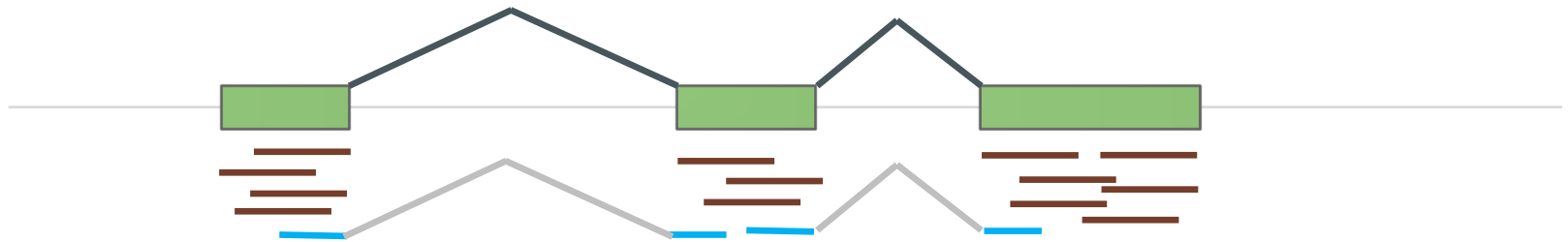


Castelo-Szekely *et al.*, 2017

# On mesure des niveaux d'expression **relatifs**

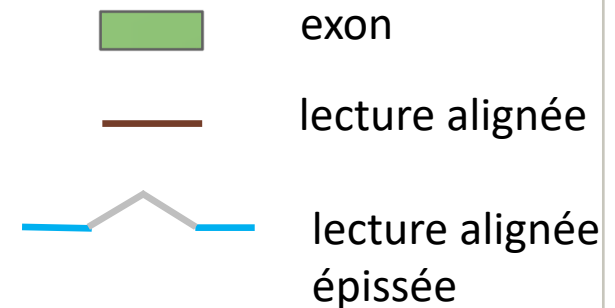
- Les niveaux d'expression obtenus avec des données de RNA-seq « classiques » sont toujours **relatifs** aux autres gènes exprimés dans l'échantillon analysé.
- On n'a pas directement accès aux nombres de molécules d'ARN.
- Comment faire pour s'approcher le plus possible des niveaux d'expression absolus ?
  - utilisation de « spike-in » RNAs : molécules en quantité contrôlée, rajoutées au pool d'ARN lors de la construction de la librairie
  - utilisation de « gènes de ménage » comme contrôle
- Ces méthodes restent très approximatives !

# Estimation du niveau d'expression des gènes avec RNA-seq

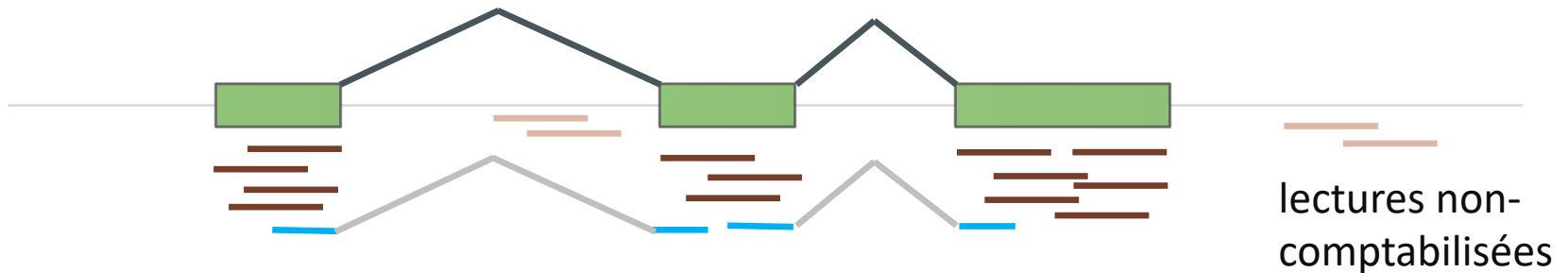


RNA-seq avec génome et annotation de référence :

- Alignement des lectures sur le génome avec un logiciel dédié (HISAT2, STAR, TopHat – identification de lectures alignées épissées)
- Comptage du **nombre de lectures** attribuées à chaque gène et/ou chaque isoforme (Rsubread, HTSeq, Cufflinks, Kallisto...)



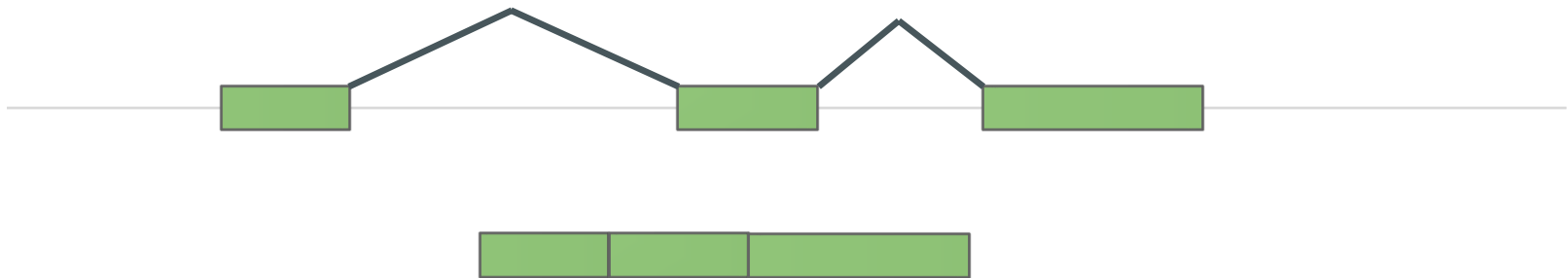
# Comment attribuer les lectures de RNA-seq aux gènes ?



- Analyse du chevauchement des alignements des lectures de RNA-seq et des coordonnées des **exons** des gènes
- Problèmes :
  - alignements ambigus (ou multiples)
  - chevauchement des gènes
- Solutions :
  - comptage des lectures alignées de manière non-ambiguë (unique)
  - attribution des lectures ambiguës proportionnellement aux nombres de lectures uniques

# Prise en compte de la longueur exonique des gènes

- Le protocole de RNA-seq comporte une étape de fragmentation ; on séquence des fragments d'ARN.
- Les molécules longues d'ARN vont être plus représentées dans la librairie de séquençage que les molécules courtes.
- Normalisation par la longueur **exonique** totale.





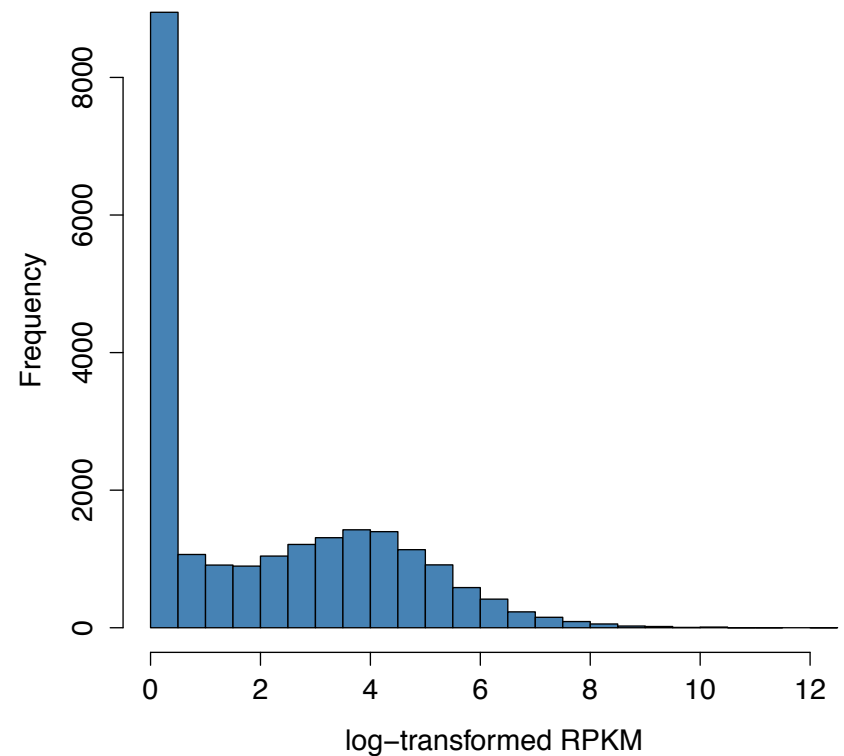
# Prise en compte de la profondeur de séquençage

---

- Le nombre de lectures varie d'un échantillon à un autre.
- Normalisation par le nombre total de lectures (après filtrage pour éliminer les lectures de mauvaise qualité, si besoin) ou par le nombre total de lectures alignées.

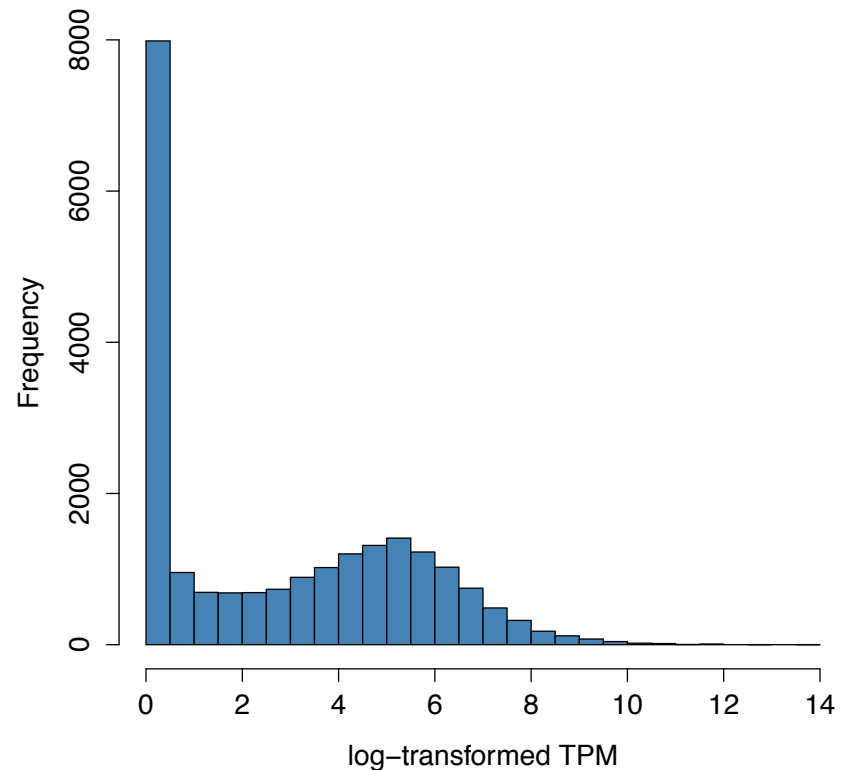
# Estimations du niveau d'expression : RPKM (FPKM)

- Première mesure du niveau d'expression : RPKM
- RPKM = **R**eads per **K**ilobase of exonic sequence per **M**illion mapped reads
- On normalise par la longueur exonique et par la profondeur de séquençage.
- Pour les représentations graphiques : transformations logarithmiques



# Estimations du niveau d'expression : TPM

- TPM = **T**ranscripts per **M**illion **T**ranscripts
- Moyen de calcul simple :
  - RPK = Reads per Kb of Exon
  - $TPM = RPK / (\text{somme RPK}) / 10^6$
- On ne normalise pas directement par la profondeur de séquençage
- La somme de tous les TPM d'un échantillon est constante ( $10^6$ )



# RPKM ou TPM : quelle mesure choisir ?

- Niveaux d'expression relatifs : la moyenne (ou la somme) doit être la même pour tous les échantillons.

| Species | Tissue/cell type              | Replicate | AvTPM  | AvRPKM |
|---------|-------------------------------|-----------|--------|--------|
| Human   | Differentiated decidual cells | 1         | 46.518 | 15.94  |
|         |                               | 2         | 46.518 | 16.13  |
| Human   | Un-differentiated dec. cells  | 1         | 46.518 | 15.27  |
|         |                               | 2         | 46.518 | 15.22  |
| Human   | Myofibroblast cells           | 1         | 46.518 | 17.66  |
|         |                               | 2         | 46.518 | 17.65  |
| Human   | Chondrocyte cells             | 1         | 46.518 | 16.57  |
|         |                               | 2         | 46.518 | 16.57  |
| Human   | Myometrial cells              | 1         | 46.518 | 17.77  |
|         |                               | 2         | 46.518 | 17.79  |
| Chicken | Forelimb digit 1 stage 28–29  | –         | 65.527 | 28.35  |
| Chicken | Forelimb digit 1 stage 31     | –         | 65.527 | 28.56  |

- Le TPM satisfait cette condition, pas le RPKM

# Seuil pour détection des gènes exprimés

---

- Présence de bruit transcriptionnel : comment dire si un gène est réellement exprimé ?
- Seuil proposé :
  - $\text{RPKM} > 1$
  - $\text{TPM} > 1$

# Logiciels pour la quantification de l'expression des gènes

---

- HTSeq (Anders et al., Bioinformatics, 2015)
  - calcule le nombre de lectures attribuées à chaque gène
  - essentiellement restreint aux lectures alignées de manière unique
- RSEM (Li & Dewey, BMC Bioinformatics, 2011)
  - peut utiliser un génome de référence ou un transcriptome de référence
  - comptages des lectures et valeurs TPM
- Rsubread (Liao et al., Nucleic Acids Res., 2019)
  - bibliothèque R
  - calcul très rapide pour l'alignement et la quantification

# Logiciels pour la quantification de l'expression des gènes

- Cufflinks (Trapnell et al., Nat Biotechnol, 2010) :
  - calcule les valeurs RPKM/FPKM par gène et isoforme, et leurs intervalles de confiances
  - correction pour alignement multiple des lectures
  - correction pour biais de séquençage/fragmentation dans RNA-seq
  - expression différentielle avec cuffdiff
  - assemblage de transcrits/gènes avec génome de référence
- Kallisto (Bray et al, Nat Biotechnol, 2016)
  - pseudo-alignement des lectures sur un transcriptome de référence
  - calcule les valeurs TPM des transcrits et des gènes
  - expression différentielle avec sleuth (en cours de développement)
  - très rapide !

# Logiciels pour la quantification de l'expression des gènes

---

- Pour les procaryotes : Rockhopper (McLure et al., Nucleic Acids Res, 2013)
  - alignement sur le génome
  - quantification de l'expression des gènes
  - expression différentielle
  - structure en opérons
  - visualisation des résultats



# Analyse de l'expression différentielle des gènes

---

- Expression différentielle : est-ce que le niveau d'expression d'un gène change significativement entre deux (ou plusieurs) conditions ?
- $H_0$  = il n'y a pas de différence entre conditions
- $H_1$  = il y a une différence significative entre conditions
- Pour construire le test, on utilise **les comptages bruts (nombre de lectures)** par gène, et non pas le RPKM ou TPM

# Logiciels pour l'analyse de l'expression différentielle

---

- Librairies (packages R) :
  - DESeq et DESeq2 (Anders et al, Genome Biology, 2010, Love et al., Genome Biology 2014)
  - edgeR (Robinson et al., Bioinformatics, 2010)
  - limma (Ritchie et al., Nucleic Acids Res, 2015)
- Pour l'analyse de l'épissage différentiel : DEXSeq (Anders et al., Genome Biology, 2012)
- On part toujours des comptages bruts, pas des valeurs RPKM/FPKM/TPM

# Comment dire si un gène est différentiellement exprimé ?

---

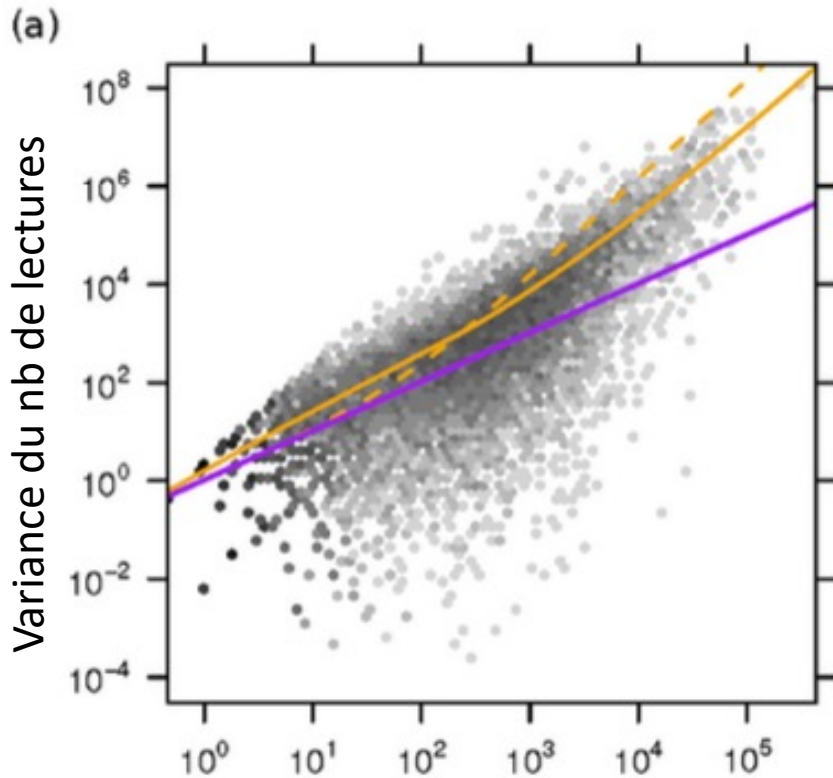
Données :

- Condition 1 : le gène est représenté par 100 lectures sur les 25,000,000 lectures alignées
- Condition 2 : le gène est représenté par 2500 lectures sur les 32,000,000 lectures alignées

Est-ce que les deux proportions sont différentes ?

Peut-on faire un simple test de Fisher ou Chi<sup>2</sup> ?

# Modélisation du nombre de lectures



Courbe violette : estimation faite en modélisant le nombre de lectures avec une loi de Poisson

- La loi de Poisson ne modélise pas suffisamment bien la variance : présence d'une sur-dispersion dans les données.

Moyenne du nb de lectures, entre réplicats d'une même condition

# Modélisation du nombre de lectures

- On modélise les comptages avec une loi binomiale négative :

$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i)$      $K_{ij}$  = nb de lectures pour le gène  $i$  et l'échantillon  $j$

$\mu_{ij} = s_j q_{ij}$      $\mu_{ij}$  = moyenne du nombre de lectures pour le gène  $i$   
et l'échantillon  $j$

$\alpha_i$  = paramètre de dispersion spécifique au gène  $i$

- La variance du nombre de lectures dépend donc de ce paramètre de dispersion :

$$\text{Var}(K_{ij}) = E[(K_{ij} - \mu_{ij})^2] = \mu_{ij} + \alpha_i \mu_{ij}^2$$

# Normalisation entre échantillons

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i)$$

$s_j$  = paramètre spécifique à l'échantillon  $j$  (size-factor)

$$\mu_{ij} = s_j q_{ij}$$

$S_j$  est estimé à partir de la médiane des nombres de lectures, sur tous les gènes dans l'échantillon  $j$

$S_j$  permet de normaliser pour la profondeur de séquençage

NB: il n'y a pas de normalisation pour la longueur exonique des gènes.

# Analyse de l'expression différentielle avec DESeq2

1. Estimation des coefficients de taille (size-factors)  $S_j$  pour chaque échantillon
  2. Estimation des paramètres de dispersion  $\alpha_i$  pour chaque gène
  3. Test statistique : test de Wald ou Likelihood Ratio Test (LRT)
- Pour estimer la variance (le paramètre  $\alpha_i$ ) il faut des réplicats biologiques. Plus on a de réplicats, mieux c'est !

# Comment sélectionner les gènes différentiellement exprimés ?

---

Il y a deux paramètres importants :

- FDR (ou p-valeur ajustée) : proportion de faux positifs dans la liste de résultats.
- Log2 fold change : intensité du changement d'expression. Il représente le ratio (log2-transformé) du niveau d'expression des gènes entre deux conditions.

**La correction pour tests multiples est absolument impérative !**

On peut faire > 20,000 tests (un par gène) dans une seule analyse.



# Comment interpréter les résultats ?

Analyse dans de l'expression différentielle dans une population de cellules (e.g., tissu disséqué *ex-vivo*, culture cellulaire, culture de micro-organismes) :

- On détecte des effets moyens au niveau de la population.
- Le gène peut changer d'expression dans toutes les cellules ou dans un sous-ensemble des cellules;
- La composition cellulaire de la population peut changer entre deux conditions (e.g., infiltration de cellules immunitaires etc.)

Quel est le mécanisme moléculaire ?

- Régulation transcriptionnelle
- Régulation post-transcriptionnelle (par ex., épissage alternatif)

# Enrichissement en catégories fonctionnelles

---

Est-ce que les gènes différentiellement exprimés sont impliqués dans un processus biologique particulier ?

Enrichissement en catégories fonctionnelles (annotation Gene Ontology, KEGG, etc.)

Principe :

- Chaque gène est associé avec 1 ou plusieurs catégories fonctionnelles.
- Pour chaque catégorie fonctionnelle, on teste si elle est plus fréquente parmi les gènes d'intérêt que parmi tous les gènes du génome (ou parmi une autre liste de référence)

# Enrichissement en catégories fonctionnelles

- $N$  gènes dans le jeu de référence
- $K$  gènes associés avec la catégorie fonctionnelle dans le jeu de référence
- $n$  gènes dans le jeu de gènes d'intérêt (différentiellement exprimés)
- $k$  gènes associés avec la catégorie fonctionnelle dans le jeu de gènes d'intérêt

Si pas d'enrichissement : distribution hypergéométrique.

P-valeur : la probabilité d'observer  $k$  ou plus gènes associés avec cette catégorie fonctionnelle, étant donnés  $N$ ,  $K$  et  $n$ .

Toujours faire la correction pour tests multiples !

# Logiciels pour l'analyse fonctionnelle

---

- Gorilla (<http://cbl-gorilla.cs.technion.ac.il>)
- DAVID (<https://david.ncifcrf.gov> )
  
- Librairie R GOfuncR
- Librairie R GOstats