SINGLE-CELL RNA-SEQ

Focus on normalization

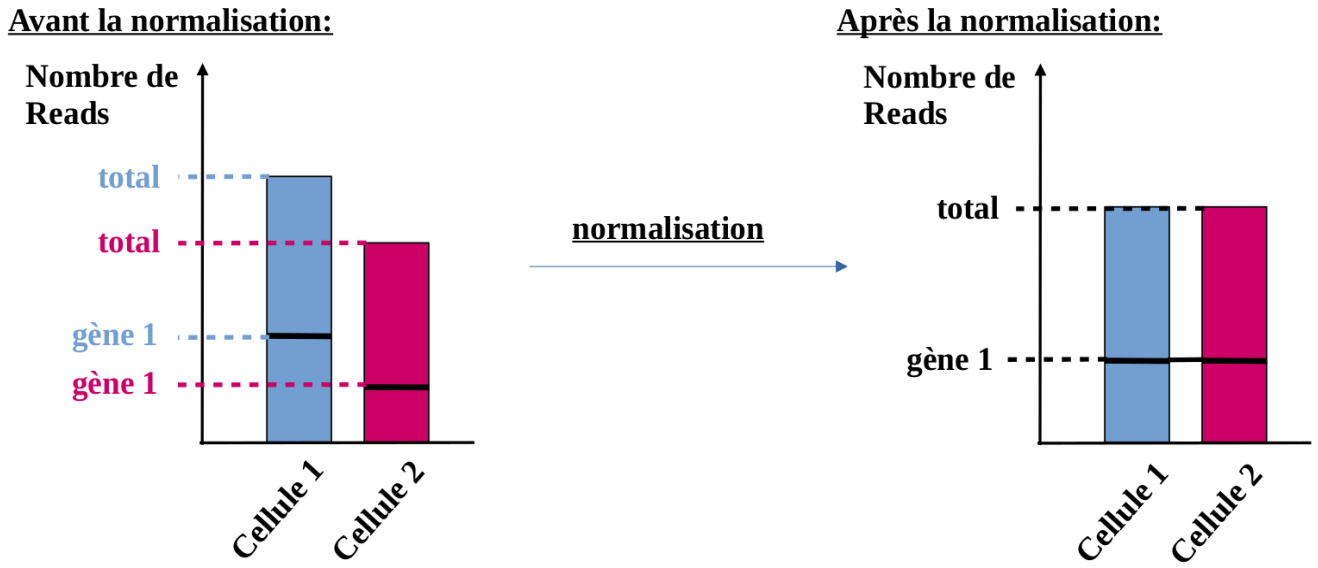
Bioinformatics Core Facility
bigr@gustaveroussy.fr
B2M+1, Gustave-Roussy

GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

Rappels : Normalisation

But : supprimer l'influence des biais techniques (profondeur de séquençage), tout en préservant les variations biologiques d'intérêt.

→ rendre les distributions comparables entre les cellules.



Seurat : LogNormalisation + HVG + scaling ou SCTransform

LogNormalisation

LogNormalisation: les mesures d'expression génique de chaque cellule sont divisées par l'expression totale de cette cellule (profondeur de séquençage) , puis multipliées par un facteur d’échelle (scale factor), suivi d'un ajout de pseudo-compte (pour éviter log(0) = −∞), et d'une transformation logarithmique.

$$f(X) = \log\left(\frac{Comptage\ du\ gène\ X\ dans\ la\ cellule\ Y}{Total\ des\ comptages\ dans\ la\ cellule\ Y} \times scale\ factor + 1\right)$$

LogNormalisation

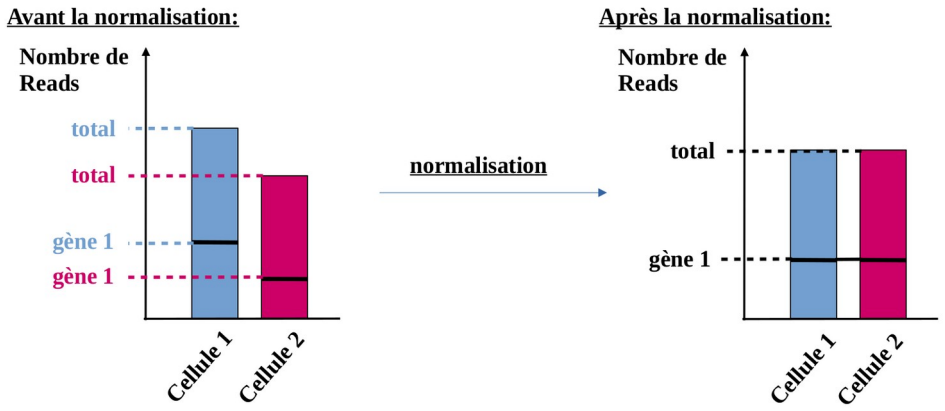
$$f(X) = \log\left(\frac{Comptage\ du\ gène\ X\ dans\ la\ cellule\ Y}{Total\ des\ comptages\ dans\ la\ cellule\ Y} \times scale\ factor + 1\right)$$

Scale factor ?

→ 10 000 par défaut, avec Seurat.

- hypothèse : les populations cellulaires sont homogènes & le niveau d'ARN est similaire dans toutes les cellules.

- en pratique : les hypothèses ne sont pas toujours vérifiées, mais les gens utilisent cette méthode.



LogNormalisation

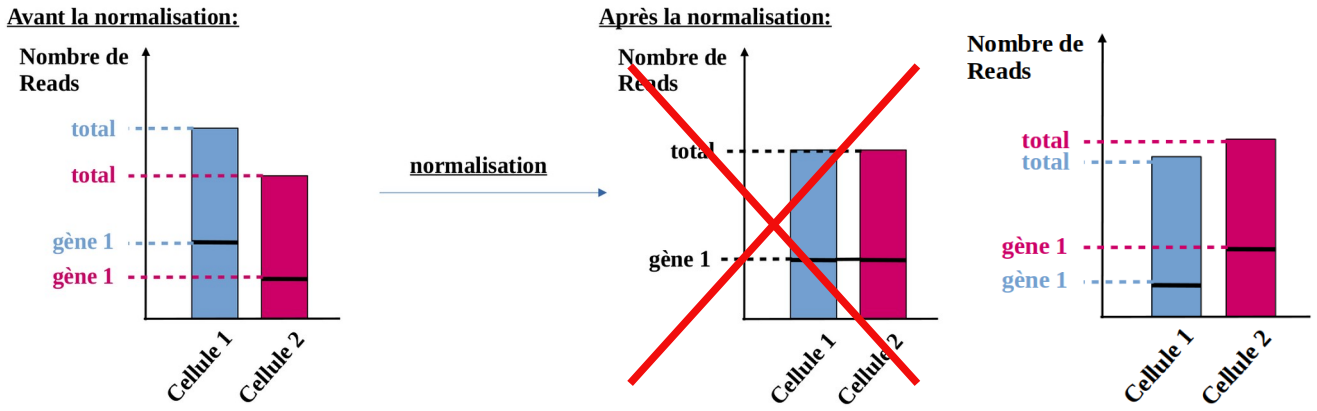
$$f(X) = \log\left(\frac{Comptage\ du\ gène\ X\ dans\ la\ cellule\ Y}{Total\ des\ comptages\ dans\ la\ cellule\ Y} \times scale\ factor + 1\right)$$

Scale factor ?

→ 10 000 par défaut, avec Seurat.

- hypothèse : les populations cellulaires sont homogènes & le niveau d'ARN est similaire dans toutes les cellules.

- en pratique : les hypothèses ne sont pas toujours vérifiées, mais les gens utilisent cette méthode.



LogNormalisation

$$f(X) = \log\left(\frac{Comptage\ du\ gène\ X\ dans\ la\ cellule\ Y}{Total\ des\ comptages\ dans\ la\ cellule\ Y} \times scale\ factor + 1\right)$$

Log ?

→ Gérer la relation Moyenne-Variance :

Problème sous-jacent : la variance d'expression d'un gène ne doit pas être liée (corrélée) à sa moyenne d'expression à travers toutes les cellules ; or, les gènes avec une moyenne d'expression élevée ont toujours une variance mesurée plus élevée (hétéroscédasticité).

	Cellule 1	Cellule 2	Moyenne	Variance	Ecart-type
Gène A	5	10	(5+10)/2= 7,5	((5-7,5)²+(10-7,5)²)/2= 6,25	sqrt(6,25)= 2,5
Gène B	50	100	(50+100)/2= 75	((50-75)²+(100-75)²)/2= 625	sqrt(625)= 25

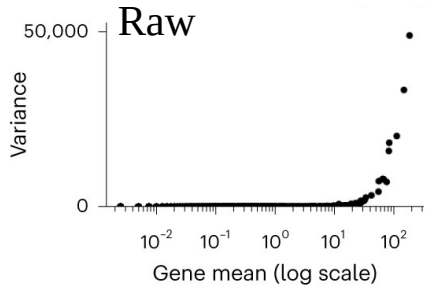
LogNormalisation

$$f(X) = \log\left(\frac{Comptage\ du\ gène\ X\ dans\ la\ cellule\ Y}{Total\ des\ comptages\ dans\ la\ cellule\ Y} \times scale\ factor + 1\right)$$

Log ?

→ Gérer la relation Moyenne-Variance :

Problème sous-jacent : la variance d'expression d'un gène ne doit pas être liée (corrélée) à sa moyenne d'expression à travers toutes les cellules ; or, les gènes avec une moyenne d'expression élevée ont toujours une variance mesurée plus élevée (hétéroscédasticité).



Or ce qui nous intéresse, c'est la variation du gène pas son niveau d'expression, car on sait bien qu'un gène faiblement exprimé peut avoir un impact majeur sur la physiologie de la cellule (et au contraire pour des gènes fortement exprimés), donc on devrait avoir une variance qui est indépendante du niveau d'expression du gène.

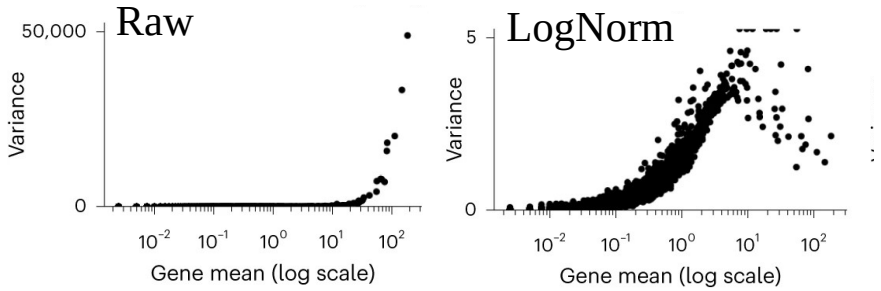
LogNormalisation

$$f(X) = \log\left(\frac{Comptage\ du\ gène\ X\ dans\ la\ cellule\ Y}{Total\ des\ comptages\ dans\ la\ cellule\ Y} \times scale\ factor + 1\right)$$

Log ?

→ Gérer la relation Moyenne-Variance :

Problème sous-jacent : la variance d'expression d'un gène ne doit pas être liée (corrélée) à sa moyenne d'expression à travers toutes les cellules ; or, les gènes avec une moyenne d'expression élevée ont toujours une variance mesurée plus élevée (hétéroscédasticité).

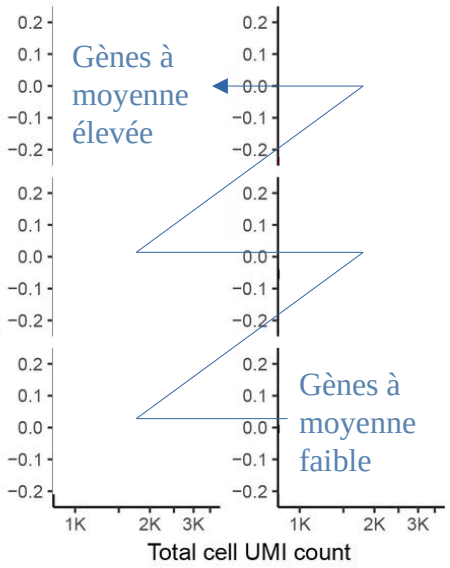


Problème : ne fonctionne que pour les expressions faibles et moyennes, mais pas pour les expressions fortes qui restent corrélées avec la variance.

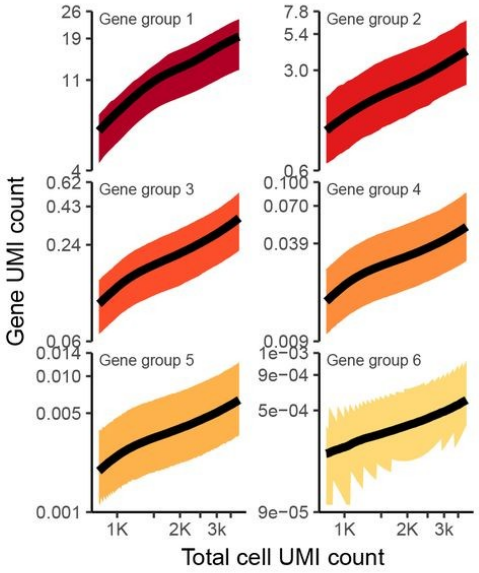
=> Donc la logNormalisation n'est pas optimale.

LogNormalisation

Sens de lecture des graphes

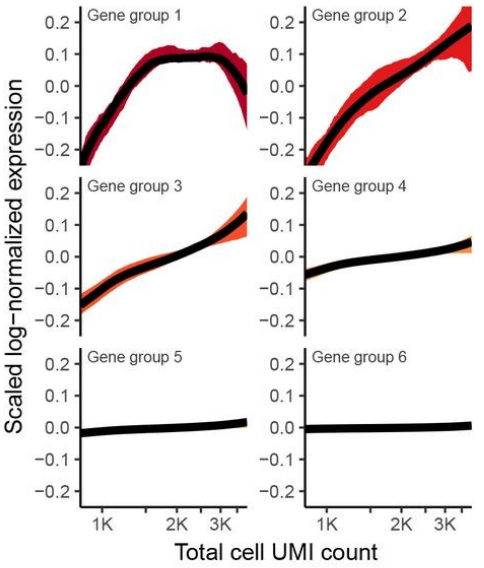


Non-normalisés



Corrélation avec la profondeur de séquençage

Log-normalisés



Ne fonctionne que pour les expressions faibles et moyennes, mais pas pour les expressions fortes qui restent corrélées avec la profondeur de séquençage.

=> Donc la logNormalisation n’est pas optimale.

HVG

Gènes Hautement Variables (HVG) : sélection des gènes d'intérêt pour identifier des types cellulaires différents.

→ en théorie : 500 < HVG < 3000

→ en pratique : HVG = 2000 pour Seurat (peut être ajusté selon le contexte d'hétérogénéité cellulaire de l'échantillon étudié).

Besoin de la variance, mais les gènes avec une moyenne élevée ont une forte variance et est corrélée avec la profondeur de séquençage, donc on a besoin de mieux normaliser.

→ Normalisation VST (Variance Stabilizing Transformation)

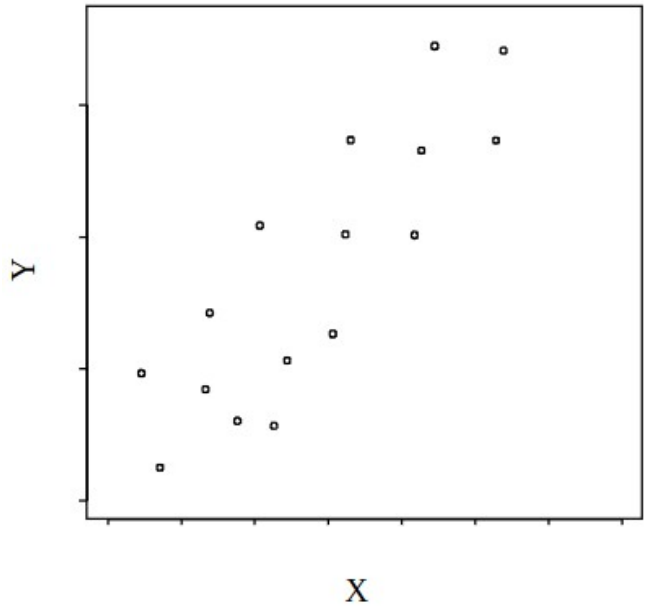
Aparté : régression linéaire

But : Expliquer une variable y (variable à expliquer) à l'aide d'une variable x (variable explicative).

→ chercher une fonction f telle que :

$$y = f(x)$$

On va faire plusieurs observations de y et x, qu'on va appeler y_i et x_i et qu'on représente sur un graphique y ~ x.



Aparté : régression linéaire

But : Expliquer une variable y (variable à expliquer) à l'aide d'une variable x (variable explicative).

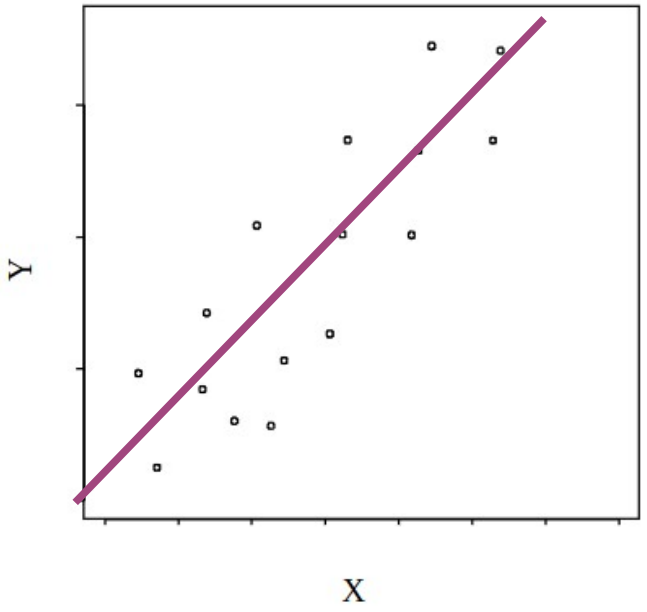
→ chercher une fonction f telle que :

$$y = f(x)$$

On va faire plusieurs observations de y et x, qu'on va appeler y_i et x_i et qu'on représente sur un graphique.
Ici la relation entre y et x semble linéaire donc la fonction f doit être du type :

$$y = f(x) = a\,x + b$$

modèle mathématique
coefficient directeur
ordonnée à l'origine



Aparté : régression linéaire

$y = f(x) = a\,x + b$

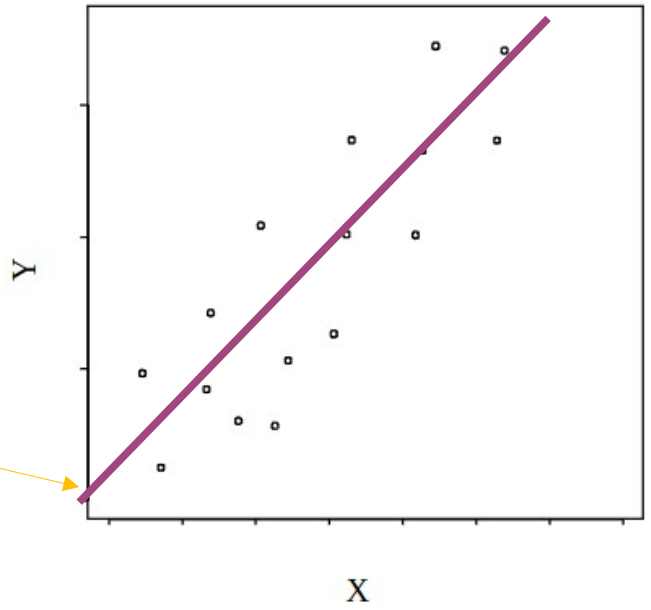
modèle mathématique
coefficient directeur
ordonnée à l'origine

Identification de **b** :

Si x = 0; y = **a***0 + **b** = **b**

Puis identification de **a** :

Puis **a** = (y - **b**)/x



Aparté : régression linéaire

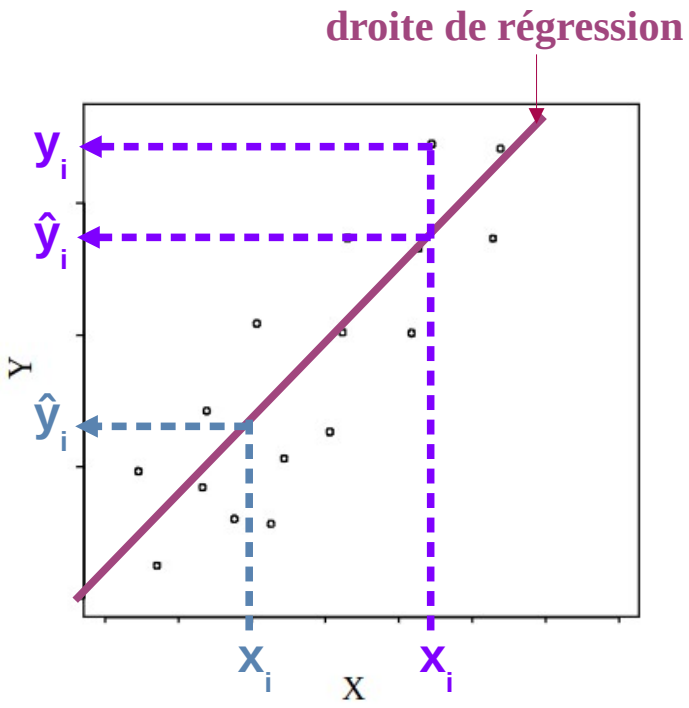
$y = f(x) = a\,x + b$

modèle mathématique
coefficient directeur
ordonnée à l'origine

But final :

- faire de la **prédiction** de $\hat{y}_i$ à partir de x_i observés.

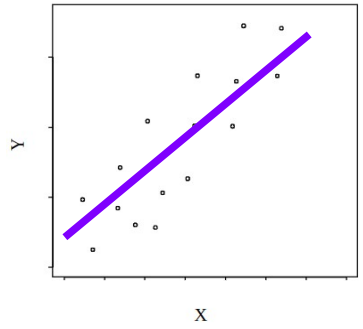
- faire de la **correction** de y_i observés à partir de x_i observés (la régression).

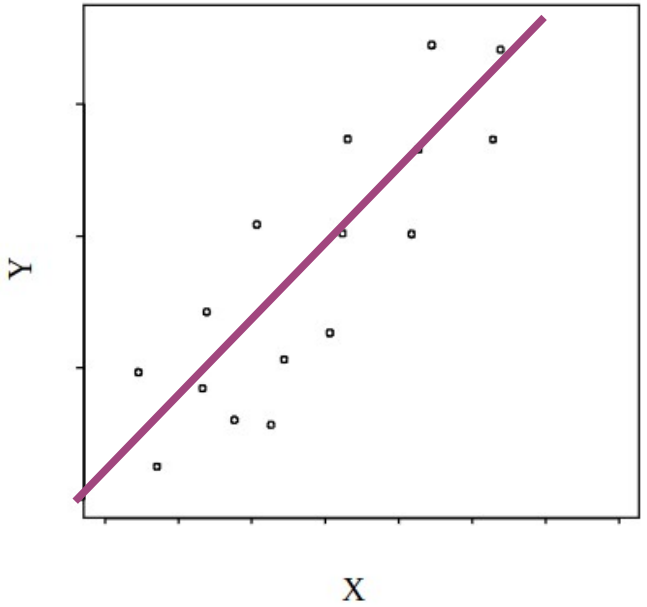


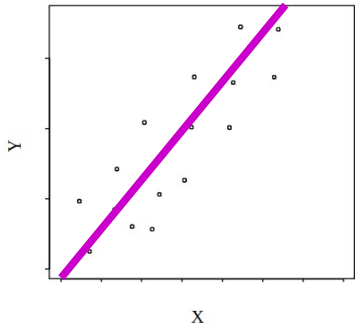
Aparté : régression linéaire

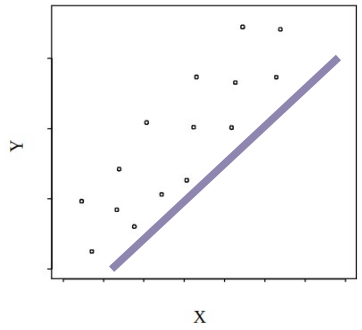
$y = f(x) = a\,x + b$

La bonne droite ?





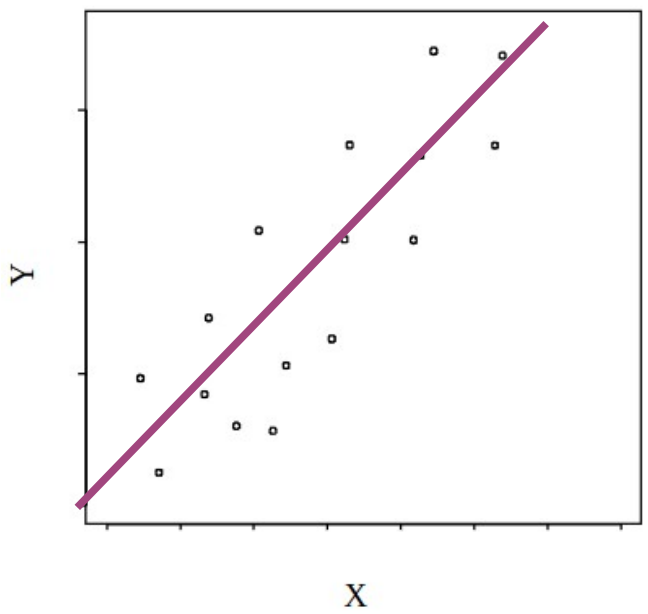


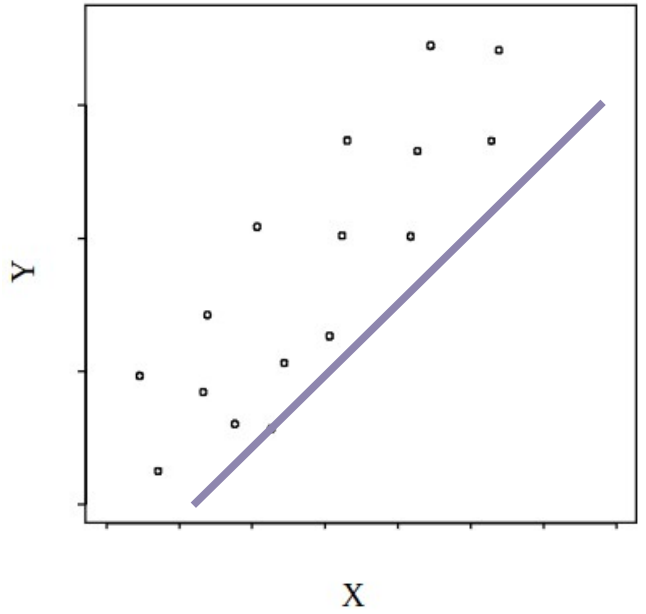


Aparté : régression linéaire

$$y = f(x) = a\,x + b$$

La bonne droite ?

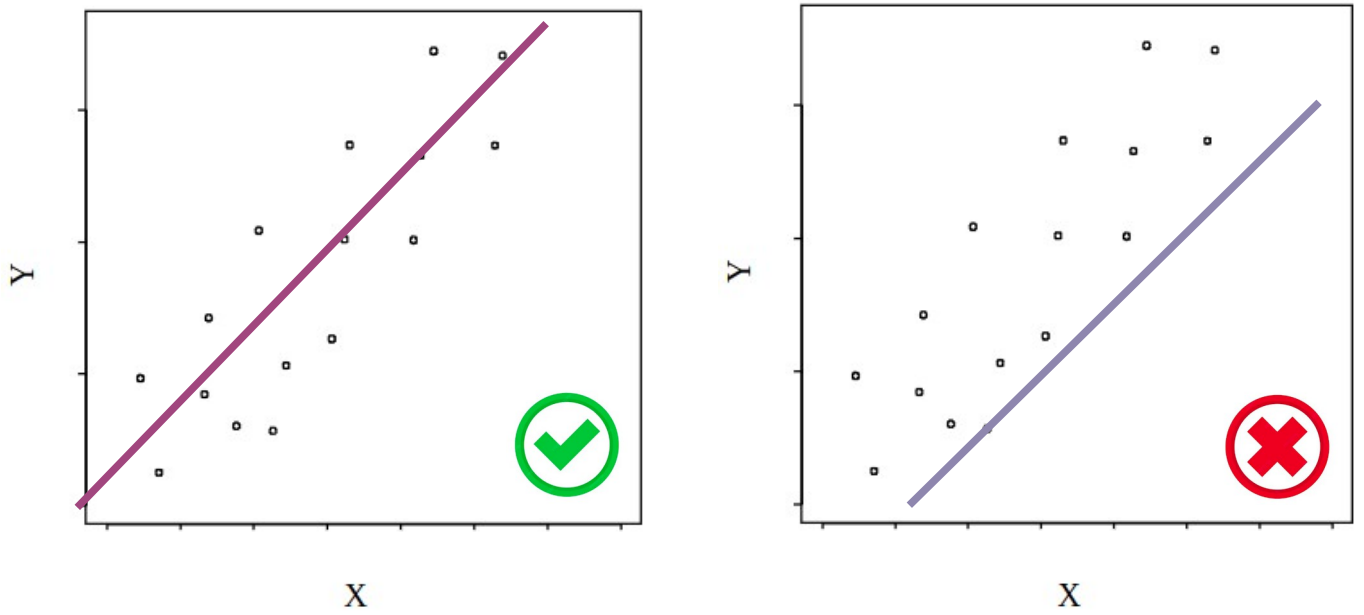




Aparté : régression linéaire

$y = f(x) = a\,x + b$

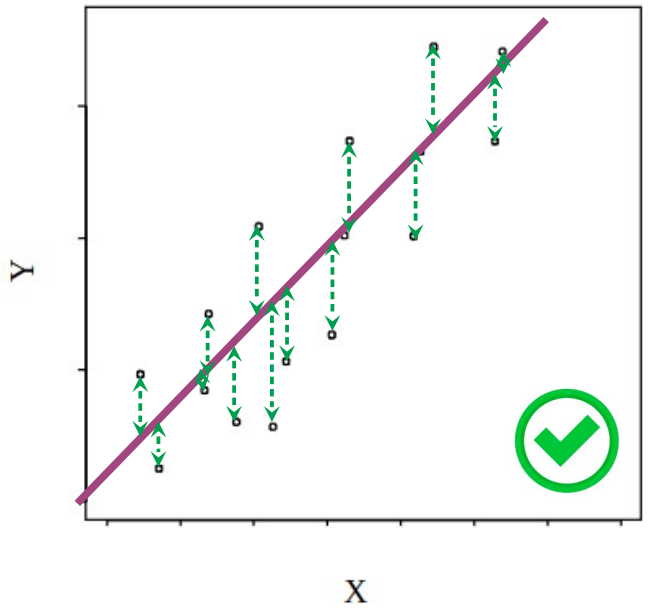
La bonne droite ?



Aparté : régression linéaire

$$y = f(x) = a\,x + b$$

La bonne droite ?



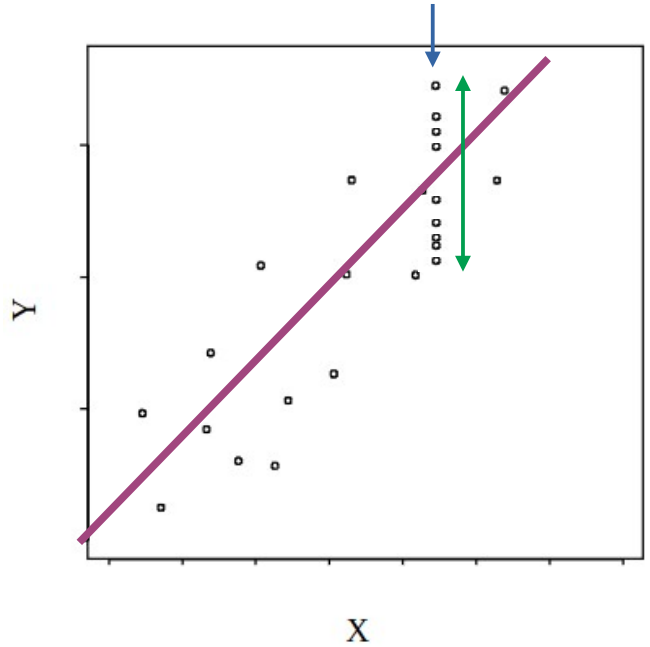
Meilleure car elle passe au plus proche des observations (points sur le graphique)
↔ minimiser les distances entre la droite et les points (résidus).

Méthodes mathématiques :
- la méthode du maximum de vraisemblance,
- la méthode des moindres carrés,
- la méthode des moments,
- les méthodes bayésiennes...

Aparté : régression linéaire

$y = f(x) = \mathbf{a}\, x + \mathbf{b}$

Et si on a plusieurs point pour un même x_i ?



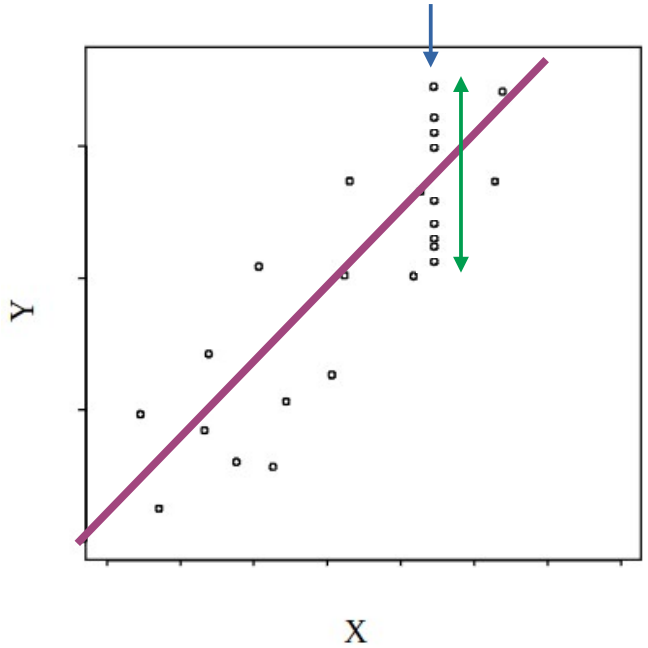
Le modèle permet d'approximer y_i mais on a une marge d'erreur qu'on va pouvoir quantifier et évaluer pour trouver la meilleure réponse pour y_i.

L'erreur s'appelle ε et a une moyenne et une variance.
Si ε est aléatoire, alors elle suit une loi normale (Gaussienne).

Aparté : régression linéaire

$y = f(x) = a\,x + b$ → $y = f(x) = a\,x + b + \varepsilon$

Et si on a plusieurs point pour un même x_i ?



Le modèle permet d'approximer y_i mais on a une marge d'erreur qu'on va pouvoir quantifier et évaluer pour trouver la meilleure réponse pour y_i.

L'erreur s'appelle ε et a une moyenne et une variance.
Si ε est aléatoire, alors elle suit une loi normale (Gaussienne).

Aparté : régression linéaire

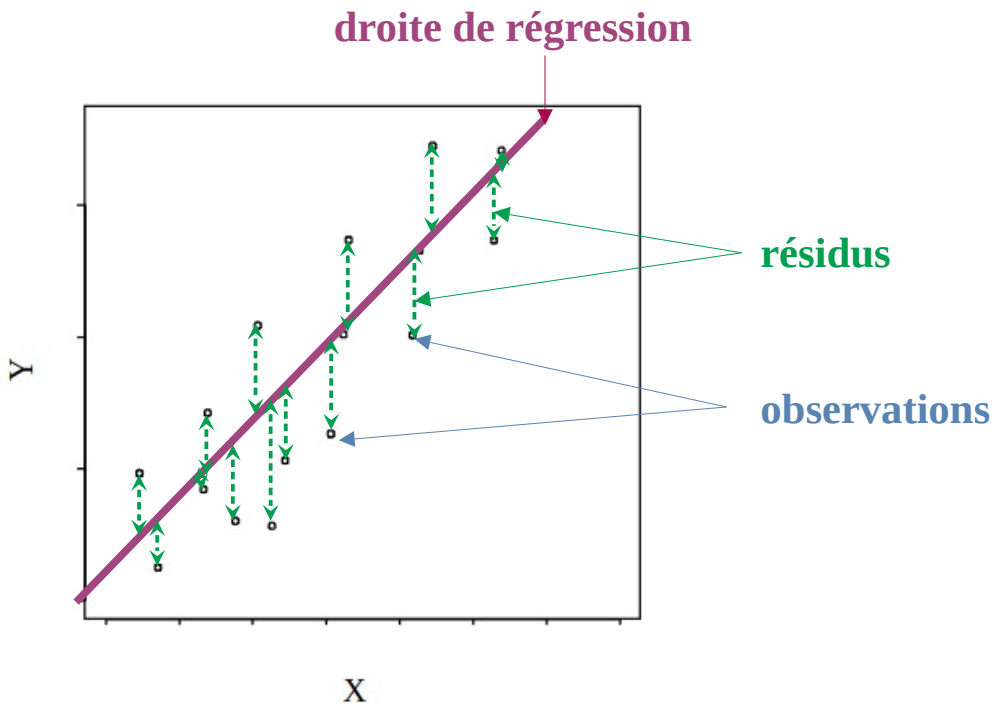
Résumé du vocabulaire :

Par convention **a** et **b** sont remplacés par β_1 et β_0:

modèle mathématique

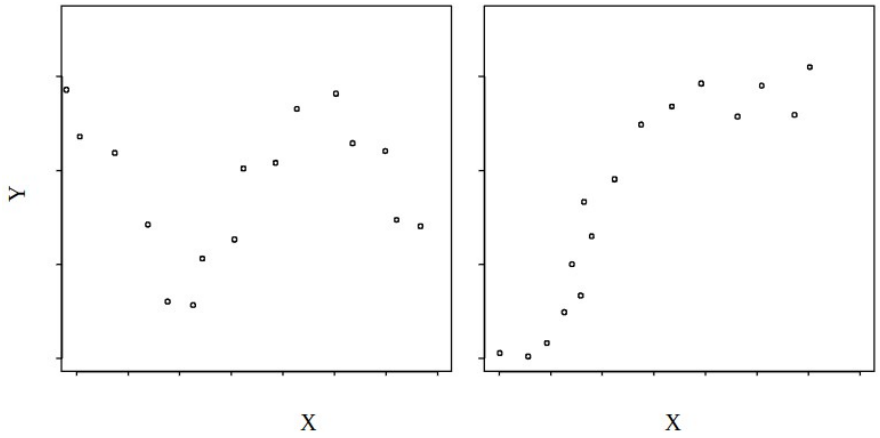
$$y = \beta_1 x + \beta_0 + \varepsilon$$

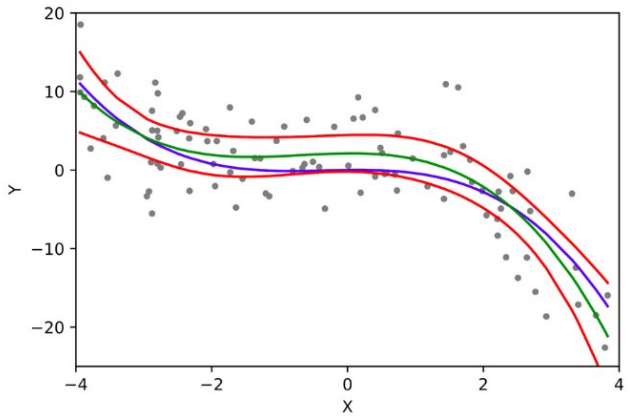
y : variable à expliquer
x : variable explicative
β_0 et β_1 : paramètres
ε : erreur



Aparté : régression linéaire

Autres types de modèles de régression:





sinusoïdale

$y = \beta_1 \sin(x)$

sigmoïdale

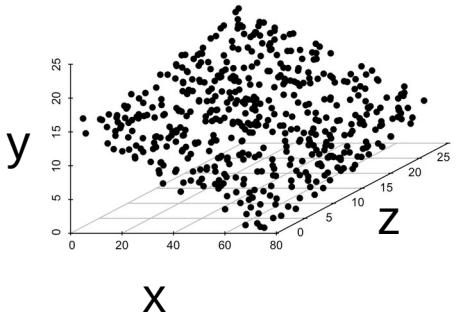
$y = 1/(1+e^{-x})$

polynomiale

$y = \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \beta_0$

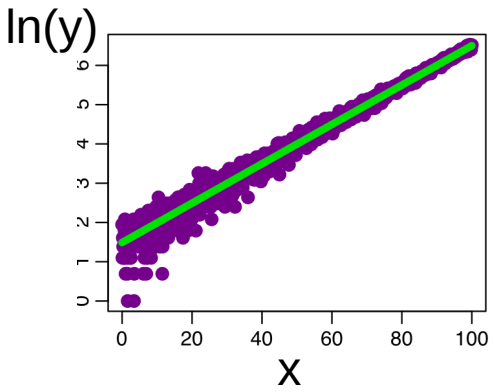
Aparté : régression linéaire

Autres types de modèles de régression:
les modèles linéaires généralisés (GLM) :



linéaire multiple

$y = \beta_1 x + ... + \beta_n z + \beta_0$

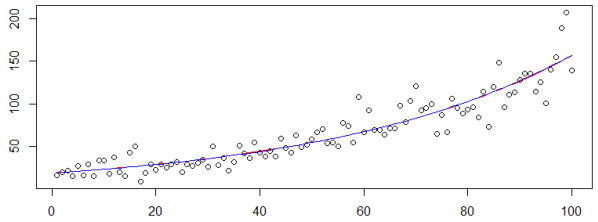


poisson

$ln(y) = \beta_1 x + \beta_0$

ou si + de variables :

$ln(y) = \beta_1 x + ... + \beta_n z + \beta_0$



binomial négatif

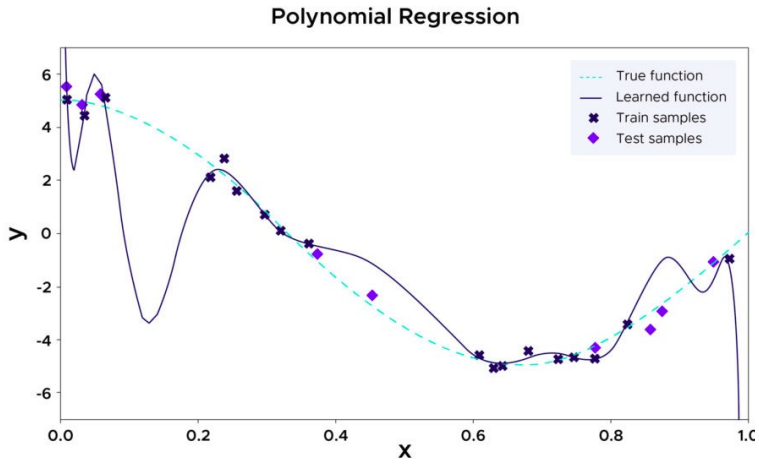
$ln(mean(y)) = \beta 1\ x + \beta 0$

ou si + de variables :

$ln(mean(y)) = \beta 1\ x + ... + \beta n\ z + \beta 0$

Aparté : régression linéaire

Qu'est-ce que l'overfitting ?
C'est quand le modèle est trop proche des données et ne permet pas de mesurer une erreur adéquat.



L'overfitting peut arriver quand :

- les données observées ne représentent pas vraiment la réalité (pas assez d'observations, ou observé dans un cas particulier). Il faut donc en tenir compte quand on identifie les paramètres du modèles.

- le modèle choisi est trop complexe par rapport à la taille du jeu de données.

HVG

→ Normalisation VST (Variance Stabilizing Transformation) :

1) Calcule la moyenne et la variance pour chaque gène en utilisant les données brutes non-normalisées.

	Cellule 1	Cellule 2	Moyenne	Variance	Ecart-type
Gène A	5	10	7,5	6,25	2,5
Gène B	50	100	75	625	25

Reprise du tableau diapo 45

Disclaimer : Les chiffres du tableau ainsi que les graphes, ne représentent pas des données réelles et permettent uniquement d'illustrer le propos.

HVG

→ Normalisation VST (Variance Stabilizing Transformation) :

1) Calcule la moyenne et la variance pour chaque gène en utilisant les données brutes non-normalisées.

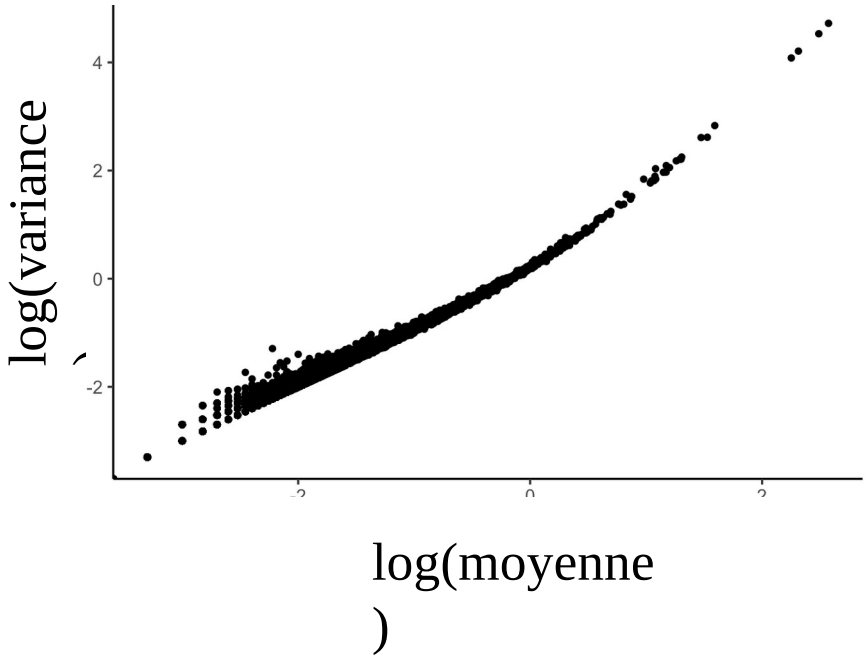
2) Calcule le log des moyennes et des variances.

	Cellule 1	Cellule 2	Moyenne	Variance	Ecart-type	log(Moyenne)	log(Variance)
Gène A	5	10	7,5	6,25	2,5	0,88	0,8
Gène B	50	100	75	625	25	1,88	2,8

HVG

→ Normalisation VST (Variance Stabilizing Transformation) :

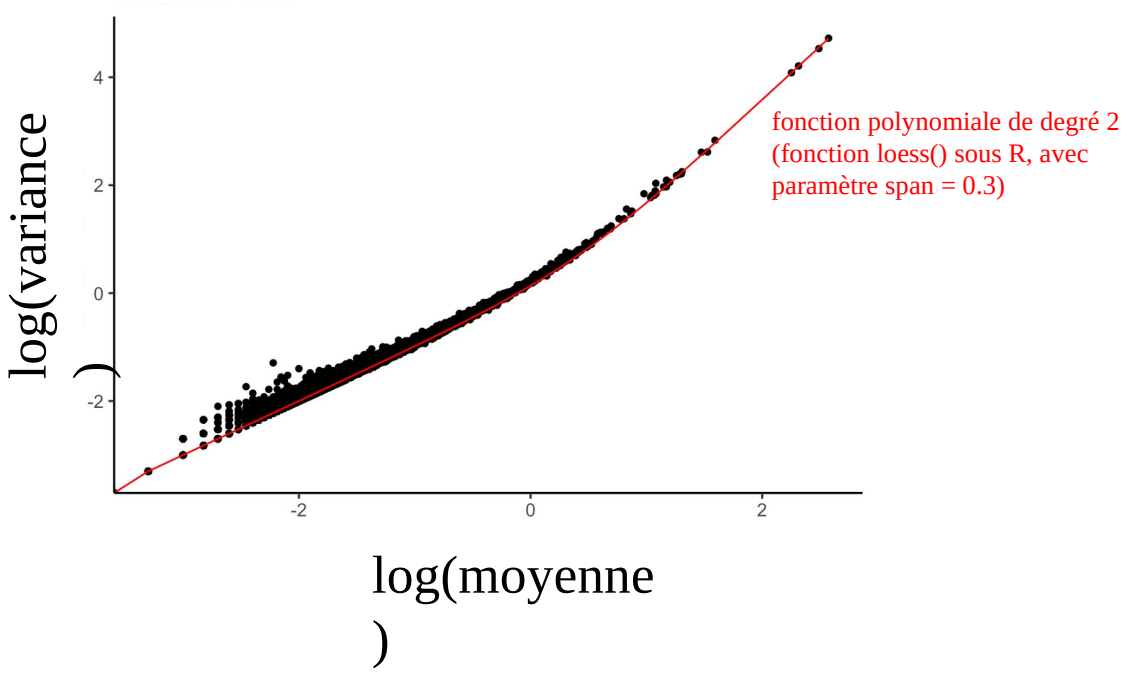
3) Trace les log(moyenne) et log(variance) sur un graphe.



HVG

→ Normalisation VST (Variance Stabilizing Transformation) :

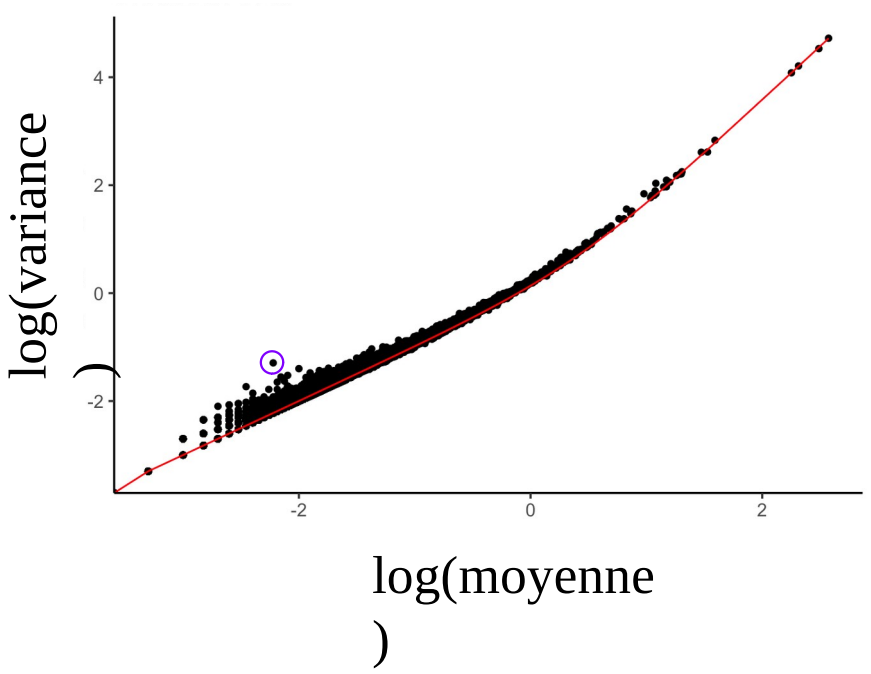
4) Identifie la courbe qui correspond le mieux au données.



HVG

→ Normalisation VST (Variance Stabilizing Transformation) :

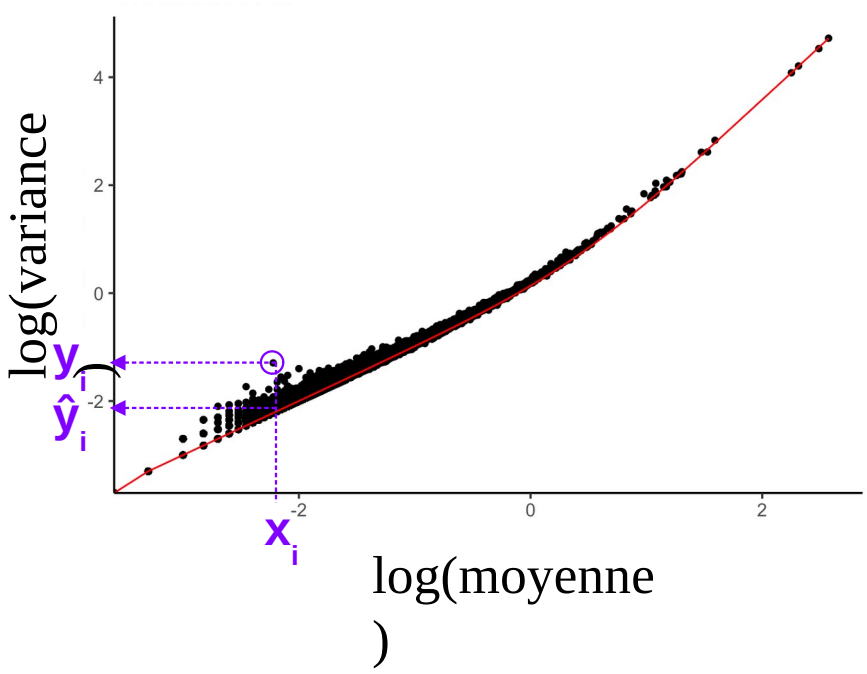
5) Estimation des variances (variances corrigées) et des écarts-types (écarts-types corrigés).



HVG

→ Normalisation VST (Variance Stabilizing Transformation) :

5) Estimation des variances (variances corrigées) et des écarts-types (écarts-types corrigés).



HVG

→ Normalisation VST (Variance Stabilizing Transformation) :

5) Estimation des variances (variances corrigées) et des écarts-types (écarts-types corrigés).

	Cellule 1	Cellule 2	Moyenne	Variance	Ecart-type	log(Moyenne)	log(Variance)
Gène A	5	10	7,5	6,25	2,5	0,88	0,8
Gène B	50	100	75	625	25	1,88	2,8

	log(Variance) corr	Variance corr	Ecart-type corr
Gène A	1,2	15,85	3,98
Gène B	3,1	1258,93	35,48

HVG

→ Normalisation VST (Variance Stabilizing Transformation) :

6) Calcul des expressions standardisées (centrées et réduites) :

$$Expression\ standardisée = \frac{(Comptage\ du\ gène\ X\ dans\ la\ cellule\ Y - Moyenne\ du\ gène\ X)}{EcartType\ estimée\ du\ gène\ X}$$

	Cellule 1	Cellule 2	Moyenne	Variance	Ecart-type	log(Moyenne)	log(Variance)
Gène A	5	10	7,5	6,25	2,5	0,88	0,8
Gène B	50	100	75	625	25	1,88	2,8

	log(Variance) corr	Variance corr	Ecart-type corr	Cellule 1 corr stand	Cellule 2 corr stand
Gène A	1,2	15,85	3,98	(5-7,5)/3,98= −0,63	(10-7,5)/3,98= 0,63
Gène B	3,1	1258,93	35,48	(50-75)/35,48= −0,7	(100-75)/35,48= 0,7

limite max = sqrt(nombre de cellules)
Ici on a 2 cellules dont sqrt(2)=1,4

HVG

→ Normalisation VST (Variance Stabilizing Transformation) :

7) Calcul des variances à partir des expressions standardisées.

	Moyenne des comptages corr stand	Variance des comptages corr stand
Gène A	(-0,63 + 0,63)/2= 0	$((-0,63-0)^2+(0,63-0)^2)/2= 0,4$
Gène B	(-0,7 + 0,7)/2= 0	$((-0,7-0)^2+(0,7-0)^2)/2= 0,49$

HVG

→ Normalisation VST (Variance Stabilizing Transformation) :

7) Calcul des variances à partir des expressions standardisées.

8) Tri des gènes selon leur nouvelle variance.

	Moyenne des comptages corr stand	Variance des comptages corr stand
Gène B	0	0,49
Gène A	0	0,4

HVG

→ Normalisation VST (Variance Stabilizing Transformation) :

7) Calcul des variances à partir des expressions standardisées.

8) Tri des gènes selon leur nouvelle variance.

9) Sélection des top 2000 gènes.

	Moyenne des comptages corr stand	Variance des comptages corr stand
Gène B	0	0,49
Gène A	0	0,4

Scaling

But :
- Rendre les gènes comparables (ce n'est plus le niveau de variance qui nous intéresse mais comment ça varie).
- Obligatoire pour faire une analyse de type PCA après.

	Cellule 1	Cellule 2	Cellule 3	Cellule 4	Cellule 5
Gène A	10	20	30	40	50
Gène B	20	40	60	80	100

Bien que le Gène B a le double d'expression que le Gène A, leur pattern d'expression est le même, et le scaling va "normaliser" leur expression pour qu'ils soit assembler dans une même composante.

Définition :
Transformation de chaque expression de gènes en z-score.
- Centre : décalage de l'expression de chaque gène, pour que la moyenne d'expression à travers les cellules vaille 0.
- Réduit : décalage de l'expression de chaque gène, pour que la variance à travers les cellules vaille 1. Cette étape donne un poids égal dans les analyses en aval, de sorte que les gènes hautement exprimés ne dominent pas.

Scaling

Les données à scaler dépendent de si vous souhaitez faire une correction de biais (pourcentage d'ARN mitochondriaux, cycle cellulaire, etc) ou pas.

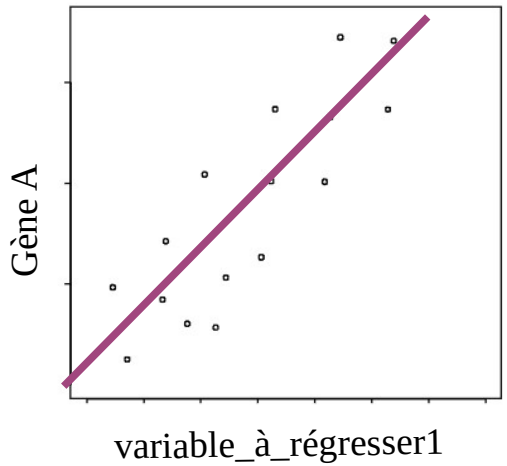
Si pas de régression à faire :
→ Centre et réduit les expressions lognormalisées des HVG (qui seront utilisés dans la PCA) :

$$\frac{\text{Expression du gène A dans la cellule Y} - \text{Moyenne d'expression du gène A}}{\text{Ecart-type d'expression du gène A}}$$

Scaling

Si régression à faire :
Pour chaque gène HVG lognormalisé :
1) Calcul d'un modèle linéaire multiple:

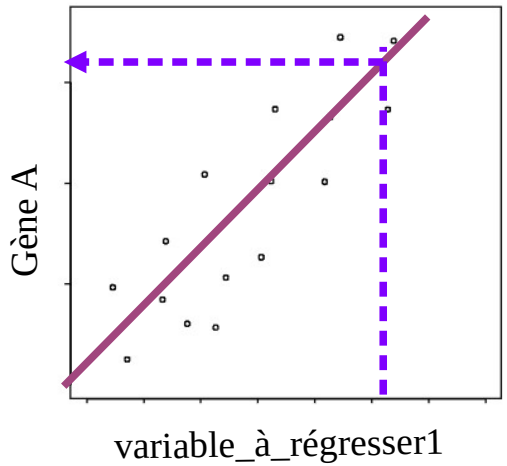
GENE ~ variable_à_régresser1 + variable_à_régresser2 + …



Gène A = β_1 variable_à_régresser1 + β_0

Scaling

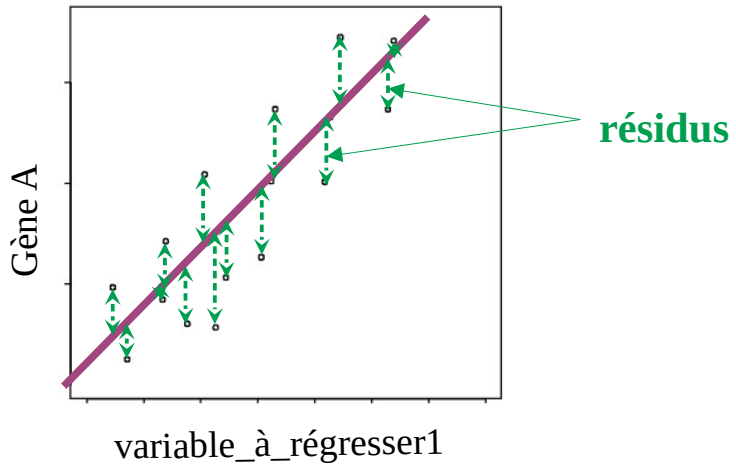
2) Calcul des expressions lognormalisées corrigées (régressées), ainsi que la moyenne, la variance et l'écart-type de ces nouvelles expressions.



Scaling

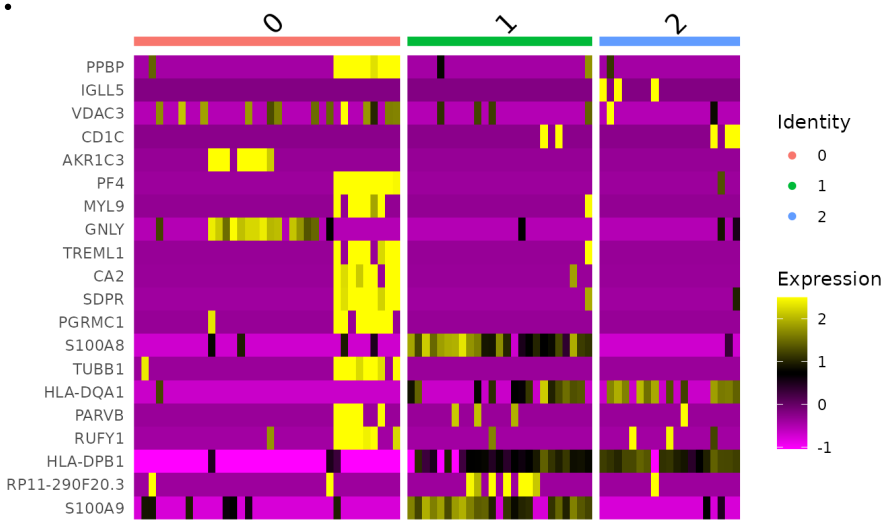
3) Calcul des résidus de Pearson (utilisés dans la PCA):

$$\text{Résidus de Pearson} = \frac{\text{Expression initiale du gène A dans la cellule Y} - \text{Nouvelle expression du gène A dans la cellule Y}}{\text{Nouvel écart-type du gène A}}$$



Scaling

Les résidus de Pearson sont très utilisés pour tracer l'expression des gènes sur une heatmap afin de comparer l'expression entre des gènes.
Si les gènes d'intérêt ne sont pas dans les HVG, leurs résidus de Pearson ne seront pas calculé.
Mais vous pouvez recalculer ces résidus à posteriori pour faire les graphes en stipulant les gènes à calculer dans la fonction ScaleData().

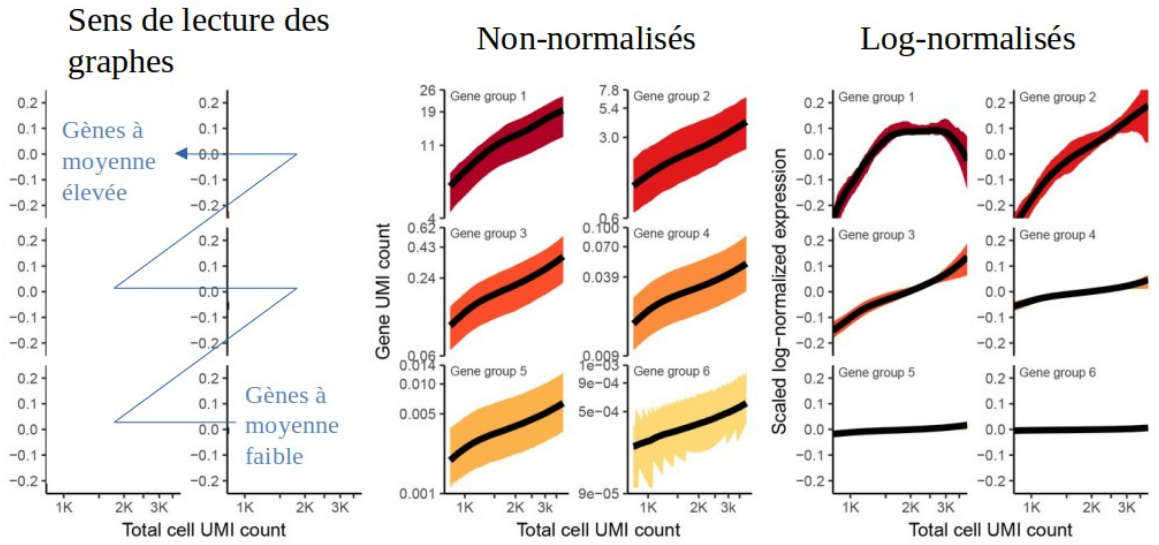


SCTransform

1) Échantillonnage de 5000 cellules (ou toutes les cellules s'il y en a moins dans l'échantillon).

2) Calculs de la moyenne géométrique d'expression de chaque gène.

3) Regroupement des gènes selon leur expression moyenne.

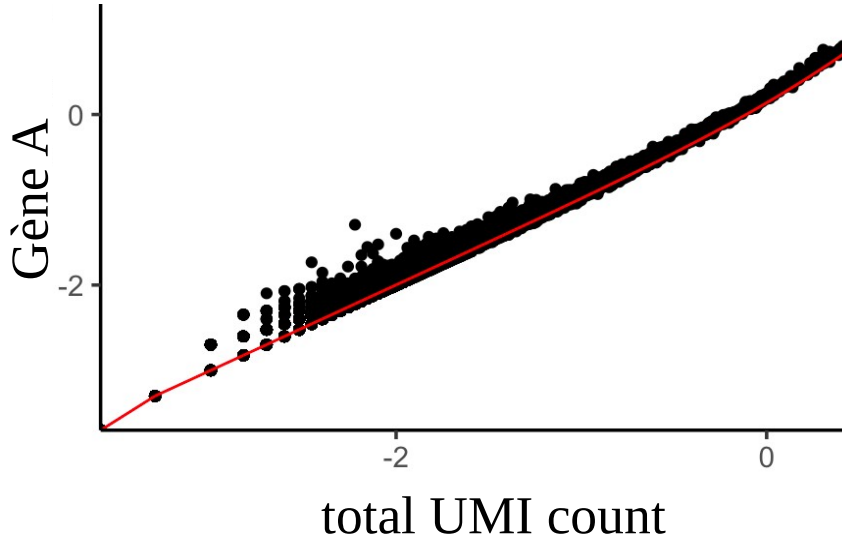


But : appliquer une normalisation différente selon la moyenne d'expression des gènes.

SCTransform

Pour chaque gène dans chaque groupe de gènes :

4) Traçage d'un graphe : GENE ~ total UMI count



Disclaimer : Les graphes ne représentent pas des données réelles et permettent uniquement d'illustrer le propos.

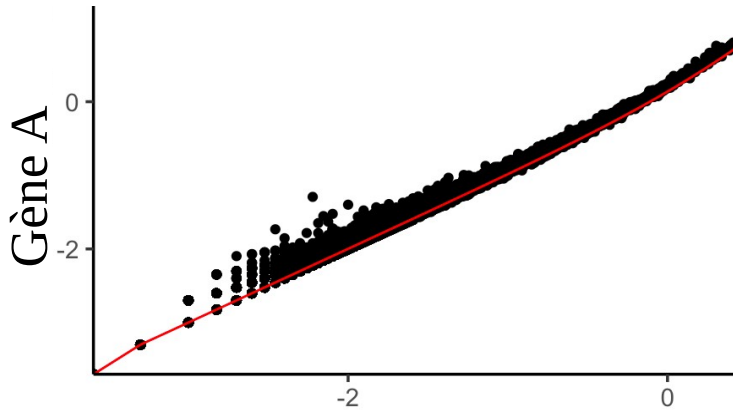
SCTransform

5) Calcul de la courbe de régression binomiale négative et identification des paramètres du modèle (intercept β_0, slope β_1, et negative binomial dispersion θ).

total UMI count

moyenne(expression_gene) = $\mu_{ij} = \exp(\beta_{0_i} + \beta_{1_i} \log_{10} m_j)$,

écart-type(expression_gene) = $\sigma_{ij} = \sqrt{\mu_{ij} + \frac{\mu_{ij}^2}{\theta_i}}$,



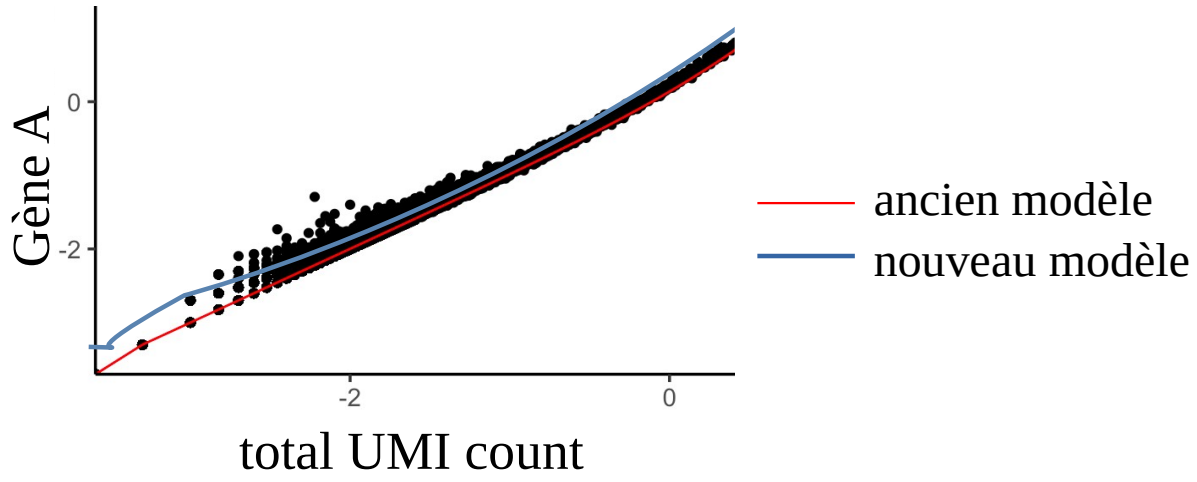
SCTransform

Pour chaque groupe de gènes :

6) Mise en commun des paramètres identifiés pour chaque gène pour créer des paramètres corrigés (régularisés) pour éviter l'overfitting, car sinon ça corrige tellement la variance que ça supprime la variabilité du signal biologique d'intérêt.

Pour chaque gène dans chaque groupe de gènes :

7) Les nouveaux paramètres sont utilisés pour recalculer une nouvelle courbe de modèle binomial négatif.

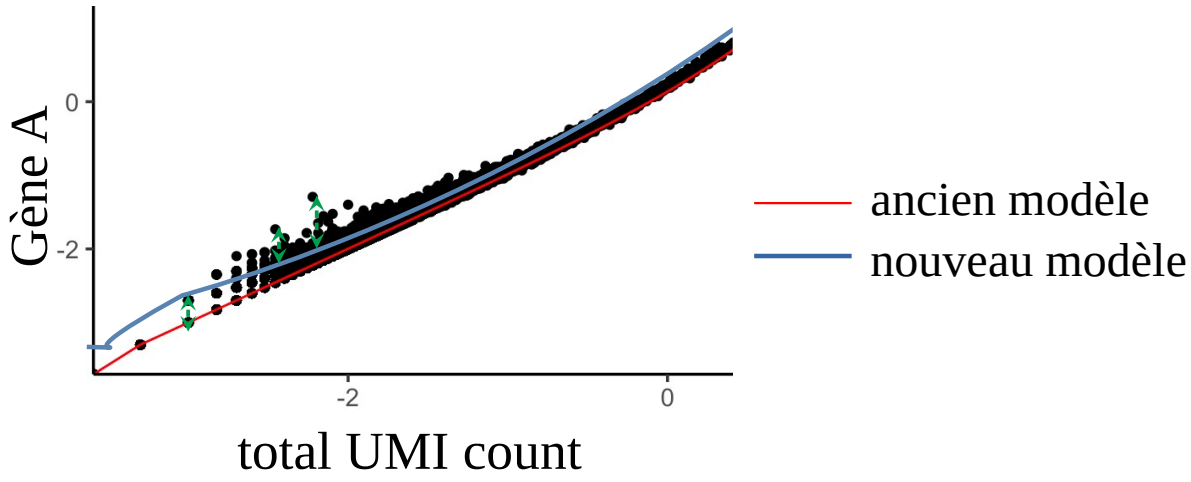


SCTransform

8) Calcul des nouvelles expressions, de leur moyenne, variance et écart-type.

9) Calcul des résidus de Pearson entre les observations initiales et le nouveau modèle:

résidus_pearson(gene) = $z_{ij} = \frac{x_{ij} - \mu_{ij}}{\sigma_{ij}}$,



SCTransform

résidus_pearson(gene) = $z_{ij} = \frac{x_{ij} - \mu_{ij}}{\sigma_{ij}},$

10) Calcul la variance des résidus
11) Tri les résidus selon leur variance, récupère le top 3000 (HVG), filtre la table de résidus sur ces 3000 gènes.

10) Calcul des log des expressions corrigées obtenues à l'étape 8.

SCTransform

Si régression des biais supplémentaires (pourcentage d'ARN mitochondriaux, cycle cellulaire, etc):

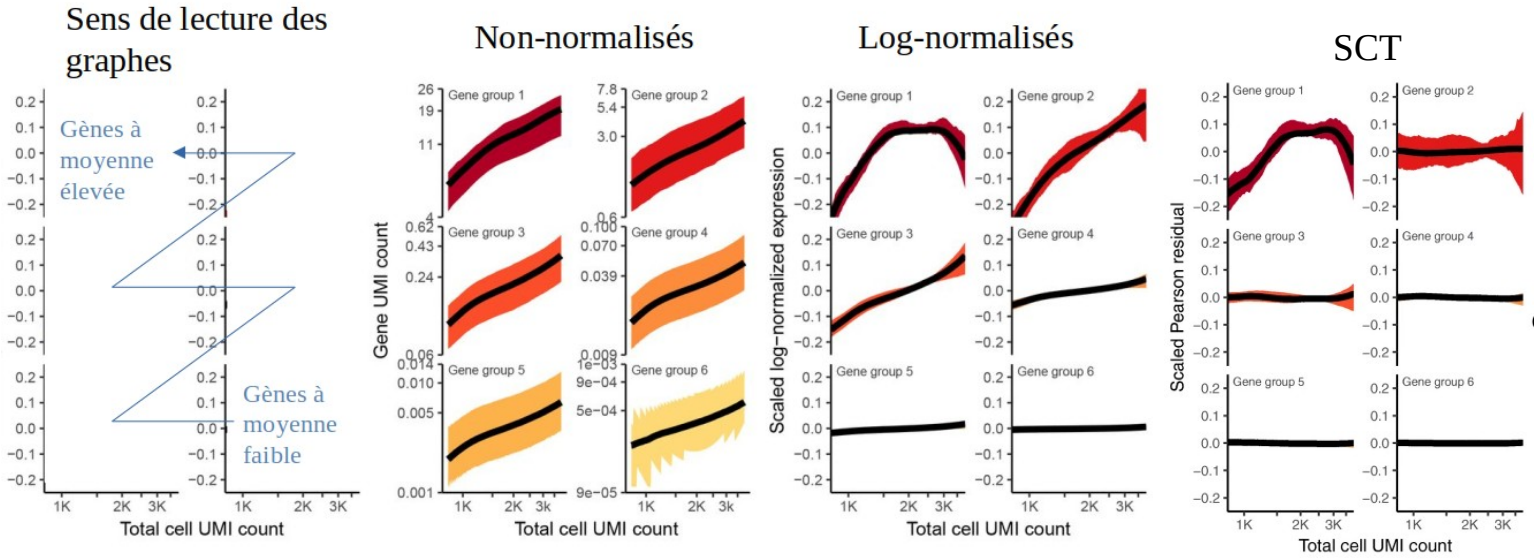
1) à 6) identiques.

7) Les nouveaux paramètres régularisés sont utilisés pour recalculer une nouvelle courbe de modèle binomial négatif, dans lequel on rajoute les nouveaux biais. Donc il faudrait juste identifier les paramètres liés aux nouveaux biais ajoutés.

$$\underbrace{\text{moyenne(expression\_gene)) = exp(} \beta_0 + \beta_1 \log_{10} m_j}_{\text{modèle initial}} \textbf{ + } \beta_2 \textbf{ biais\_à\_régresser1 + } \beta_3 \textbf{ biais\_à\_régresser2 + ...)}$$

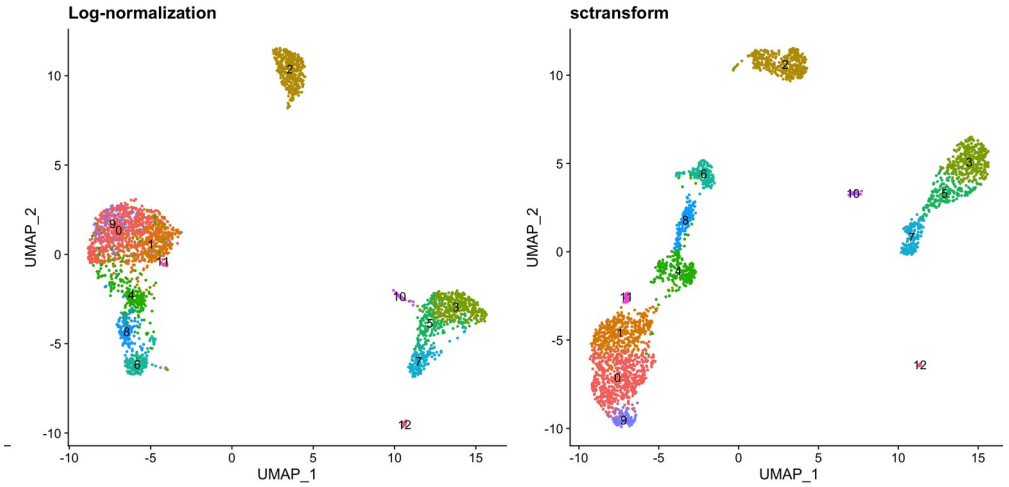
10) à 11) identiques.

SCTransform

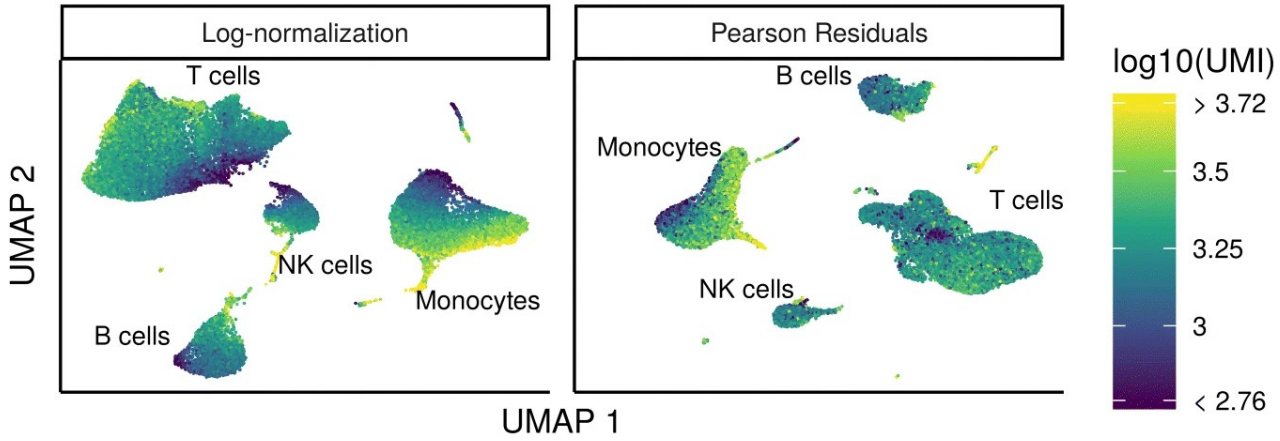


La méthode SCTransform permet de mieux décorréler l'expression avec la profondeur de séquençage.
Mais les gènes avec une moyenne d'expression très élevée restent corrélés.

SCTransform



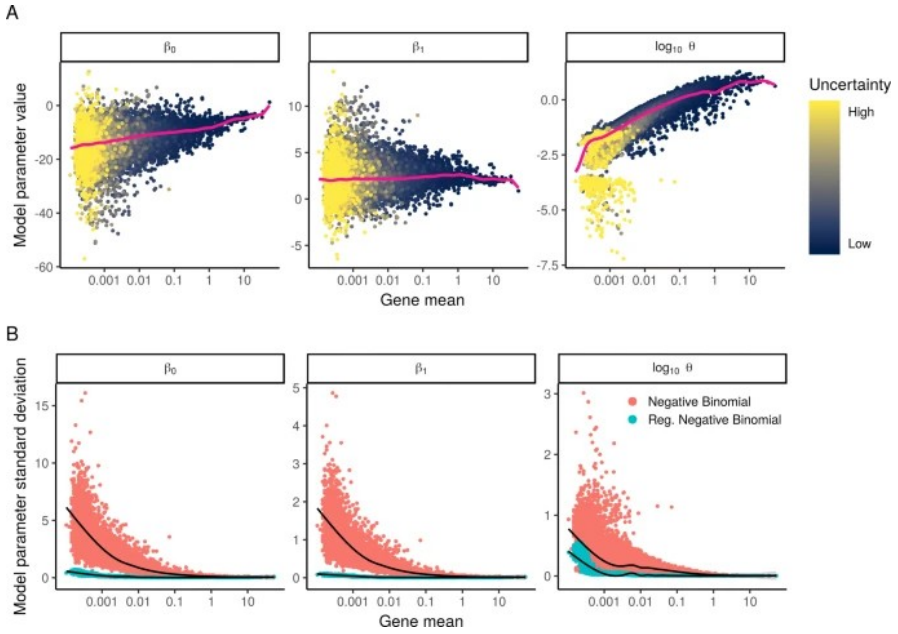
Avantages du SCT:
1) Pas besoin de fixer un size factor.
2) La régularisation des paramètres minimises l'overfitting des données.
3) Les résidus de Pearson sont indépendent de la profondeur de séquençage et peuvent être utilisés pour la sélection des gènes hautements variables ainsi que pour la PCA.



SCTransform V2

2019 : SCTransform
Christoph Hafemeister and Rahul Satija (2019)

2021 : Comparaison de SCTransform vs VST vs modèle GLM-PCA
Jan Lause, Philipp Berens & Dmitry Kobak (2021)



→ l'estimation des paramètres du modèle binomial négatif de SCTransform est bruitée sans régularisation.
→ β_0 et β_1 sont fortement corrélés, donc on peut donner une valeur par défaut à β_1 pour simplifier le modèle et éviter la régularisation des paramètres.
→ le modèle simplifié correspond à un cas particulier du modèle GLM-PCA.
→ discussion sur la valeur de θ.
→ ...

SCTransform V2

2019 : SCTransform
Christoph Hafemeister and Rahul Satija (2019)

2021 : Comparaison de SCTransform vs VST vs modèle GLM-PCA
Jan Lause, Philipp Berens & Dmitry Kobak (2021)

2022 : SCTransform V2
Saket Choudhary and Rahul Satija (2022)

→ discussion sur la valeur de θ.

→ simplification du modèle binomial négatif en suivant les recommandations de Lause et al (2021) mais la régularisation est conservée car elle permet de modéliser le niveau de bruit intrinsèque (turn-over des ARN).

→ les gènes faiblement exprimés ne sont plus défini par un modèle binomial négatif mais par un modèle de Poisson et ne sont plus utilisés dans la régularisation.

→ et avec moins de cellules en entrée, ça fonctionne aussi bien (SCTransform v2 = 2000 cellules pour l'étape 1).

→ ...

Normalisation

Normalisation en général:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5549838/
https://www.nature.com/articles/s41592-023-01814-1
https://divingintogeneticsandgenomics.com/post/negative-bionomial-distribution-in-single-cell-rnaseq/

Seurat: LogNormalisation + HVG + Scaling:
https://www.cell.com/cell/fulltext/S0092-8674(19)30559-8

VST :
https://htmlpreview.github.io/?https://github.com/satijalab/sctransform/blob/supp_html/supplement/variance_stabilizing_transformation.html

Régression Linéaire :
https://fermin.perso.math.cnrs.fr/Files/Chap3.pdf

SCTransform :
https://htmlpreview.github.io/?https://github.com/satijalab/sctransform/blob/supp_html/supplement/seurat.html
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1

Comparaison de SCTransform vs VST vs modèle GLM-PCA :
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02451-7

SCTransform v2 :
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02584-9

Merci de votre attention

Merci de votre attention