Expression différentielle avec SARTools

Matthias Zytnicki, Nathalie Villa-Vialaneix
avec l'appui des supports de H. Varet et I. González

MIAT INRA

31 janvier 2018

Étapes

- question biologique
- plan expérimental* †
- extraction
- préparation des librairies
- séquençage
- filtrage et quantification*
- analyse* †
- validation et interprétation*

- * : besoins en stat
- † : évoqué ici

Plan

1. Plan expérimental
2. Travaux pratiques
3. Exploration des données
4. Normalisation
5. Test

Plan expérimental

But

Protocole répondant à une question donnée.

En gros

Voir si, dans deux conditions différentes (un facteur), l'expression d'un gène change (estimer la variabilité).

Moyen

- Estimer la variabilité de votre facteur.
- Contrôler la variabilité des autres facteurs.

Fichier de design

id	cond
WT1	WT
WT2	WT
WT3	WT
KO1	KO
KO2	KO
KO3	KO

Fichier de design

id	cond	day
WT1	WT	1
WT2	WT	2
WT3	WT	3
KO1	KO	1
KO2	KO	2
KO3	KO	3

Fichier de design

id	cond	day	tech.
WT1	WT	1	Alice
WT2	WT	2	Bob
WT3	WT	3	Alice
KO1	KO	1	Bob
KO2	KO	2	Alice
KO3	KO	3	Bob

Ce que l'on ne fera pas ici

Données pairées

patient	tissue
1	normal
1	cancer
2	normal
2	cancer
3	normal
3	cancer

Données temporelles

patient	heure
1	0h
1	24h
1	48h
2	0h
2	24h
2	48h
3	0h
3	24h
3	48h

Données complexes (+ rép. !)

souche	stress
A	Y
A	N
B	Y
B	N

Ce que l'on ne fera jamais

Facteurs confondants

cond	âge	tech.	lane
sain	21	Alice	1
sain	23	Alice	1
sain	22	Alice	1
sain	24	Alice	1
atteint	54	Bob	2
atteint	52	Bob	2
atteint	53	Bob	2
atteint	51	Bob	2

Variabilité technique

Importance des effets

- Effet lane $<$ effet flowcell $<$ effet run $\ll$ effet biologique
- Le multiplexing permet toutefois de réduire les effets.
- L'effet librairie et multiplexing sont également très importants.

Réplicats

Pourquoi

- Le but est d'estimer la variabilité. Elle doit donc être estimée.
- « Pooler » les données ne permet d'estimer la variabilité.

Réplicats biologiques vs techniques

- réplicats techniques : plusieurs mesures d'un même échantillon
    - plusieurs extractions d'un même échantillon
    - plusieurs séquençages d'une même librairie
- réplicats biologiques : plusieurs mesures de la variabilité observée

Conclusion

- Ne vous lancez pas dans un plan que vous ne pourrez pas analyser.
- Parfois, plus simple, c'est mieux.
- Pensez comme un statisticien :
  - réplication : permet d'estimer les variabilités ;
  - randomisation : diminue les biais non pris en compte dans le design ;
  - blocking : lorsque la randomisation ne permet pas de prendre en compte un effet connu, on ajoute un facteur dans le design.

Plan

1. Plan expérimental
2. Travaux pratiques
3. Exploration des données
4. Normalisation
5. Test

Étape 1 — Récupérer les données

But

Récupérer les comptages pour 3 WT et 3 MT sur le génome complet.

Comment faire ?

- Ouvrir un nouvel historique.
- Allez dans `Shared Data > Published histories`
- Trouvez `TP RNA-Seq -- SARTools`
- Sélectionnez `Import history`
- Choisissez le nom d'un nouvel historique qui va bien.

Étape 2 — Regrouper les données

But

Créer la matrice de comptage et le plan expérimental.

Comment faire ?

Il s'agit d'entrer chaque séquençage un à un.

- Entrez un nom pour le premier groupe (de référence) : `WT`
- Sélectionnez le premier fichier de réplicat : `wt1.txt`
- Donnez-lui un identifiant : `WT1`
- De même pour les autres réplicats/groupe. Ajoutez des réplicats quand nécessaire.
- Go !

Attention

- Souvenez-vous du nom de groupe de référence.
- Chaque identifiant doit être unique.

Étape 3 — Lancer l'analyse

Comment faire ?

Seuls champs importants :

- `Design/target file` : le fichier de design généré précédemment.
- `Zip file containing raw counts files` : les comptages générés précédemment.
- `Reference biological condition` : le nom du groupe de référence.

Plan

1. Plan expérimental
2. Travaux pratiques
3. Exploration des données
4. Normalisation
5. Test

Pourquoi ?

Une étape obligatoire

- Une bonne partie des erreurs/biais peut se voir dès cette étape.
- On gagne du temps à chercher d'abord les problèmes.
- Les analyses stats ont des suppositions fortes sur les distributions des comptages, qu'il faut vérifier *a priori*.

Une analyse exploratoire prend plus de temps que l'analyse des données, mais elle est nécessaire !

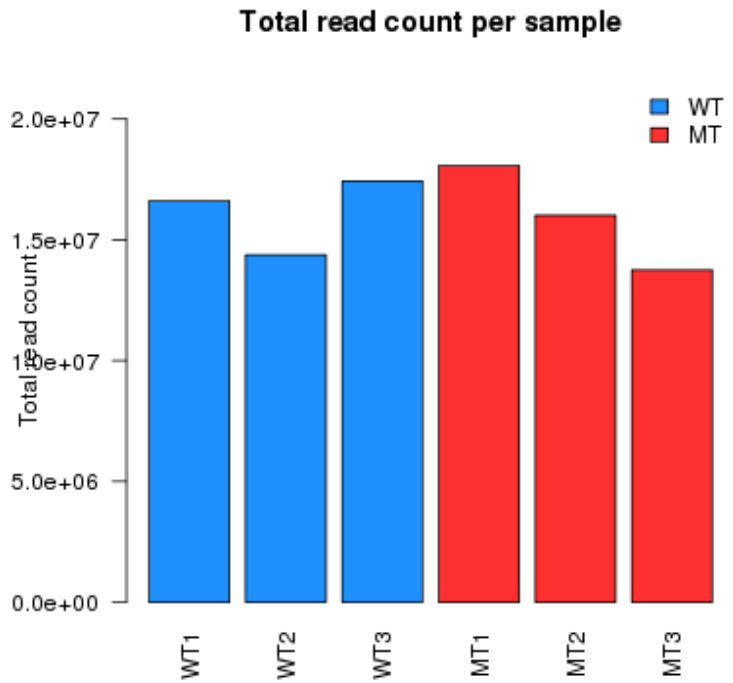
Taille de librairie

Definition (Taille de librairie / *library size*)

- Nombre de lectures brut / normalisé d'un échantillon.
- Ici, le nombre de lectures sur les gènes.

But

Détecter des déséquilibres dans la profondeur de séquençage.

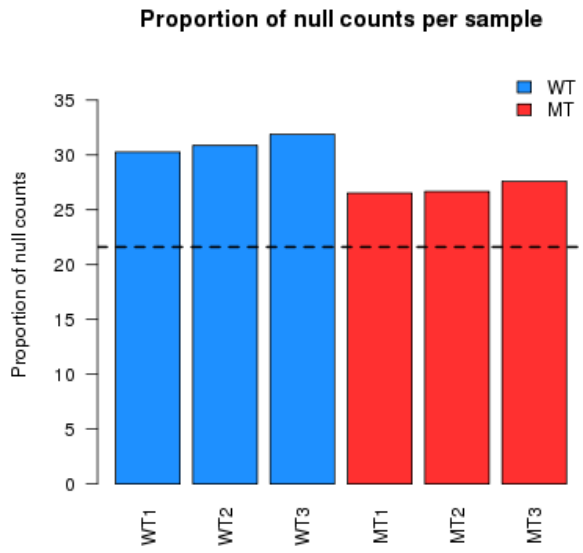


Nombre de gènes non-exprimés

(Ici, avec 0 lecture.)

But

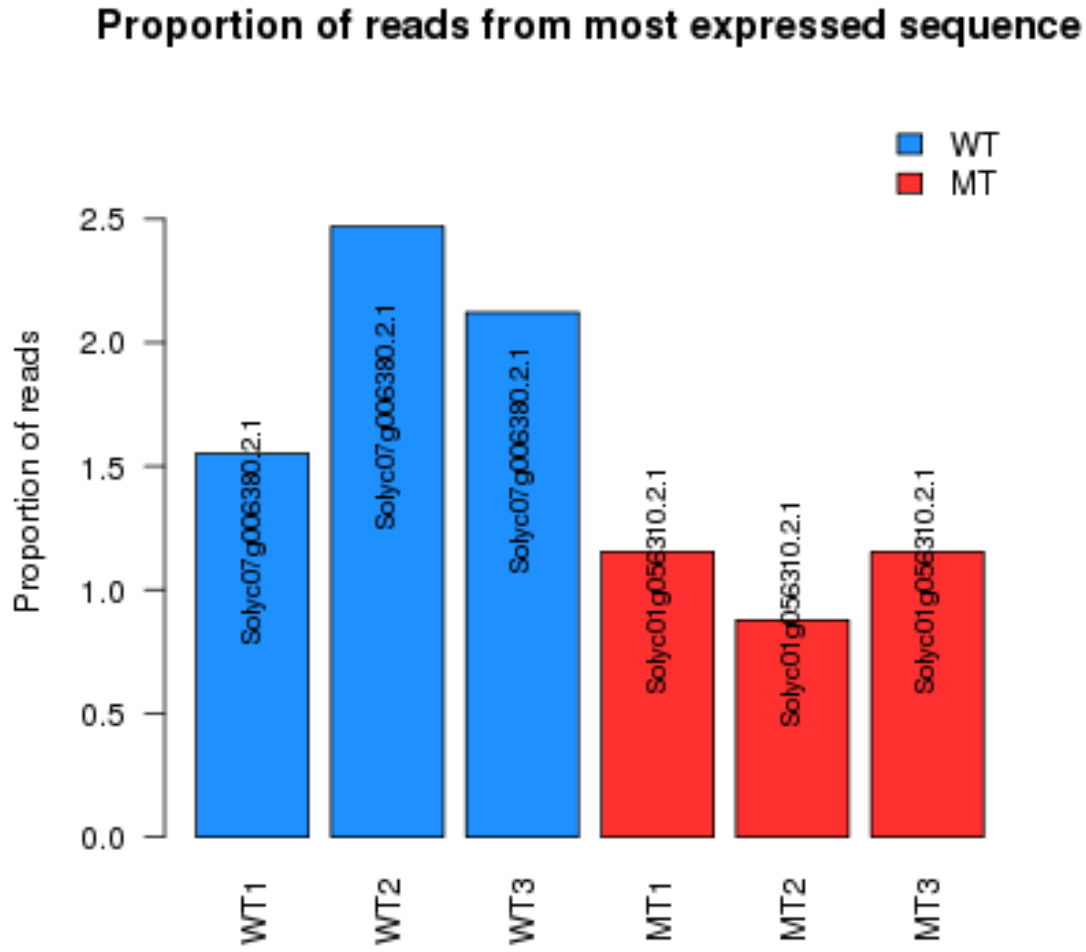
Vérifier la distribution homogène des lectures. Détecter des biais d'amplification.



Proportion de lectures venant du gène le plus exprimé

But

Détecter une contamination ou un enrichissement problématique.



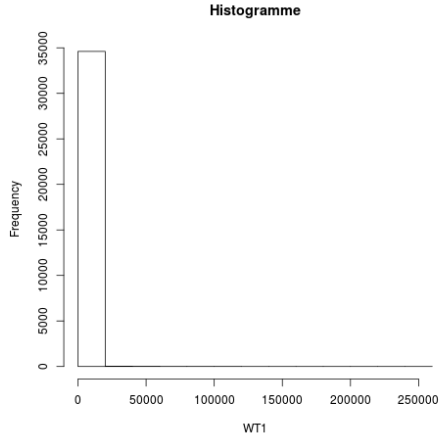
Densité en nombre de comptages

But

Vérifier une distribution homogène des lectures. Détecter des biais d'amplification.

Visualisation

- Classique : un histogramme.



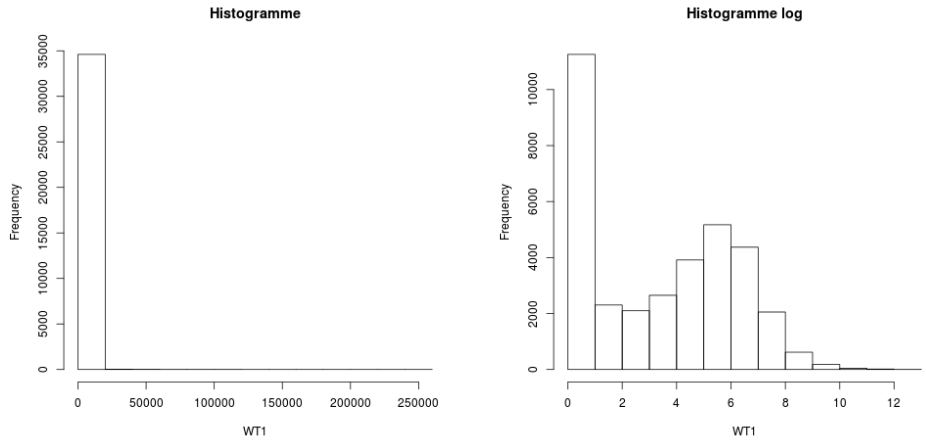
Densité en nombre de comptages

But

Vérifier une distribution homogène des lectures. Détecter des biais d'amplification.

Visualisation

- Classique : un histogramme.
- Pour mieux visualiser, on *transforme* la distribution (en $\log_2$).



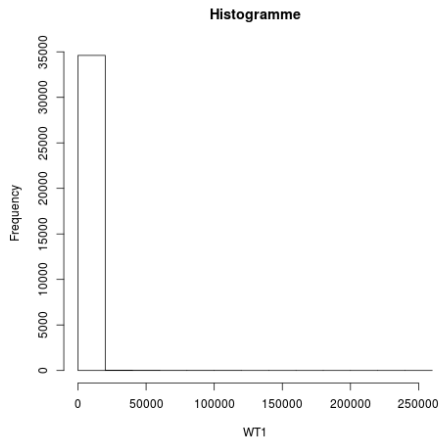
Densité en nombre de comptages

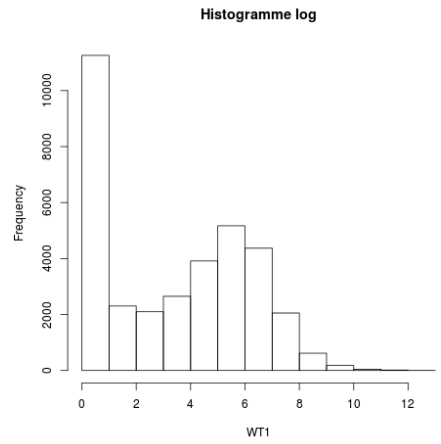
But

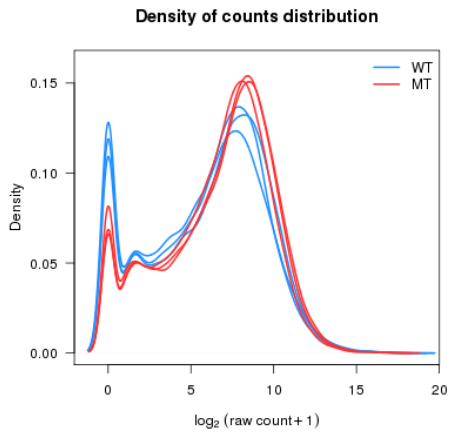
Vérifier une distribution homogène des lectures. Détecter des biais d'amplification.

Visualisation

- Classique : un histogramme.
- Pour mieux visualiser, on *transforme* la distribution (en $\log_2$).
- Plus joli (?) on donne la densité (ici, échelle log/log).







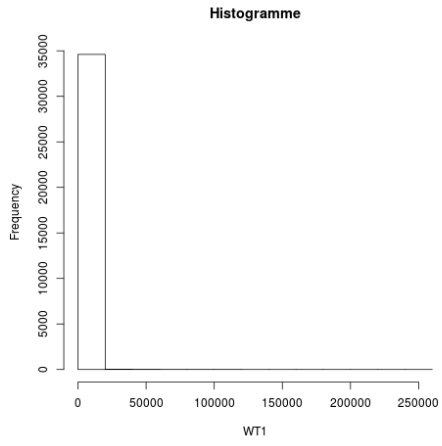
Densité en nombre de comptages

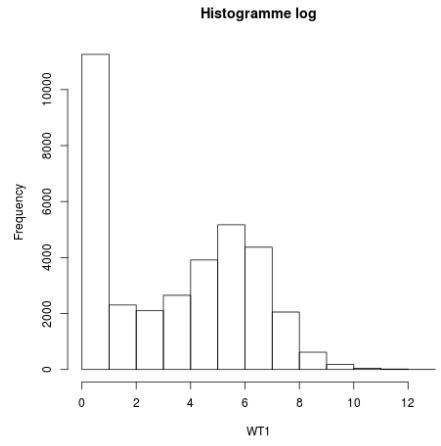
But

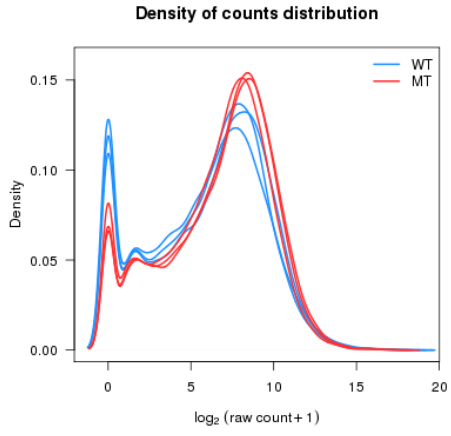
Vérifier une distribution homogène des lectures. Détecter des biais d'amplification.

Visualisation

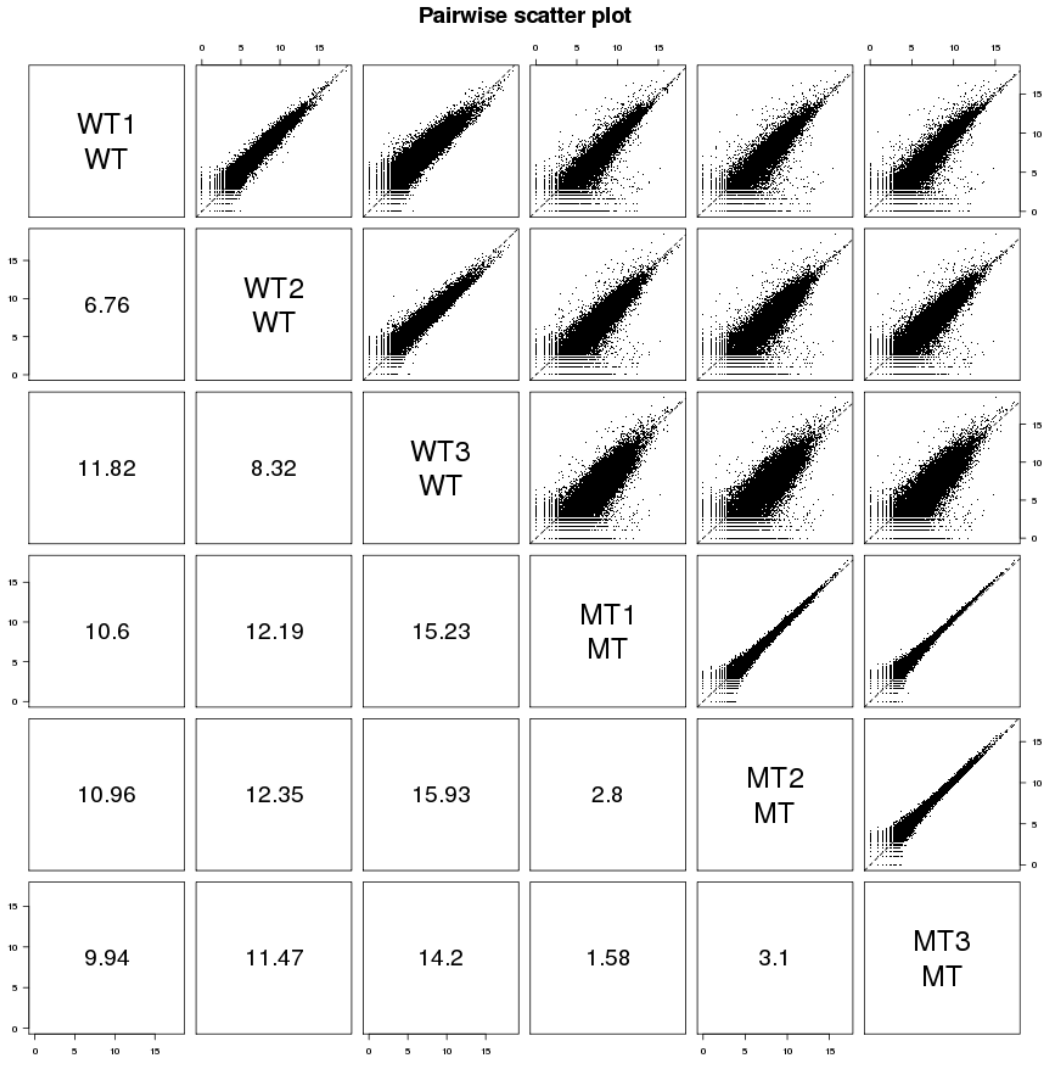
- Classique : un histogramme.
- Pour mieux visualiser, on *transforme* la distribution (en $\log_2$).
- Plus joli (?) on donne la densité (ici, échelle log/log).
- On peut également utiliser des boîtes à moustaches.







Pairwise scatter plot



Pairwise scatter plot

But

Vérifier les corrélations d'échantillons attendues.

Comment le lire ?

- Rectangles de la diagonale supérieure-droite :
  - un carré est une comparaison de 2 échantillons A vs B,
  - un point est un gène,
  - absisse : $\log_2$(comptages dans A $+1$),
  - ordonnée : $\log_2$(comptages dans B $+1$).
- Rectangles de la diagonale inférieure-gauche : coefficients de SERE :
  - 0 : identique
  - ≈ 1 : réplicat technique
  - > 1 : réplicats biologiques
  - $\gg 1$: autre condition

Coefficients SERE

Proposé dans Schulze *et al.*, BMC Genomics. 2012.

- Avantages :
  - pas sensible à la taille de la librairie,
  - détecte des réplicats « trop semblables »,
  - plus adapté que les coefficients de Pearson ou Spearman pour le RNA-Seq.
- Inconvénients :
  - sensible aux forts comptages (rRNA),
  - pas facile d'interpréter un coefficient de 5.

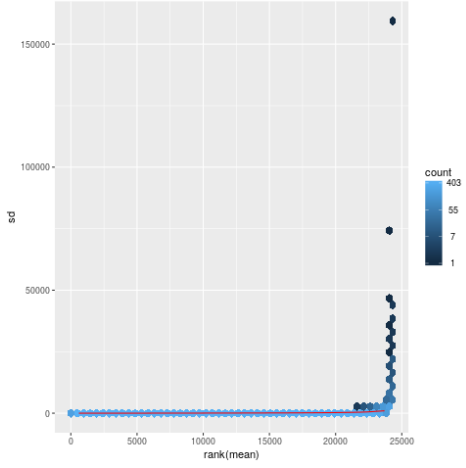
Dendogramme

But

Agréger les données de manière non-supervisée (et comparer avec l'attendu).

Problème

La variance (attendue) « tire » l'agrégation.



Dendogramme

But

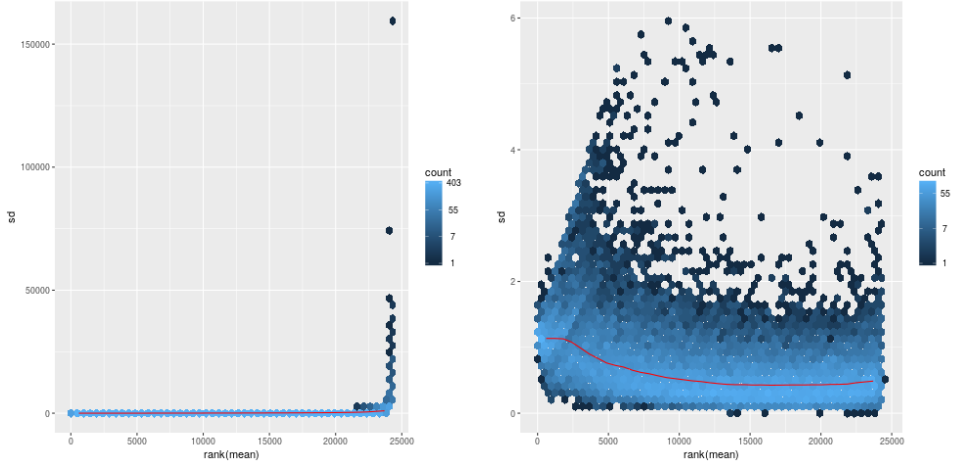
Agréger les données de manière non-supervisée (et comparer avec l'attendu).

Problème

La variance (attendue) « tire » l'agrégation.

Solution

Stabiliser la variance.



Dendogramme

But

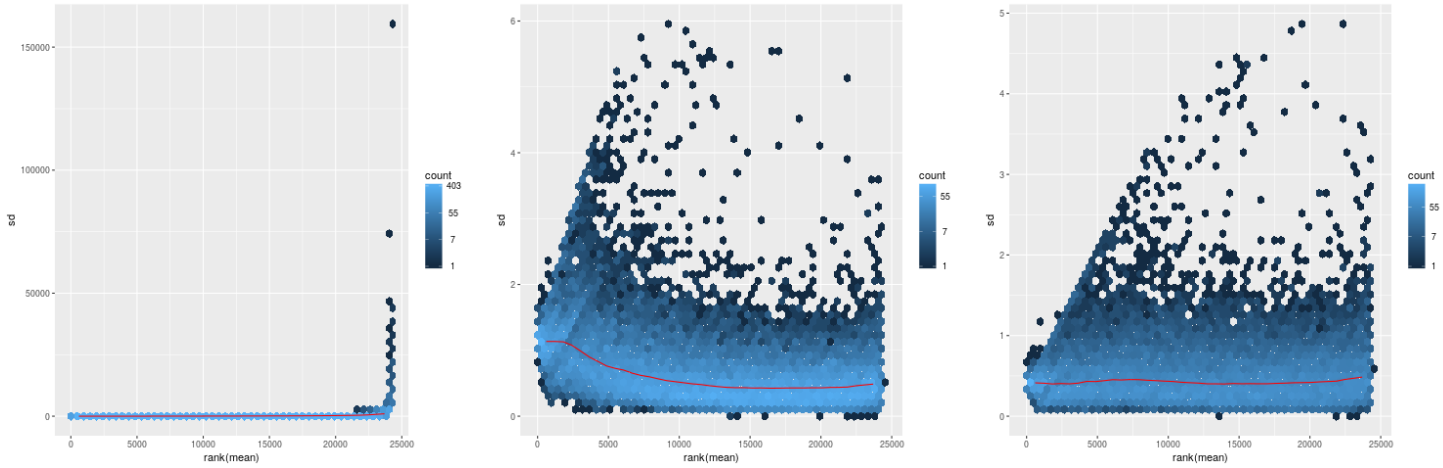
Agréger les données de manière non-supervisée (et comparer avec l'attendu).

Problème

La variance (attendue) « tire » l'agrégation.

Solution

Stabiliser la variance.



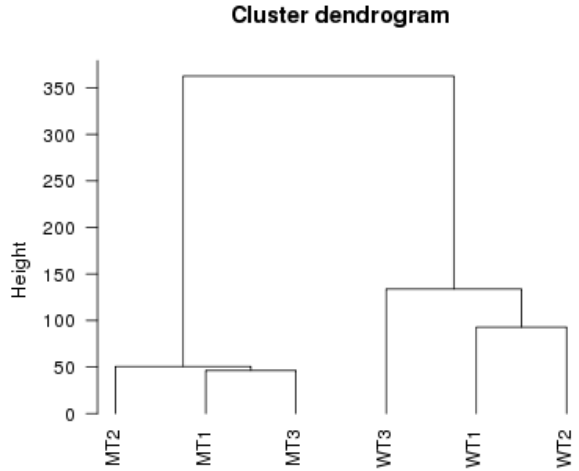
Dendogramme

But

Agréger les données de manière non-supervisée (et comparer avec l'attendu).

Méthode

- Définir une distance en 2 échantillons : distance euclidienne ($\sqrt{\sum_g (c_{1,g} - c_{2,g})^2}$).
- Choix des échantillons à fusionner : méthode de Ward (la paire d'échantillons avec la plus petite distance).



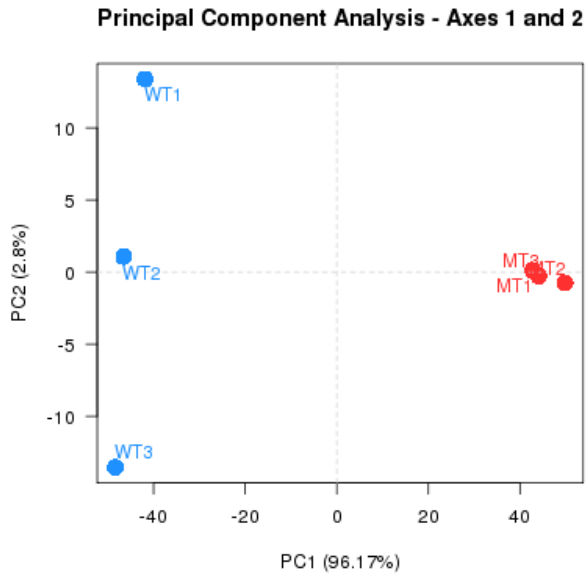
ACP

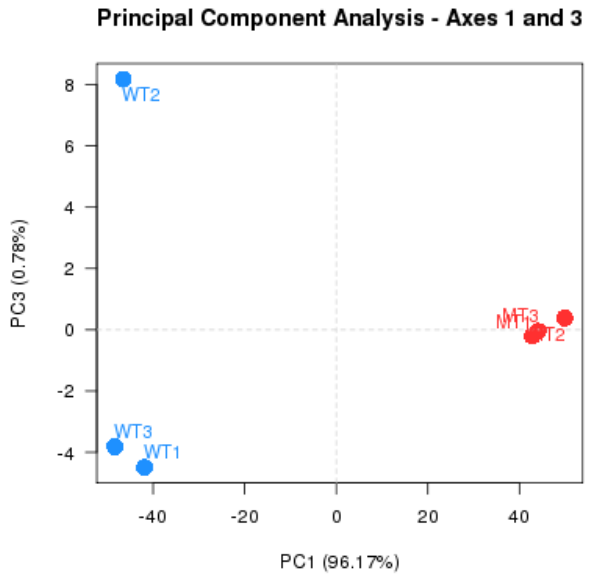
But

Agréger les données de manière non-supervisée (et comparer avec l'attendu).

Méthode

- Représenter les comptages dans un espace à n dimensions (où n est le nombre de gènes).
- Projeter sur des axes qui maximisent la dispersion.





Plan

1. Plan expérimental
2. Travaux pratiques
3. Exploration des données
4. Normalisation
5. Test

But

Comparer les comptages :

- deux gènes d'un même séquençage (intra-échantillons), ou bien
- le même gène dans deux séquençages (inter-échantillons).

Normalisation intra-échantillon

Exemple

gène	taille	nb. lectures
A	1kb	100
B	2kb	100

Quel gène est plus exprimé ?

Mesures

n_A : nombre de lectures pour le gène A, N : nombre total de lectures, l_A : taille de A.

- RPKM (reads per million reads per kilobase) : $\frac{n_i}{\frac{N}{10^6} \times \frac{l_A}{10^3}}$
- FPKM (fragments per million reads per kilobase) : même chose, sauf que l'on compte les fragments (2 lectures en PE).
- TPM (transcripts per million) : $\frac{\frac{n_A}{l_A}}{\sum_i \frac{n_i}{l_i}} \times 10^6$.

Normalisation inter-échantillon

Pourquoi ne pas utiliser les méthodes précédentes ?

La taille de la librairie est *une* normalisation possible.

Exemple

- Imaginez les séquençages suivants :
  - séquençage A : $n_1 = \ldots = n_{1000} = 1, n_{1001} = 1000$,
  - séquençage B : $n_1 = \ldots = n_{1000} = n_{1001} = 2$.
- Même nombre de lectures (ou presque). Tous les gènes sont-ils différentiellement exprimés ?

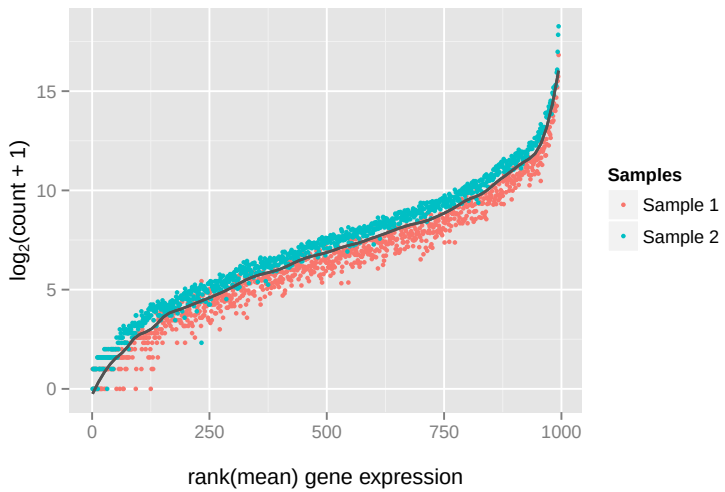
Méthode

- Supposer qu'un ensemble de gènes (en général, la majorité) des gènes n'est pas différentiellement exprimé.
- Appliquer un coefficient multiplicatif par échantillon pour que ces gènes aient les mêmes comptages.

Normalisation inter-échantillon — DESeq2

Méthode

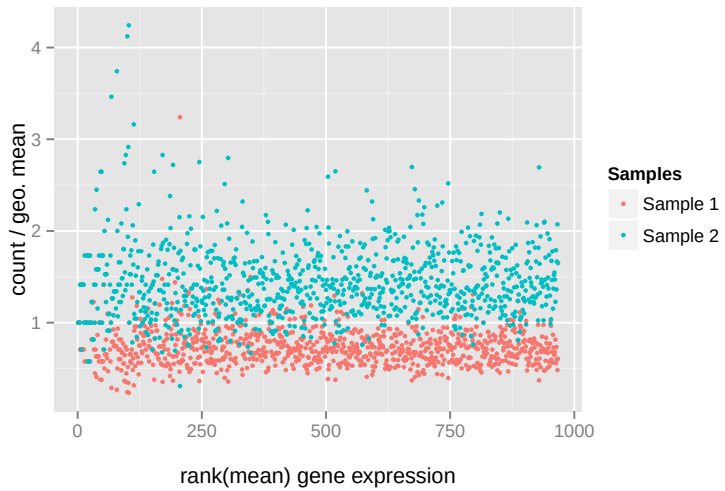
- Définir une expression moyenne pour chaque gène (moyenne géométrique) : $r_g = (\Pi_i c_{g,i})^{1/N}$



Normalisation inter-échantillon — DESeq2

Méthode

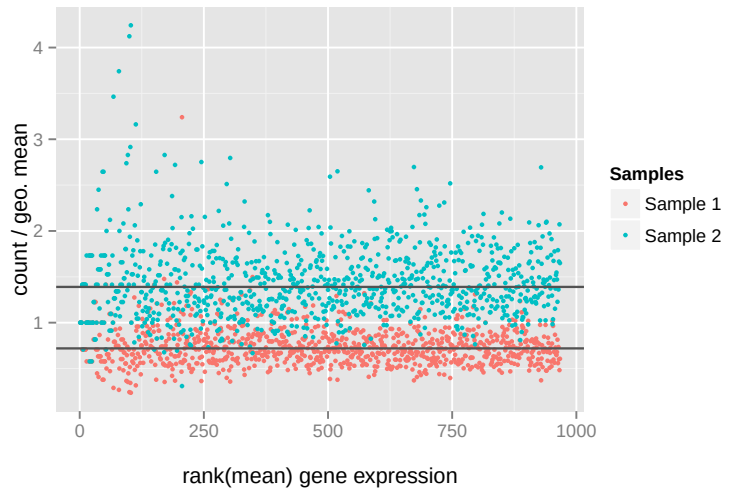
- Définir une expression moyenne pour chaque gène (moyenne géométrique) : $r_g = (\Pi_i c_{g,i})^{1/N}$
- Comparer chaque expression par rapport à la moyenne : $\tilde{c}_{g,i} = \frac{c_{g,i}}{r_g}$



Normalisation inter-échantillon — DESeq2

Méthode

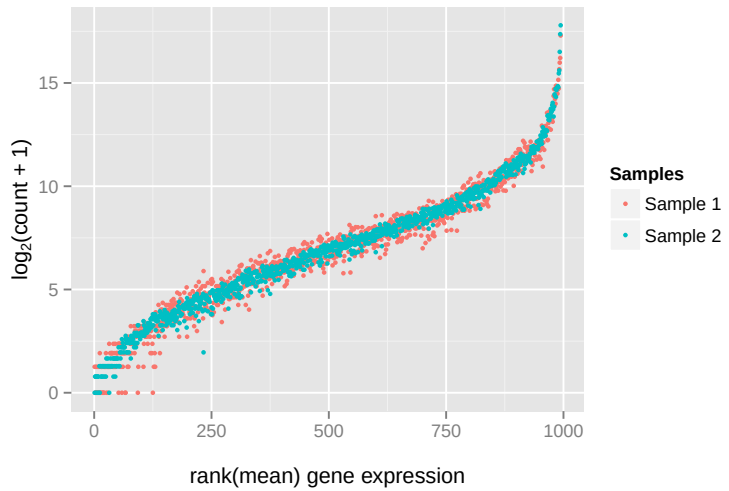
- Définir une expression moyenne pour chaque gène (moyenne géométrique) : $r_g = (\Pi_i c_{g,i})^{1/N}$
- Comparer chaque expression par rapport à la moyenne : $\tilde{c}_{g,i} = \frac{c_{g,i}}{r_g}$
- Trouver le coefficient moyen : $s_i = med_g(\tilde{c}_{g,i})$



Normalisation inter-échantillon — DESeq2

Méthode

- Définir une expression moyenne pour chaque gène (moyenne géométrique) : $r_g = (\Pi_i c_{g,i})^{1/N}$
- Comparer chaque expression par rapport à la moyenne : $\tilde{c}_{g,i} = \frac{c_{g,i}}{r_g}$
- Trouver le coefficient moyen : $s_i = med_g(\tilde{c}_{g,i})$
- Appliquer les coefficients : $c'_{g,i} = c_{g,i} \times s_i$

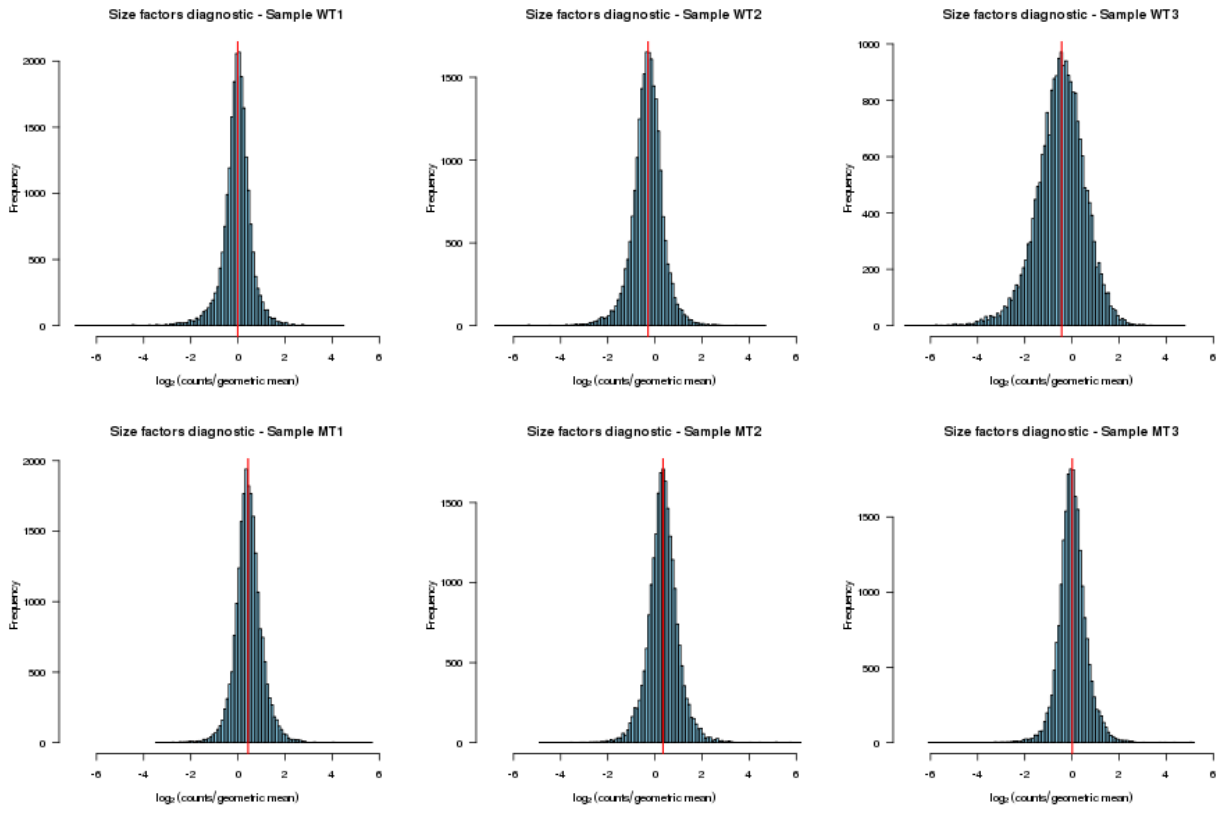


Normalisation inter-échantillon — DESeq2

Hypothèse

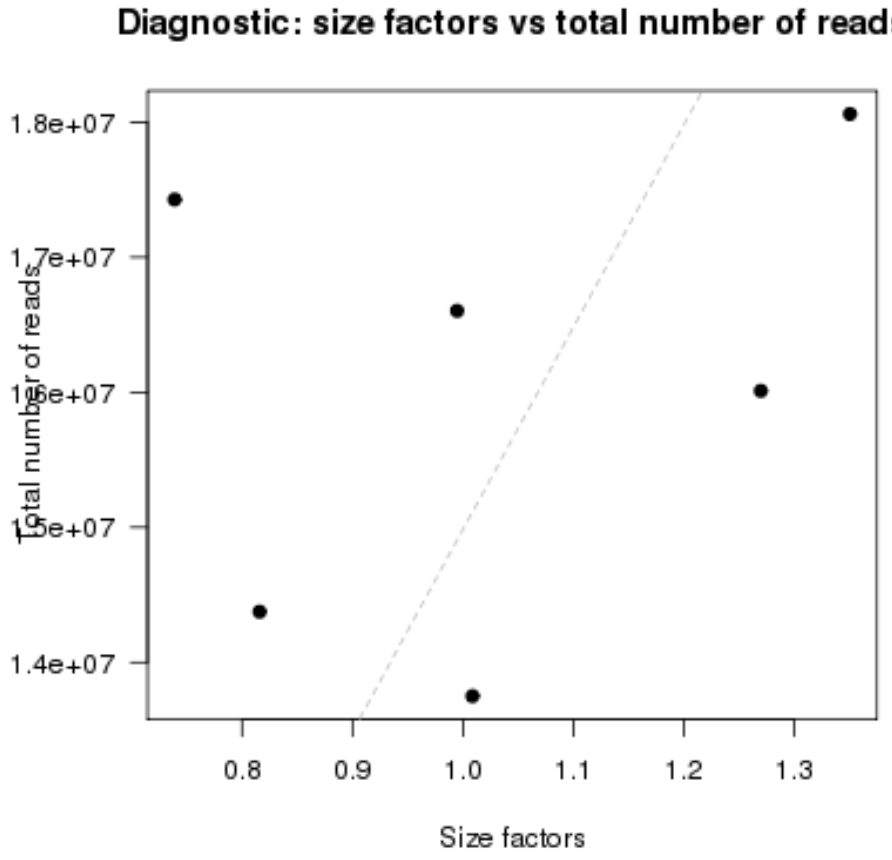
Cela suppose que les valeurs $\tilde{c}_{g_i}$ sont plutôt symétriques. Idéalement, la médiane est également le mode.

Vérification



Normalisation inter-échantillon

Corrélation entre facteurs de normalisation et taille de librairie

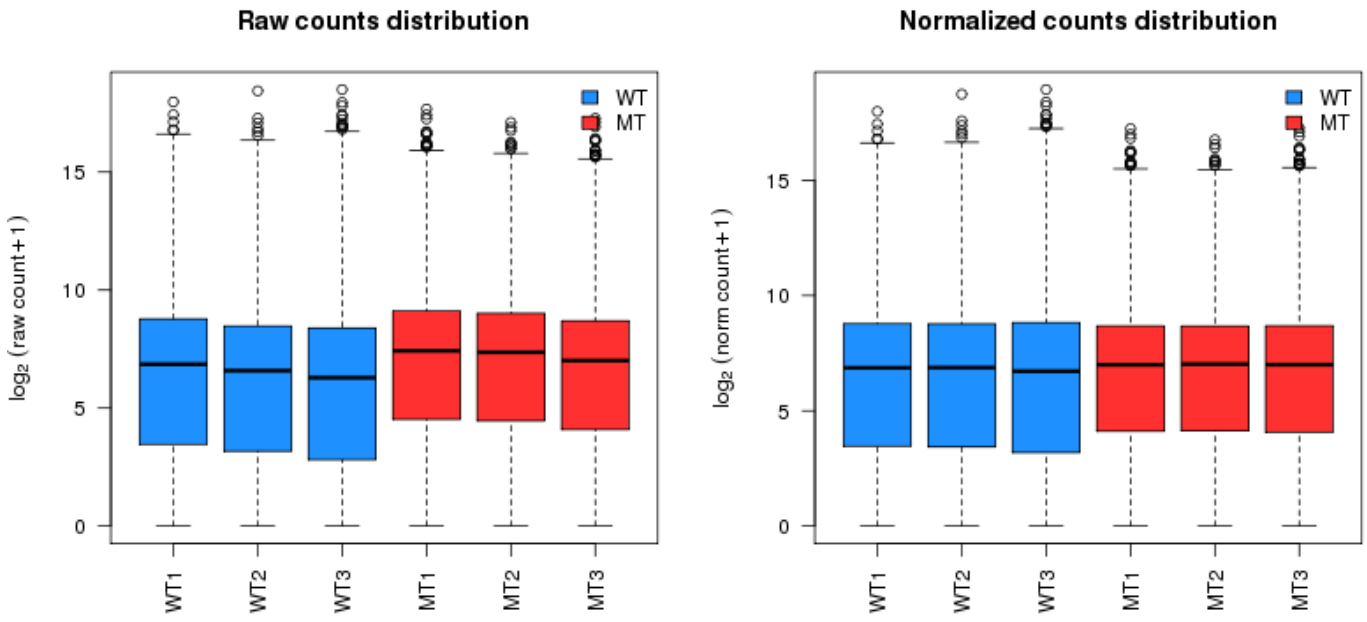


Notez qu'il n'y a pas toujours corrélation évidente.

Vérification de la normalisation

But

Vérifier que les distributions sont maintenant relativement similaires.



Plan

1. Plan expérimental
2. Travaux pratiques
3. Exploration des données
4. Normalisation
5. Test

Données

Ce que l'on a

gene_id	cond1_rep1	cond1_rep2	...
gene_A	12	14	
gene_B	42	56	
gene_C	0	0	

et un design expérimental.

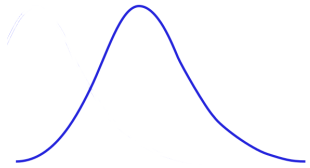
Ce que l'on veut

Connaître les gènes différentiellement entre les 2 conditions.

Méthode

Modèle

L'expression est modélisée par une loi paramétrique.

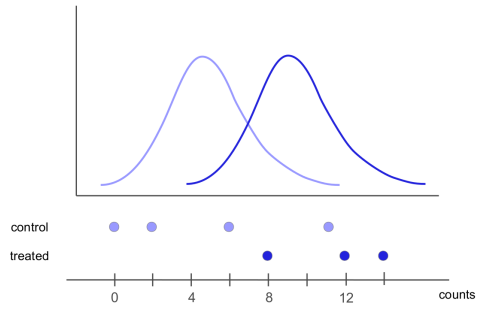


Hypothèse nulle

Le gène g n'est pas différentiellement exprimé.

Fitting

Les paramètres sont estimés.



Statistique

Une statistique quantifie la différence dont la loi sous l'hypothèse nulle est connue.

Test

Le test dit si l'on peut rejeter l'hypothèse nulle ou pas à un seuil donné.

Modèle

Correspondance

Données	Modèle
DNA extrait et clivé	urne avec des boules
séquençage	sélection de N boules
gène A	boule avec marqué A

Question

- J'ai 3 tirages.
- J'ai tiré 42, 45 et 56 boules A sur 1M de boules au total.
- Combien y avait-il de boules A dans l'urne ?

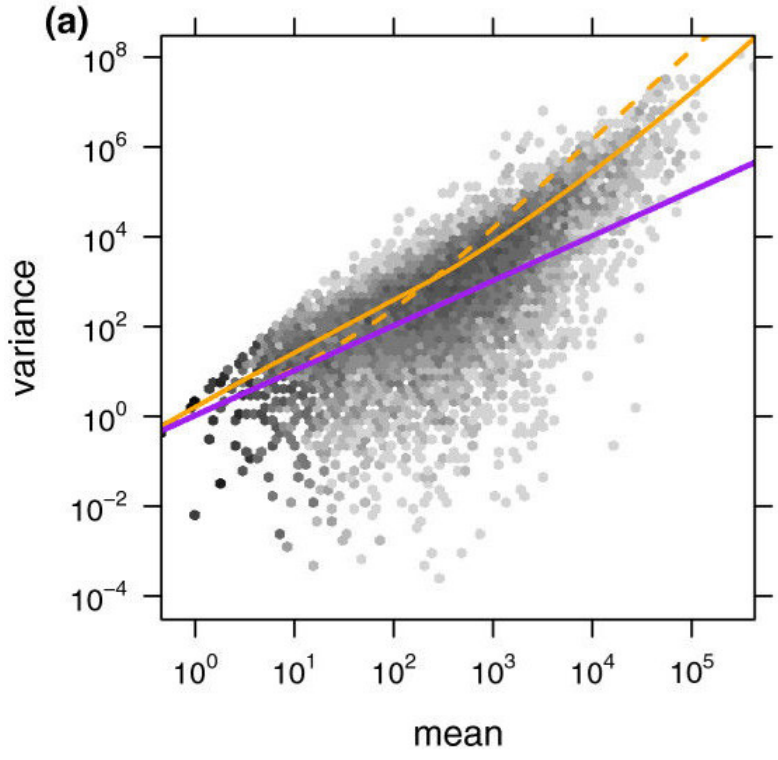
Réponse

- μ_A : proportion de A.
- 1 tirage : loi de Bernoulli de loi μ_A.
- N tirages : le nombre de boules A suit une loi Binomiale (N, μ_A).
- $\approx$ loi de Poisson $(N\mu_A)$.

Binomiale négative

Problème

La moyenne et la variance d'une loi de Poisson sont identiques.



Anders & Huber, BMC Gen. Biol., 2010

Binomiale négative

Définition

La binomiale négative *BN* a deux paramètres :

- la moyenne μ,
- le paramètre de dispersion α,

à estimer.

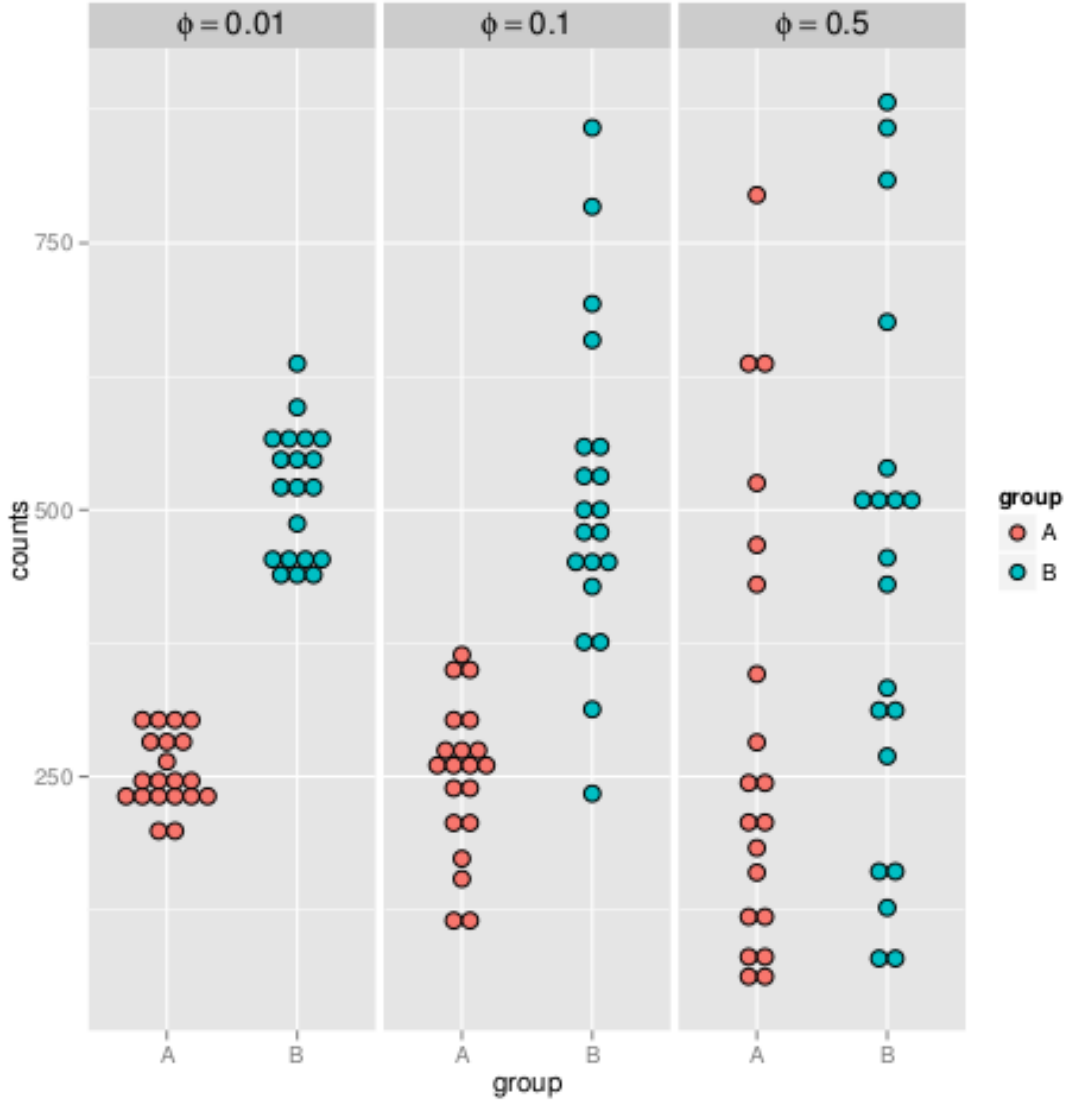
Hypothèse nulle

La moyenne de l'expression du gène *A* dans les conditions 1 et 2 est identique.

$$\mu_{A1} = \mu_{A2}$$

Binomiale négative — dispersion

Effet de la dispersion



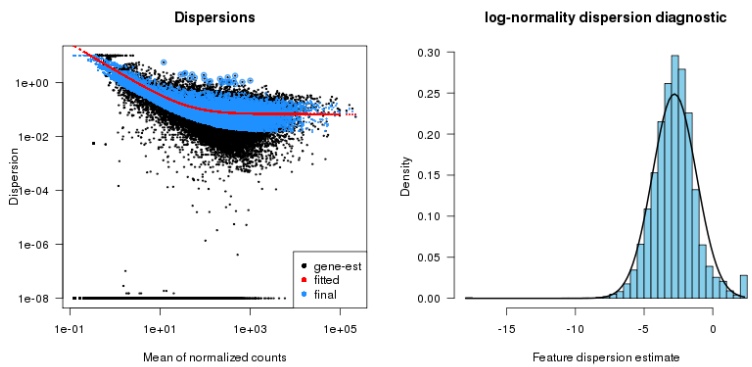
Estimation de la dispersion

Méthode

Par défaut, DESeq2 estime la relation moyenne/variance par un tendance qui peut être choisie comme linéaire ou bien « arbitraire » (LOESS, lissage local).

But

Vérifier que la relation est bien estimée.



p-valeur

Définition

La p-valeur est la probabilité d'observer un élément au moins aussi extrême, sous l'hypothèse nulle.

Ce n'est pas

la probabilité que l'hypothèse nulle soit vraie.

Intérêt

Proposer une quantification dans le choix du rejet de l'hypothèse nulle (contrôler le risque de rejeter l'hypothèse nulle à tord).

Remarques

- La p-valeur peut être déduite de la loi modélisant l'hypothèse nulle et des observations.
- Un seuil de 5% ou 1% est simplement une convention.

Distribution des p-valeurs

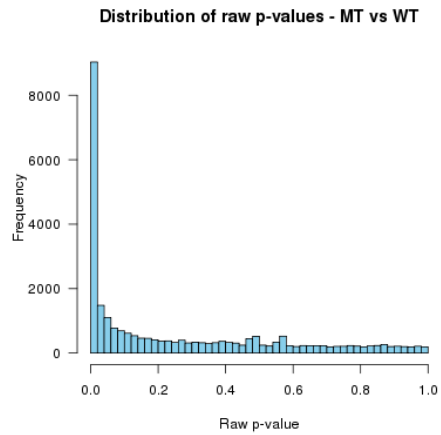
But

Vérifier que le test s'est correctement réalisé.

Méthode

- Sous l'hypothèse nulle, la distribution des p-valeurs est uniforme sur $[0, 1]$.
- Les gènes différentiellement exprimés doivent avoir une p-valeur proche de 0.

$\Rightarrow$ La distribution doit être un mélange des deux précédentes distributions.



Comment se tromper

Deux types d'erreur

		vérité	
		diff. exp.	non diff. exp.
décision	détecté	OK	type I
	non-détecté	type II	OK

Remarque

Accepter l'hypothèse nulle lorsque p-valeur $\leq \alpha$ revient à contrôler l'erreur de type I au seuil α.

Tests multiples

Exemple

Expression d'un gène dans des populations sauvages et mutantes.

sauvages											
10	20	30	30	30	30	30	30	30	30	40	50
mutants											
10	20	30	30	30	30	30	30	30	30	40	50

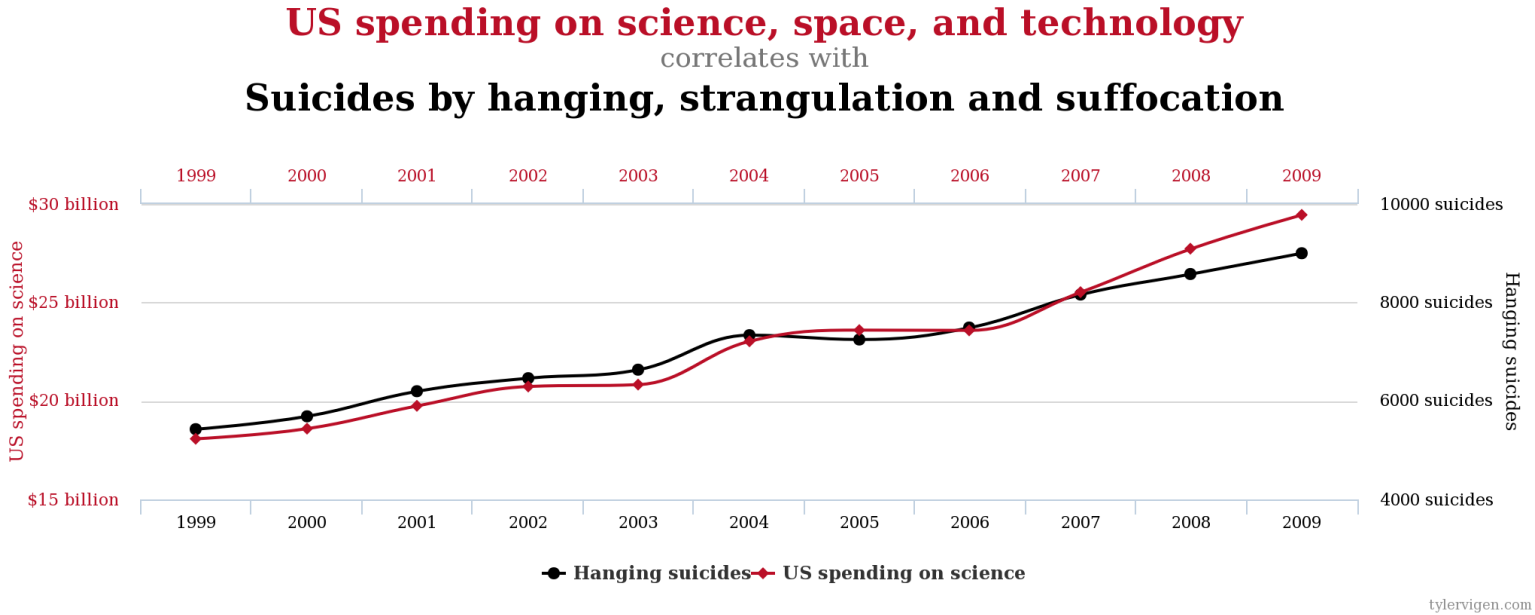
On échantillonne (au hasard) les populations :

- Si l'on prend 30, 30, 30 vs 30, 30, 30 : pas d'expression différentielle.
- Si l'on prend 10, 20, 30 vs 30, 40, 50 : expression différentielle.

Conclusion

On a toujours une chance de trouver des gènes différentiellement exprimés (par hasard).

Spurious correlation



http://tylervigen.com

Correction des tests multiples

But

Contrôler le FDR (false positive rate), le nombre d'erreurs de type I, sur l'ensemble des tests.

Méthode

Algorithme de Benjamini–Hochberg donnant des p-valeurs ajustées.

MA-plot

But

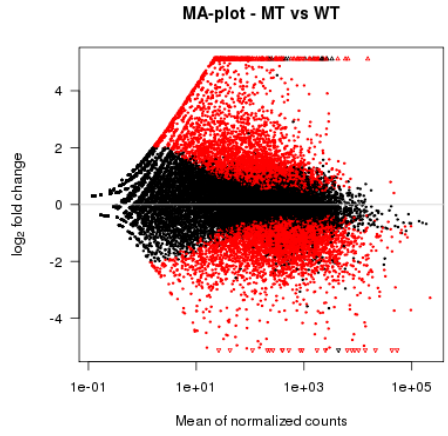
Observer les tendances de différence d'expression.

Méthode

Un point est un gène :

- abscisse : moyenne (A)
- ordonnée : log-ratio (M)

Les points rouges sont les gènes différentiellement exprimés. Les points rouges et noirs peuvent se superposer.



Volcano plot

But

Observer les tendances de différence d'expression.

Méthode

Un point est un gène :

- abscisse : log-ratio
- ordonnée : p-valeurs

Les points rouges sont les gènes différentiellement exprimés.