Conférences Axe RSME

Méthodes statistiques pour l'analyse de données génomiques

Loïc Mangnier, PhD

Biostaticien Arnaud Droit Lab

loic.mangnier@gmail.com

https://statsxomics.blog/

github.com/lmangnier

"Statistics without science is incomplete, science without statistics is imperfect." K.V Mardia

À QUOI SERT LA STATISTIQUE EN RECHERCHE ?

À quoi sert la statistique en recherche ?

Constat sans appel: La statistique est indispensable pour publier aujourd'hui

- Inférence vs Prédiction
  - ***Inférence***: Déterminer l'association entre une ou plusieurs variables $\rightarrow$ Significativité statistique (Conférence 1)
  - ***Prédiction***: Prédire de nouvelles observations sur la base de données actuelles$\rightarrow$ Machine Learning (Conférence 2)
- Statistique fréquentiste vs Statistique Bayésienne
  - ***Statistique fréquentiste***: Larges Tailles d'échantillon
  - ***Statistique Bayésienne***: Notion d'à-priori

Concepts centraux liés à l'inférence

Tests statistiques vs modèles de régression

- ***Tests statistiques***: Cadres **simples** pour tester des hypothèses **simples** $\rightarrow$ ne permet pas d'ajustement pour certains facteurs de confusion et de plans expérimentaux complexes.
- ***Modèles de régression***: Cadres **plus complexes** pour tester des hypothèses **plus complexes** $\rightarrow$ permet l'ajustement pour la présence de facteurs de confusions et plans expérimentaux complexes.

Significativité statistique: Valeur-p et intervalles de confiance

Concepts centraux liés à l'inférence

Tests statistiques paramétriques usuels

- *t-test/ANOVA*: Comparaisons de moyennes pour données normales
- Tests statistiques non-paramétriques usuels
- *Test de Wilcoxon/Mann-Whitney*: Test de comparaisons de "médianes" pour tous types de données
- *Test exact de Fisher/Chi2*: Test de comparaisons de fréquences pour tableaux de contingence

Tests paramétriques fonctionnent bien même avec un N petit mais font des hypothèses fortes sur la distribution des données

Tests non-paramétriques fonctionnent mieux pour un N grand mais ne font pas d'hypothèses sur la distribution des données

Concepts centraux liés à l'inférence

Modèles de régression usuels

- *Régression linéaire*: Étude d'association entre prédicteurs et une variable réponse **continue**
- *Régression Logistique*: Étude d'association entre prédicteurs et une variable réponse **binaire**
- *Régression de Poisson/Binomiale négative*: Étude d'association entre prédicteurs pour données de **comptages** avec ou sans présence de **surdispersion**

Ces modèles peuvent être étendus à des devis longitudinaux (modèles mixtes) et ont leur équivalent semi- ou non-paramétriques (GEE ou GAM).

Au final, les modèles de régression sont à privilégier au profit des tests statistiques standards

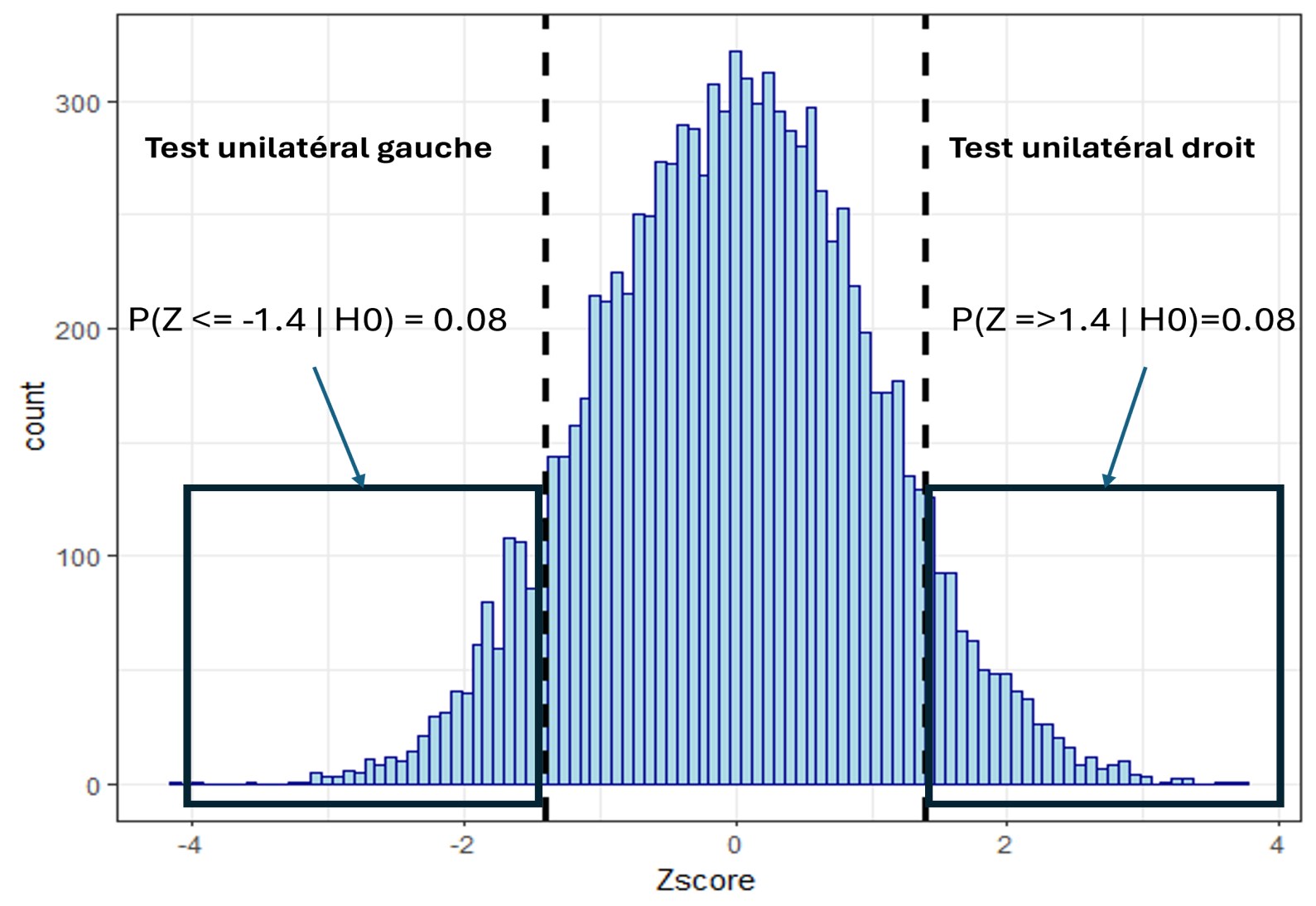
C'est quoi la significativité statistique ?

Concepts centraux liés à l'inférence

Valeur-p et intervalle de confiance

- ***valeur-p***: Probabilité d'observer un résultat au moins aussi extrême qu'**attendu** sous l'**hypothèse nulle**
- ***Intervalle de confiance***: Intervalle contient la vraie valeur du paramètre avec un certain niveau de confiance. Un intervalle de confiance à 95% contient la vraie valeur 95% du temps.

Valeur-p: approche graphique



À retenir

La valeur-p n'est pas:

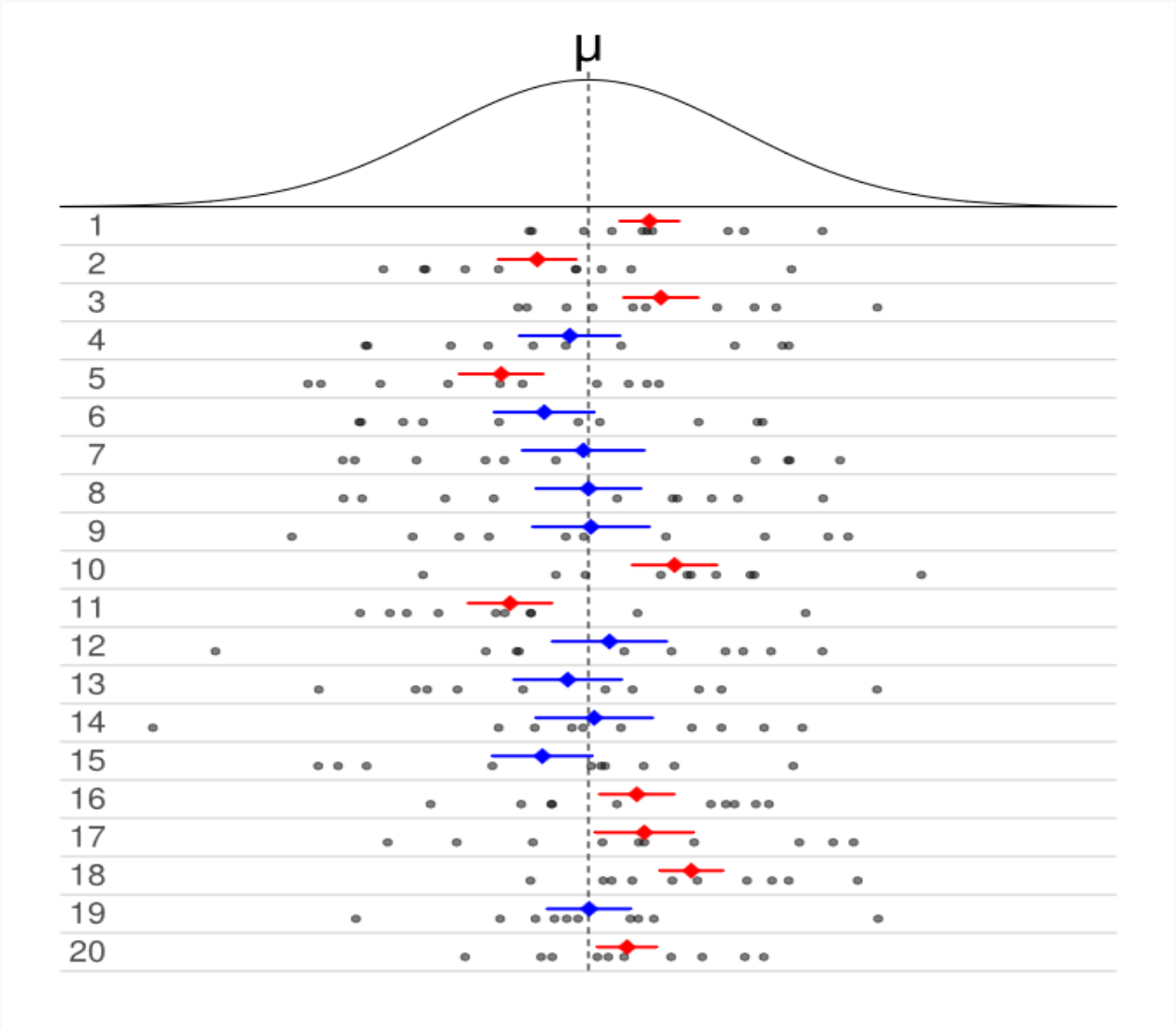
- La probabilité qu'une hypothèse soit vraie $\rightarrow$ peut être obtenue par une approche bayésienne
- La probabilité que le résultat soit obtenue par la chance seulement

La valeur-p est:

- Un degré d'**adéquation** entre les **données** et une **hypothèse**: Plus la valeur-p est petite plus le résultat est improbable sous l'hypothèse d'intérêt

Aussi définir un seuil de significativité à 5% signifie que l'on s'attend **maximalement** à 5% de faux-positifs si H0 est vraie

Intervalle de confiance: approche intuitive



[°] https://en.wikipedia.org/wiki/Confidence-interval

Pourquoi faire de l'inférence en biologie ?

Pour mesurer la variabilité d'une association: distinguer le signal du bruit biologique

- **Exemple 1**: Analyse d'Expression différentielle pour données de RNA-Seq
- **Exemple 2**: Analyse d'enrichissement différentielle pour données de chiP-Seq

Analyse d'expression différentielle

Structure des données de transcriptomiques:

- **Défi 1:** Données de comptage avec présence de surdispersion: **Loi binomiale négative**
- **Défi 2:** Souvent un nombre faible de réplicats

Outil de référence: DESeq2

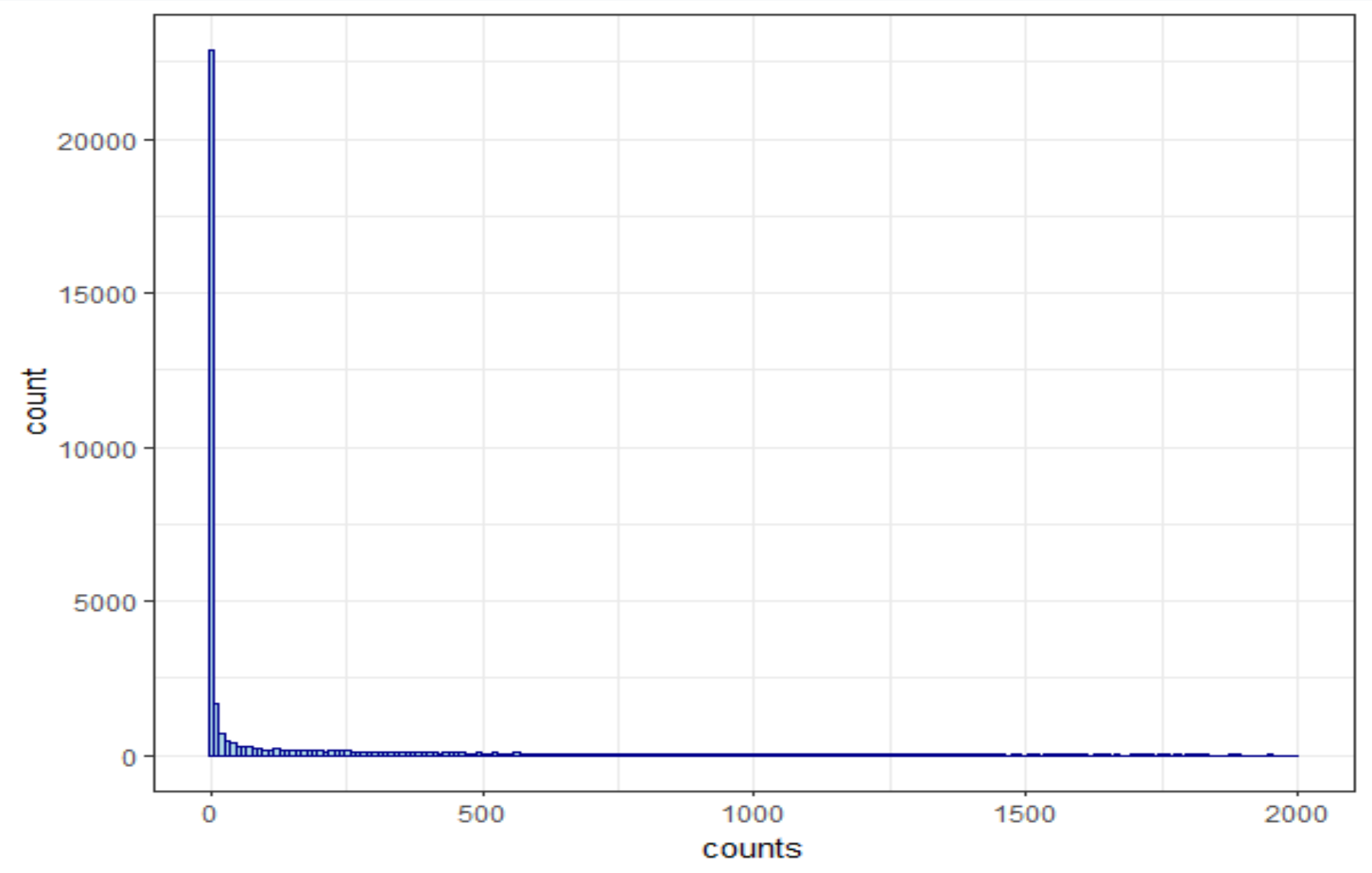
Analyse d'expression différentielle

Structure des données de transcriptomiques:

	Rep1	Rep2	Rep3	Rep1	Rep2	Rep3
Gene1	0	1	10	0	4	6
Gene2	18	2	2	3	3	9
Gene3	6	0	12	7	7	4
Gene4	2	12	19	1	0	8
Gene5	7	2	2	8	9	0
Gene6	12	0	2	3	52	12

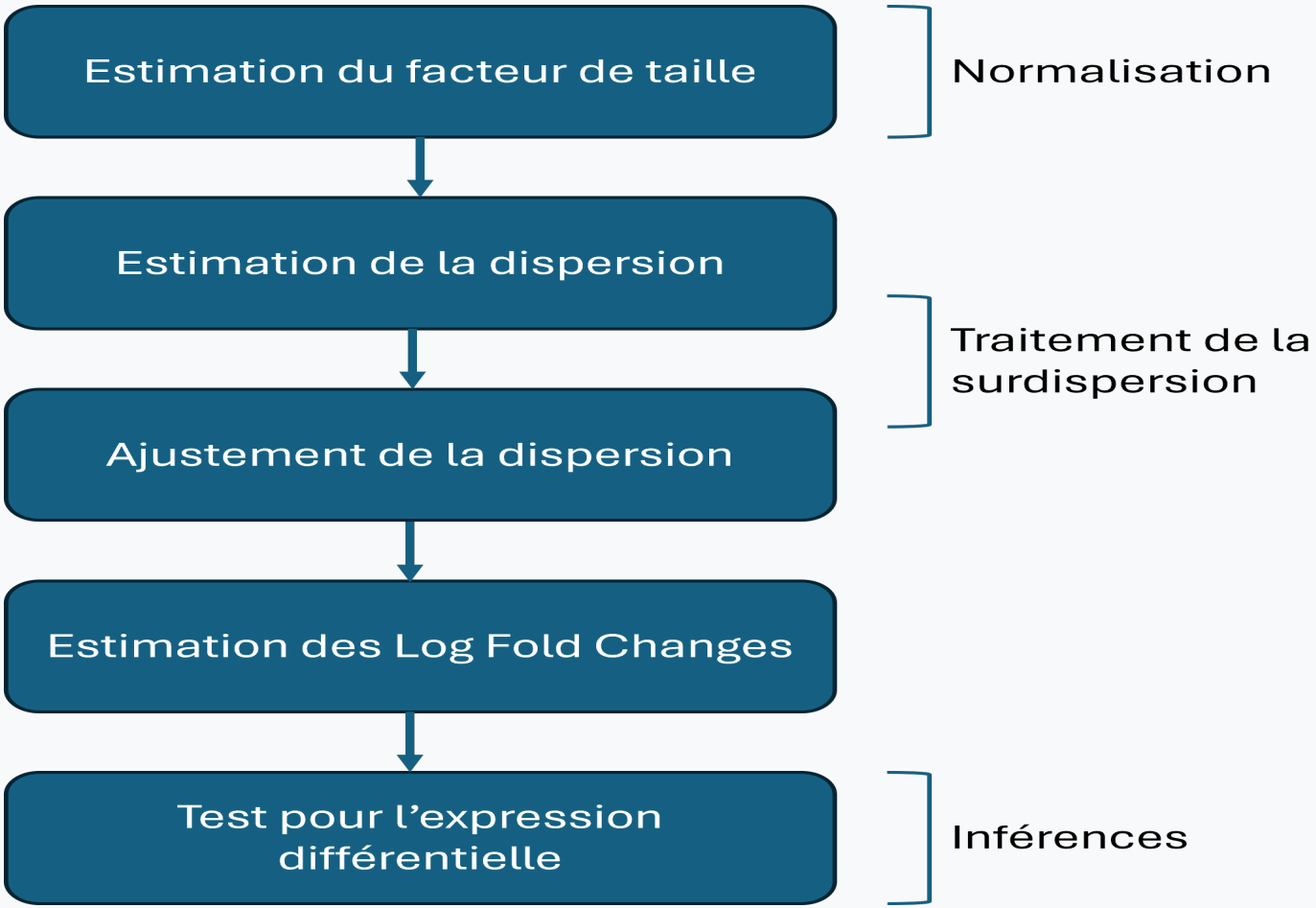
Table: Exemple d'une table de comptage pour données de RNA-Seq

Structure des données

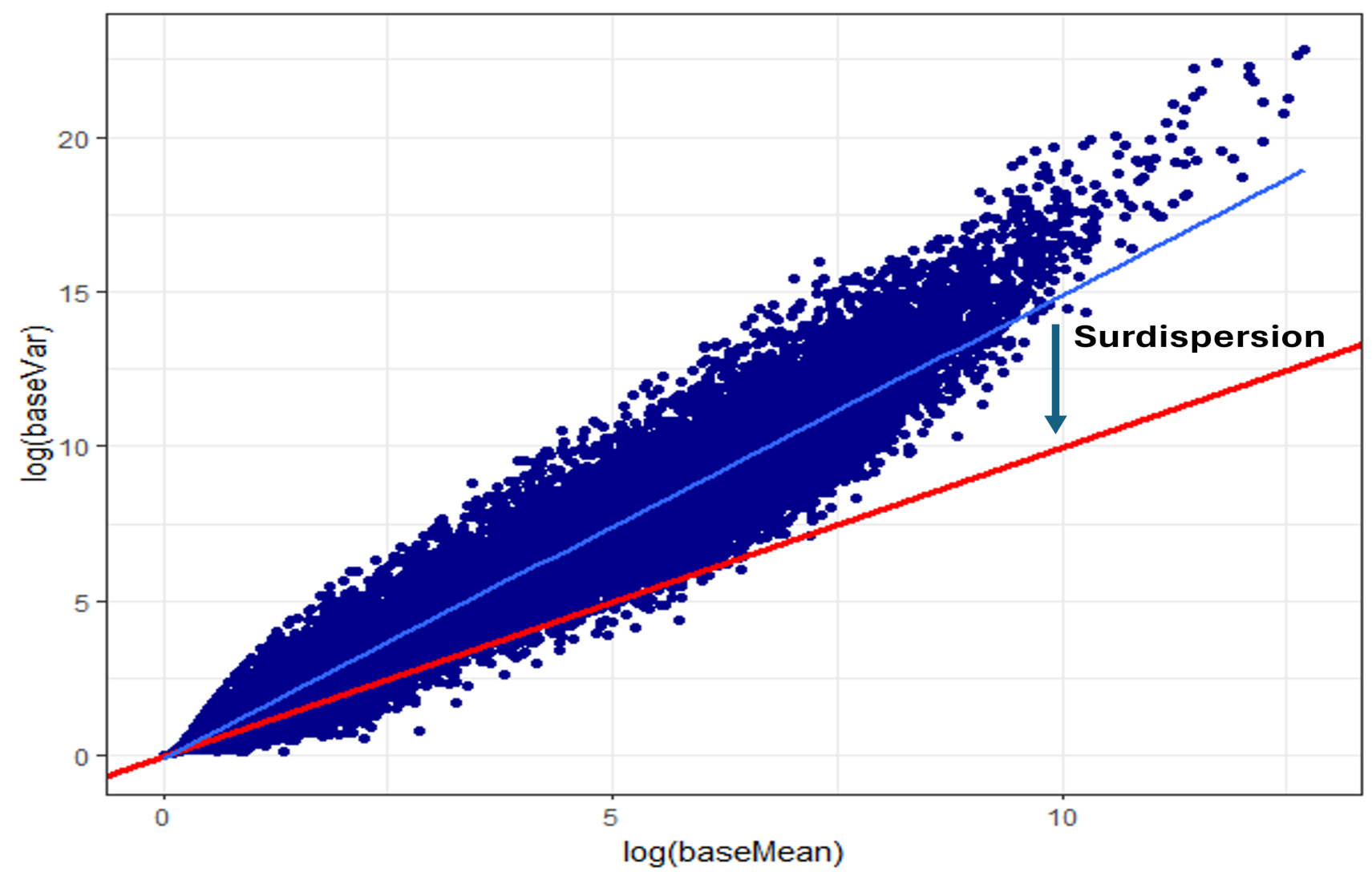


Analyse d'expression différentielle

DESeq 2: vue d'ensemble



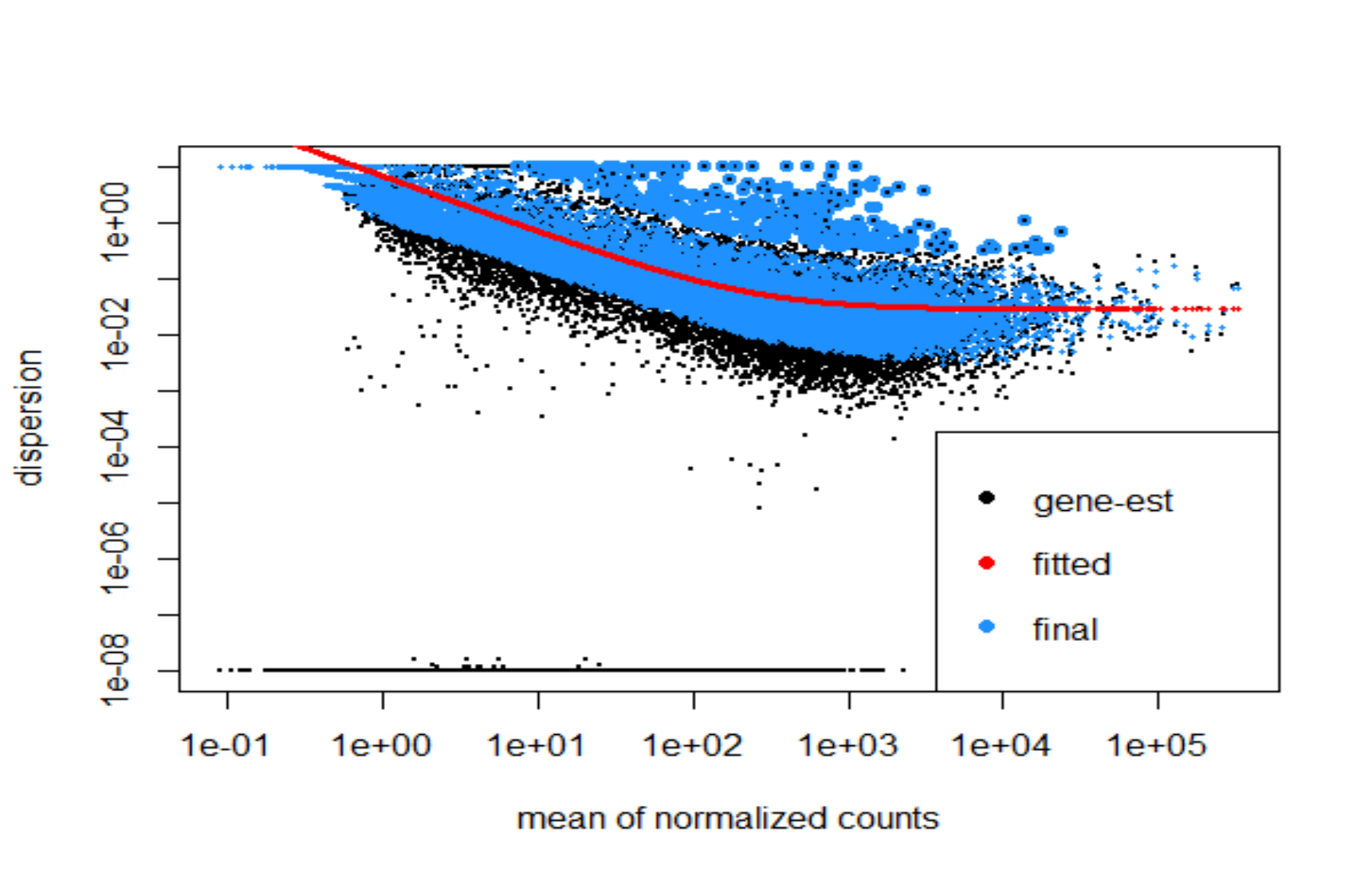
Surdispersion



Traitement de la Surdispersion

$$Var(K_{ij}) = \mu_{ij} + \mu_{ij}^2 \alpha_{ij}$$

Traitement de la Surdispersion



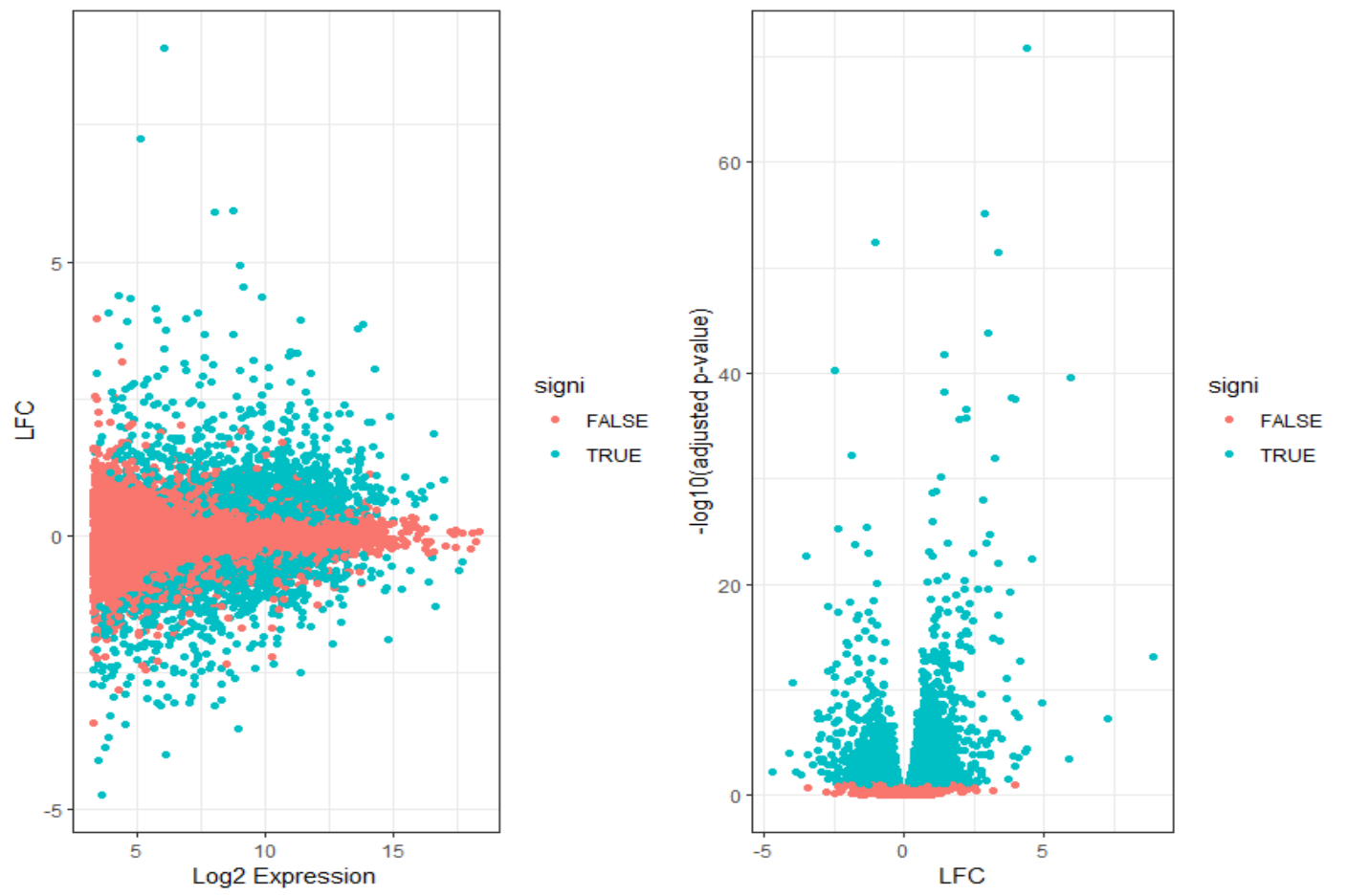
Analyse d'expression différentielle

Interprétation des résultats

- Log-Fold Change: donne la variation d'un gène par le fait de passer de la condition de référence à la condition d'intérêt
- Si le Log-Fold Change < 0 alors l'expression du gène diminue (downregulated)
- Si le Log-Fold Change > 0 alors l'expression du gène augmente (upregulated)
- Par défaut l'utilisation du shrinkage est conseillé $\rightarrow$ Correction supplémentaire pour la surdispersion + présence d'un faible nombre de réplicats: Favorise la réplicabilité et comparabilité des résultats

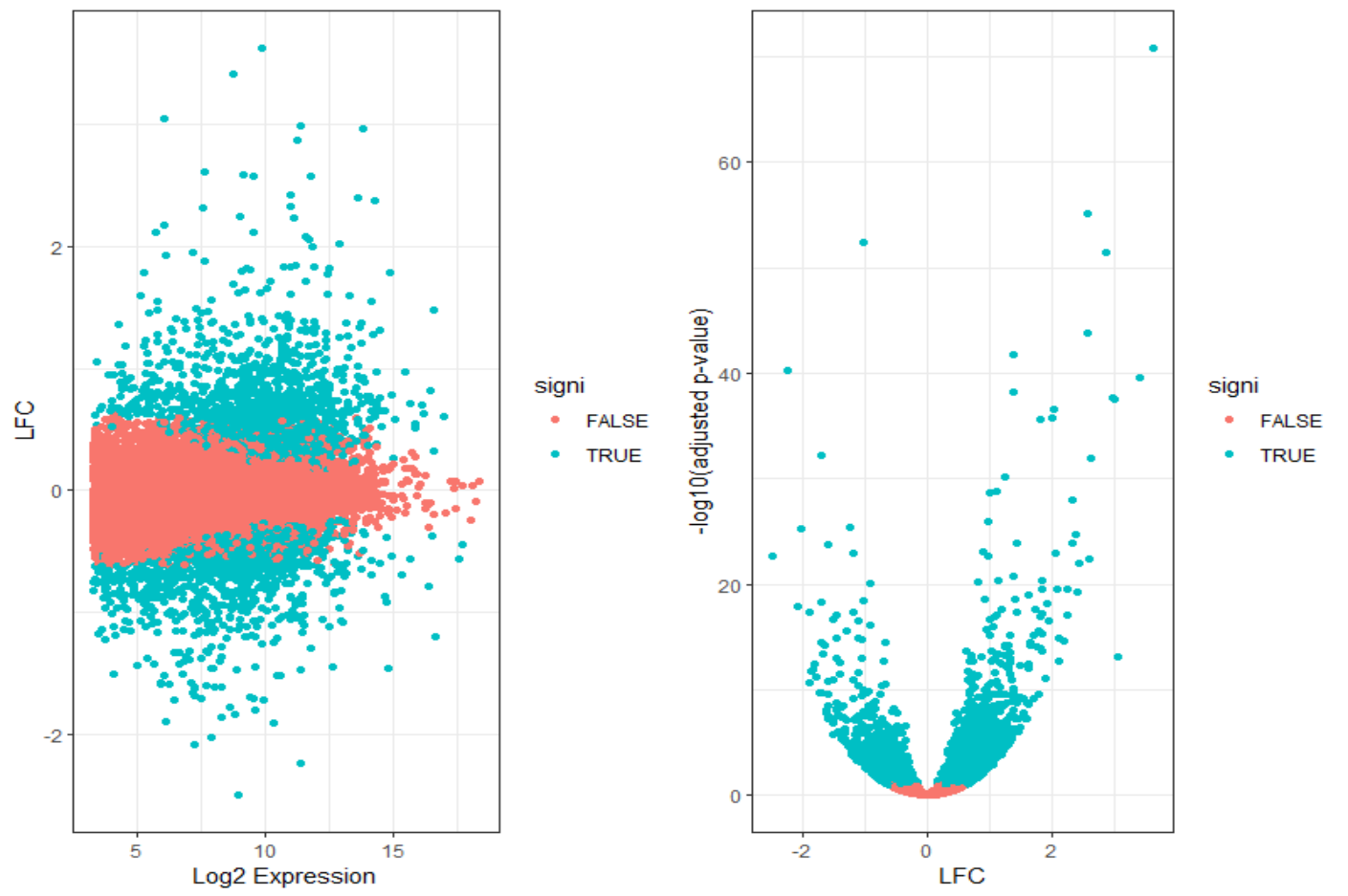
Analyse d'expression différentielle

Interprétations des résultats: Approche non-pénalisée



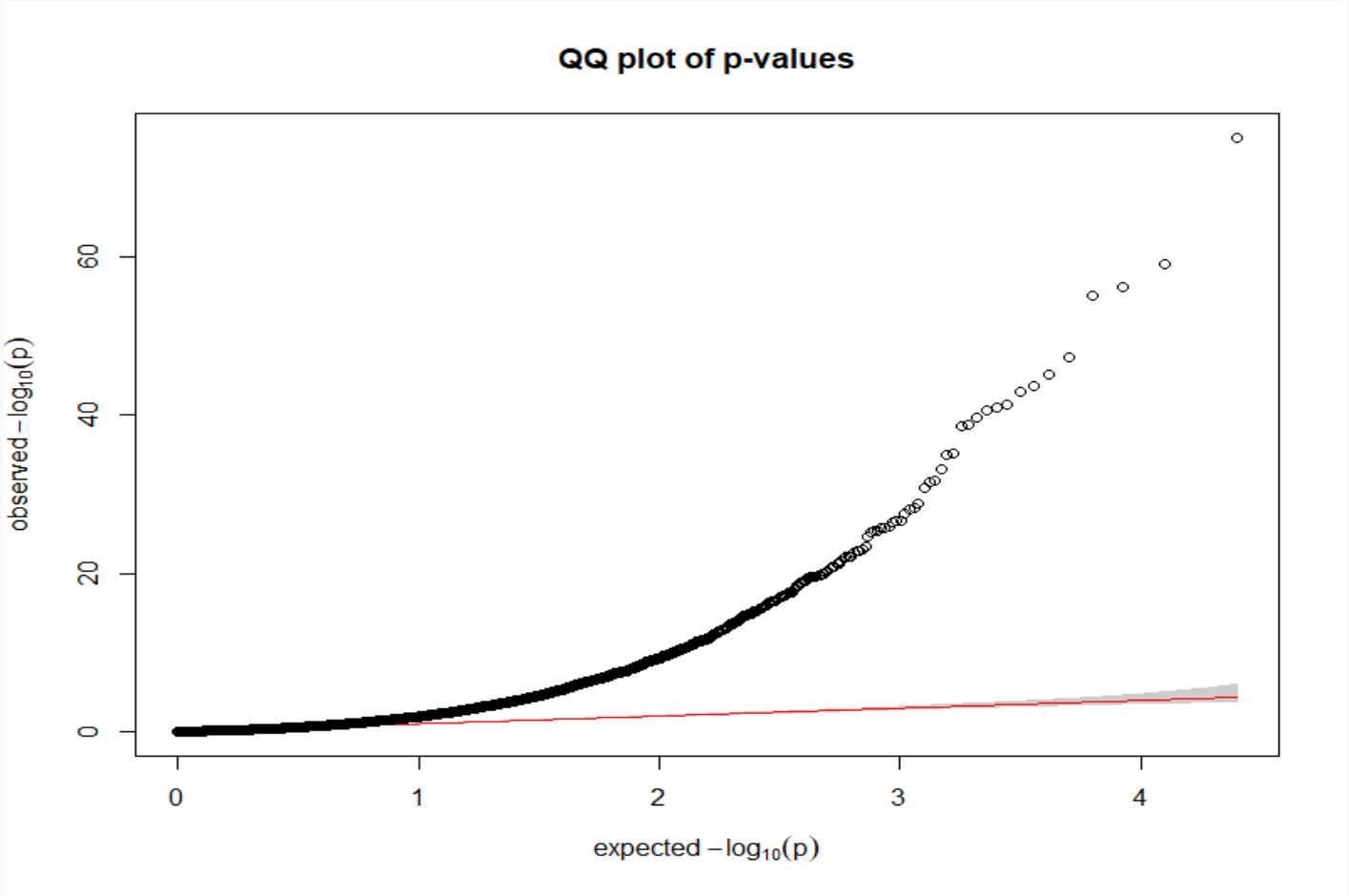
Analyse d'expression différentielle

Interprétations des résultats: Approche pénalisée



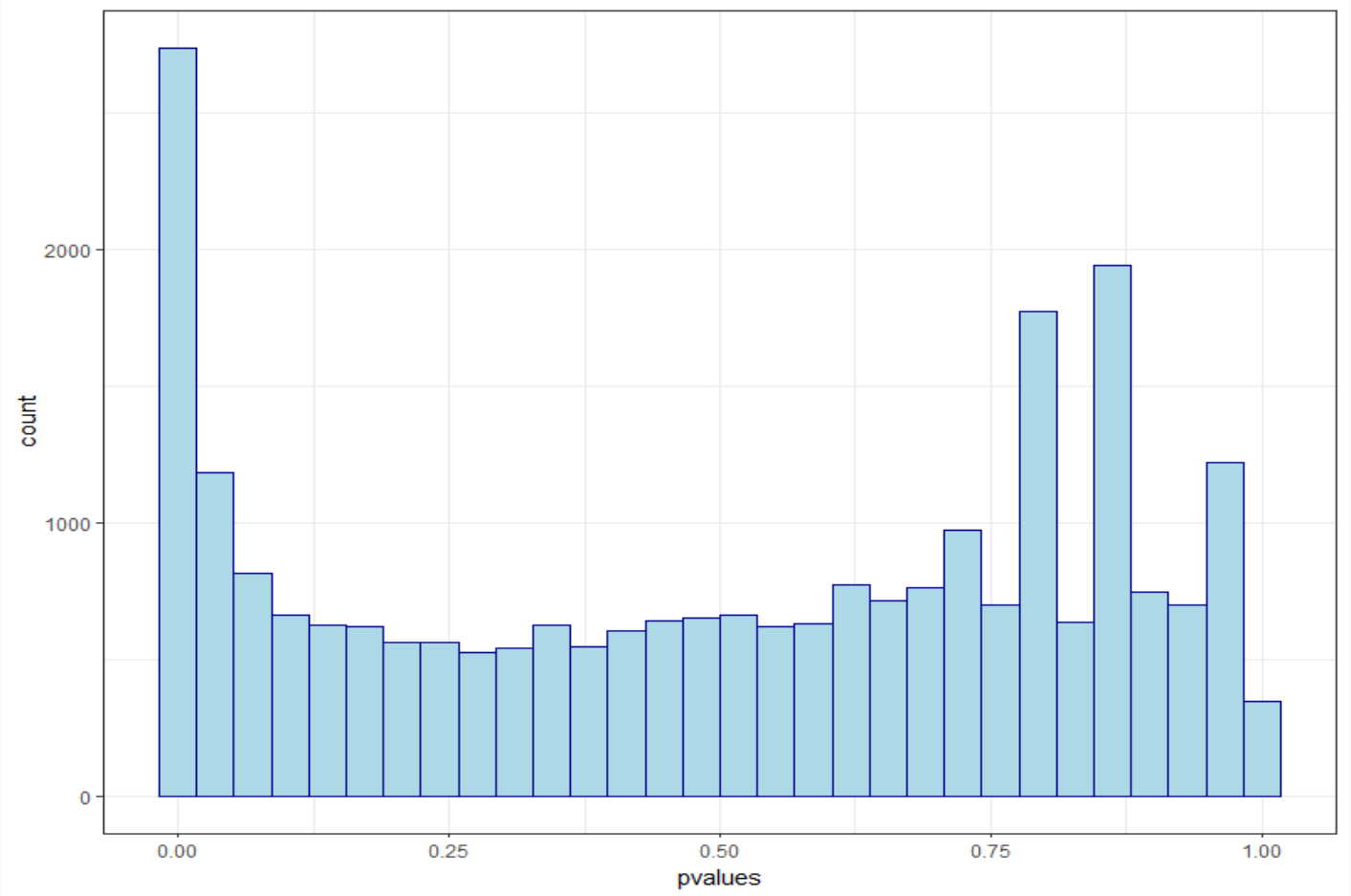
Analyse d'expression différentielle

Validité des résultats: QQPlot



Analyse d'expression différentielle

Validité des résultats: Histogramme



Remèdes à l'inflation des valeurs-p

- Descriptif ? Mécanistique ?
- Emploi d'un modèle plus approprié: limma, edgeR, ou tout autre modèle pertinent en lien avec la structure sous-jacente des données
- D'autres orientations peuvent êtres employées: analyses descriptives
- La correction pour les tests multiples n'est justifiée que dans le cadre où les valeurs-p non ajustées produisent le comportement attendu

À retenir

- DESeq2 traite explicitement la surdispersion par l'emploi d'un modèle binomial négatif
- La question de la variabilité induite par les faibles comptes est traitée par (1) une estimation de la dispersion utilisant tous les gènes (2) une approche pénalisée
- Avant de se lancer dans l'analyse des résultats: Vérification des qqplots et histogrammes de valeurs-p
- Le choix de la visualisation et de la méthode de correction pour les tests multiples va dépendre des objectifs de recherche.

Remarques générales

- Le modèle présenté ici est valide pour d'autres données de comptages: ChiP-Seq par exemple
- Le modèle ne traite pas l'inflation de zéros: À éviter dans le cas de données de single-cell ou de conformation de la chromatine (Hi-C)
- Le modèle n'est pas valide pour les données de métagénomique

Pas d'approche systématique. Comprenez vos données avant tout !

Merci pour votre attention !
Questions ? Commentaires ?

Suivez moi sur mes réseaux sociaux