Master 1 - Statistique et Sciences des Données 2019-2020

Rapport du Projet tutoré

Modélisation de l'expression des gènes tissu-spécifiques avec une distribution “Zero-inflated”

Tuteur

Florent CHUFFART

Membres du groupe de projet

Marie-Anne LIEU - Thi Thuy Anh DOAN - Anais ROSSETTO - Kodjo Mawuena AMEKOE

Table des matières

Remerciements **4**

Introduction **5**

1. Analyse descriptive **6**

1.1. Les données 6

1.2. Préparation des données 6

1.3. Co-facteurs 7

1.4. Analyses sur les données brutes 7

1.5. Analyses sur les données log-transformées 8

1.5. Analyses des gènes OLFA 9

▸Heatmap 10

1.6. Quatre classes de gènes 13

1.7. Études des cofacteurs 14

▸Répartition en fonction du sexe 15

▸Répartition en fonction du tissu 15

2. Présentation des modèles **19**

2.1. Loi Normale 19

2.2. Loi de Poisson 19

2.3. Loi Négative Binomiale 20

2.4. Loi Zero Inflated de Poisson 20

2.5. Loi Zero Inflated Negative Binomiale 21

2.6. Loi de Bernoulli 22

2.7. Qualité d'ajustement, tests et choix entre différents modèles 22

2.7.1. Qualité d'ajustement 22

2.7.2. Test du rapport de vraisemblance 23

2.7.3. Autres critères 23

3. Modélisation **24**

3.1. "mise en contexte" 24

3.2. Outils utilisés 24

3.3. Modélisation à l'aide d'une loi Normale 25

◆ Procédure utilisée 25

◆ Modélisation de la loi normale sur les quatre classes de gènes 25

◆ Modèle avec tissu : anova 27

3.3. Modélisation par la loi de Poisson 28

◆ Procédure utilisée 28

▸La méthode des moments 28

▸La méthode du maximum de vraisemblance 28

◆ Résultats de la modélisation de poisson 29

◆ Modélisation de poisson sur les quatre classes de gènes 30
◆ Influence du tissus 31
3.4. Modélisation par la loi Négative Binomiale 31
◆ Procédure utilisée 31
◆ Comparaison des deux méthodes 32
▸ Résultats de la méthode des moments 32
▸ Résultats par la méthode du maximum de vraisemblance 32
◆ Modélisation NB sur les quatre classes de gènes 33
◆ Influence du tissu 34
3.5.Modélisation par une loi Zero Inflated de Poisson 34
◆ Procédure utilisée 34
▸ Estimation par la méthode des moments 35
▸ Estimation par la méthode du maximum de vraisemblance 35
◆ Résultats 35
◆ Modélisation ZIP sur les quatre classes de gènes 37
◆ Influence du tissu 38
3.6. Modélisation par une loi Zero Inflated Négative Binomiale 38
◆Procédure utilisée 39
◆ Résultats 39
◆ Modélisation ZINB sur les quatre classes de gènes 40
◆ Influence du tissu 41
3.7 Modélisation par une loi de Bernoulli 41
◆Procédure utilisée 41
▸Méthode des moments 41
▸Méthode du maximum de vraisemblance 42
◆ Modélisation de Bernoulli sur les quatre classes de gènes 43
◆ Caractéristiques des gènes OLFA 44

4. Comparaison des modélisations 45
4.1. Evaluation de la modélisation de la proportion de zéro 45
4.2. Évaluation à l'aide du rapport de vraisemblance 47
◆ Négative Binomiale vs Poisson 47
◆ ZIP vs ZINB 47
4.3. Évaluation à l'aide du test de Vuong 47
◆ ZIP vs Poisson 48
◆ ZINB vs NB 48
4.4. AIC et proportion de zéro 49

Conclusion du projet 51

BIBLIOGRAPHIE 52

ANNEXES 54

Remerciements

Au terme de ce travail, nous tenons à exprimer notre profonde gratitude à notre professeur et encadrant Monsieur Florent CHUFFART pour son suivi, et son énorme soutien, qu'il n'a cessé de nous prodiguer tout au long du projet.
Nous remercions également Madame Adeline LECLERQ SAMSON Responsable du Master pour son soutien et ses directives.
Nous adressons aussi nos vifs remerciements pour Monsieur Rémy DROUILHET Co-Responsable du Master et chargé du module de programmation en R, pour nous avoir apporté les bases de cet outil.

Introduction

Le mot ectopique a pour préfixe, “ecto”, qui signifie qu’il n’est pas à sa place habituelle. Une expression ectopique est une anomalie de l’expression d’un gène dans un type de cellule, un tissu, ou une étape de développement dans lequel le gène n’est généralement pas exprimé.

Cette définition nous aide pour la compréhension de la thématique dans lequel s'inscrit notre projet. L’équipe epimed[1] du laboratoire de l’IAB[2] mène une étude qui porte sur l’expression ectopique de gènes dans les cellules tumorales. Ces gènes ectopiques apportent de l’information quant au type de cancer et leur degré d'agressivité.

Plus particulièrement, le sujet porte sur une classe spécifique de gènes: la famille des gènes des récepteurs olfactifs. Ce sont des gènes qui participent au système olfactif et qui sont exprimés uniquement dans les tissus olfactifs. Ceux-ci ont la particularité d’être on ou off, c’est à dire qu’ils vont s’exprimer que dans certains tissus et être non exprimés dans beaucoup d’autres. Cette particularité intéresse fortement les chercheurs. De plus, ces gènes ont été détectés actifs dans des tissus où ils ne devraient normalement pas s’exprimer. Et c’est la raison pour laquelle, ils portent un intérêt pour ces gènes qui peuvent être porteurs d’information dans le cadre de la recherche contre le cancer.

Notre projet est donc de chercher à caractériser cette expression si particulière.

Pour mieux comprendre, intéressons-nous dans un premier temps à l’analyse du jeu de données qui nous a été fourni. Cela nous permettra de mieux caractériser notre sujet en termes de modélisation statistique.

[1] EPIMED: EPIgénétique MEDicale et Bioinformatique.

[2] L'IAB (Institute for Advanced Biosciences) est un institut de renommée internationale dans la recherche biomédicale fondamentale et translationnelle basée à Grenoble.

1. Analyse descriptive

1.1. Les données

Les données sont issues de la base GTEX [3] Genotype-Tissue Expression. Les 2753 prélèvements de tissus ont été fait sur presque 1000 individus et ont porté sur 30 organes différents. Pour chacun de ces prélèvements, on a mesuré le nombre de fois où l'ADN est transcrit en ARN. C'est cet ARN qui a été découpé et séquencé et qui a permis de savoir combien de fois on a un read pour un gène donné. Pour chaque échantillon, on a mesuré cette quantité pour chacun des 22904 gènes constitutif du génome humain. Ainsi la matrice de données à explorer est constituée des 2753 échantillons par 22904 gènes. Les 384 gènes *olfa* qui nous intéressent représente une sous-famille de gènes.

1.2. Préparation des données

On charge les données fournies au début du projet et on produit :

- 4 matrices, nommées data_raw, data_norm , data_zero et data_lnorm, au format gènes (lignes) x échantillons (colonnes), ce format est imposé par le projet.

df_raw correspond aux données brutes.

df_zero est une normalisation des données brutes. Les données sont transformées en un si elles ont une valeur différente de zéro sinon on conserve le zéro. Ainsi ce dataframe ne contient que des zéros et des uns.

df_norm ce sont les données normalisées par la méthode de normalisation deseq2. Il s'agit d'une normalisation et recherche de gènes différentiellement exprimés en se basant sur un modèle de distribution binomiale négative.

df_lnorm pour lequel on utilise les données précédemment normalisées, on leur ajoute un et on applique le log2.

- 4 dataframes, nommées df_raw, df_zero, df_norm et df_lnorm au format échantillons (lignes) x gènes (colonnes), nous manipulons ce format qui correspond plus à ce que nous avons l'habitude de voir en statistiques (un individu par ligne).

Le jeu de données comporte 22904 gènes et 2753 échantillons.

On remarque que les différentes transformations apportées aux données n'ont ni ajouté, ni retiré, ni déplacé de zéro.

[3] https://gtexportal.org/home/

1.3. Co-facteurs

Plusieurs co-facteurs enrichissent les données, notamment:

- *sex* : indique le sexe de la personne de laquelle l'échantillon est issu
- *tissue_group_level1* : 31 groupes qui représentent 31 types de tissus (ex: cerveau, foie, …)
- *tissue_group_level2* : 17 groupes de tissus (niveau supérieur des types de tissus, qui englobe le précédent)
- *tissue_group_level3* : 2 groupes de tissus (niveau supérieur des types de tissus, qui englobe les autres, constitué des cellules germinales ou somatiques)
- *organism_part* : 30 organes (ce sont les organes desquels sont extraits les prélèvements de tissu)

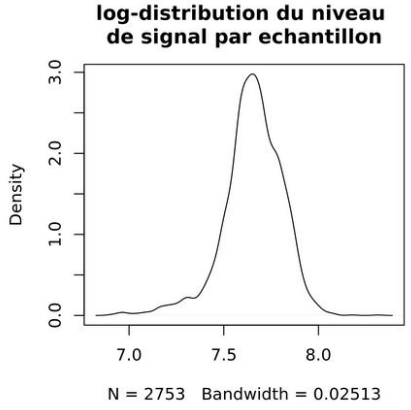
Nous avons fait le choix d'utiliser uniquement les deux co-facteurs suivants : le sexe et le groupe de tissu nommé *tissue_group_level1*. Les variables *tissue_group_level2* et *tissue_group_level3* étant des parties dans lesquelles sont imbriqués le groupe 1. La variable *organism_part* est très similaire à *tissue_group_level1* avec quelques petites différences. Mais, il est commun dans ce type d'analyse d'utiliser *tissue_group_level1*.

1.4. Analyses sur les données brutes

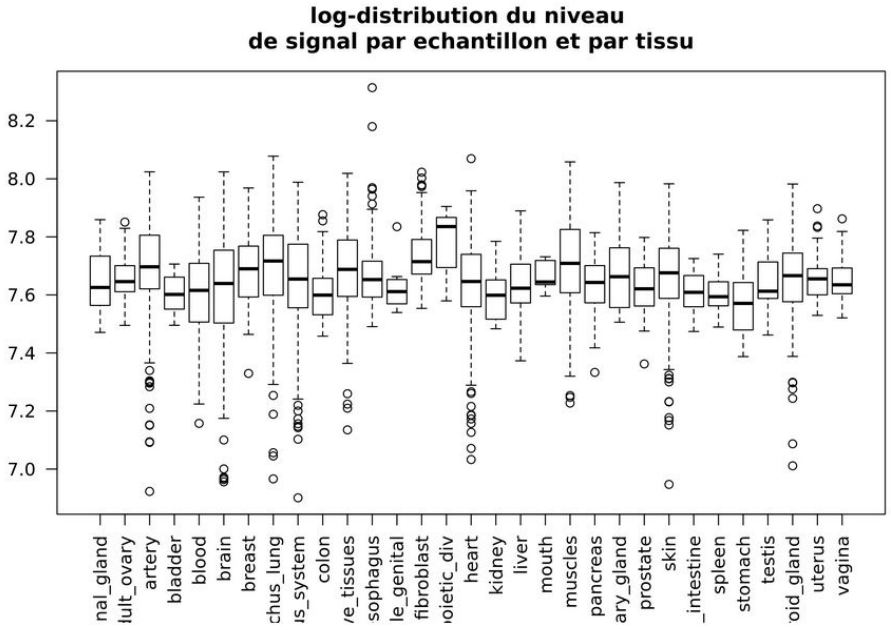
Les données brutes sont présentées dans une dataframe de 2753 échantillons (lignes) et 22904 gènes (colonnes).

Pour chaque échantillon, on calcule le nombre total de read. L'objectif étant de détecter les échantillons avec un faible niveau de signal (une faible couverture en termes de nombre de reads et potentiellement un grand nombre de zéro "techniques").

D'après le graphique ci-dessous, on peut voir que la couverture des échantillons est à la fois bonne (10^7 reads) et homogène (un facteur de 10 entre le plus et le moins couvert).



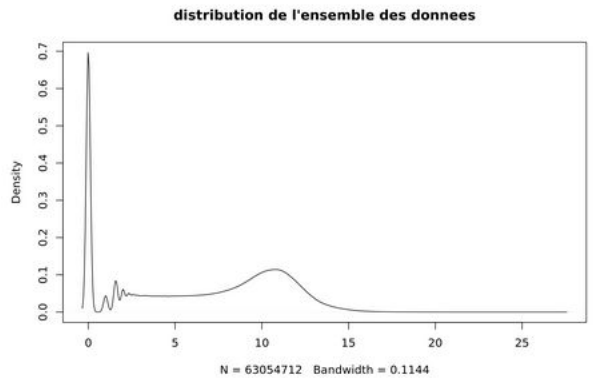
Intéressons nous maintenant à la représentation du nombre total de read par échantillon et par tissu.

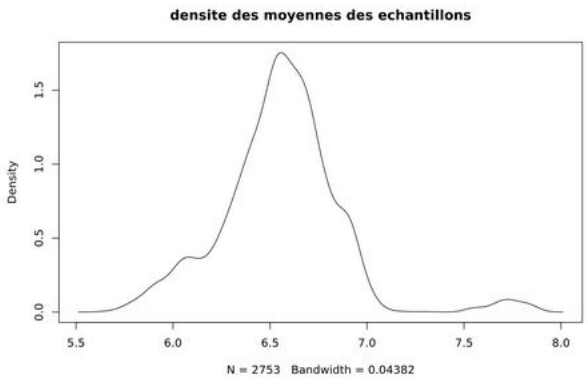


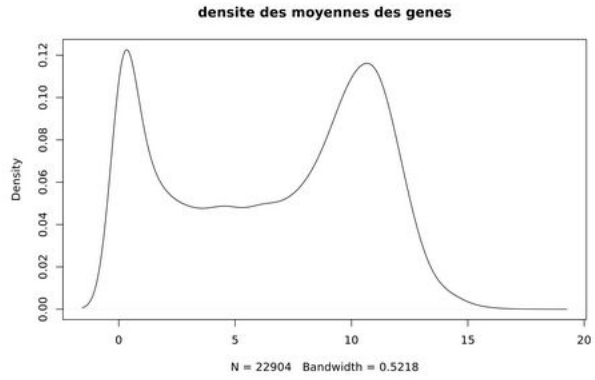
La couverture est également bonne et homogène.

1.5. Analyses sur les données log-transformées

Les graphiques suivants, issus du dataframe des données pseudo log transformées, permettent de mieux comprendre le jeu de données.







Le premier graphique, affiche la densité de l'ensemble des valeurs de la matrice. Ainsi, on peut voir que les données comportent beaucoup de zéro et qu'il y a une forte dispersion.

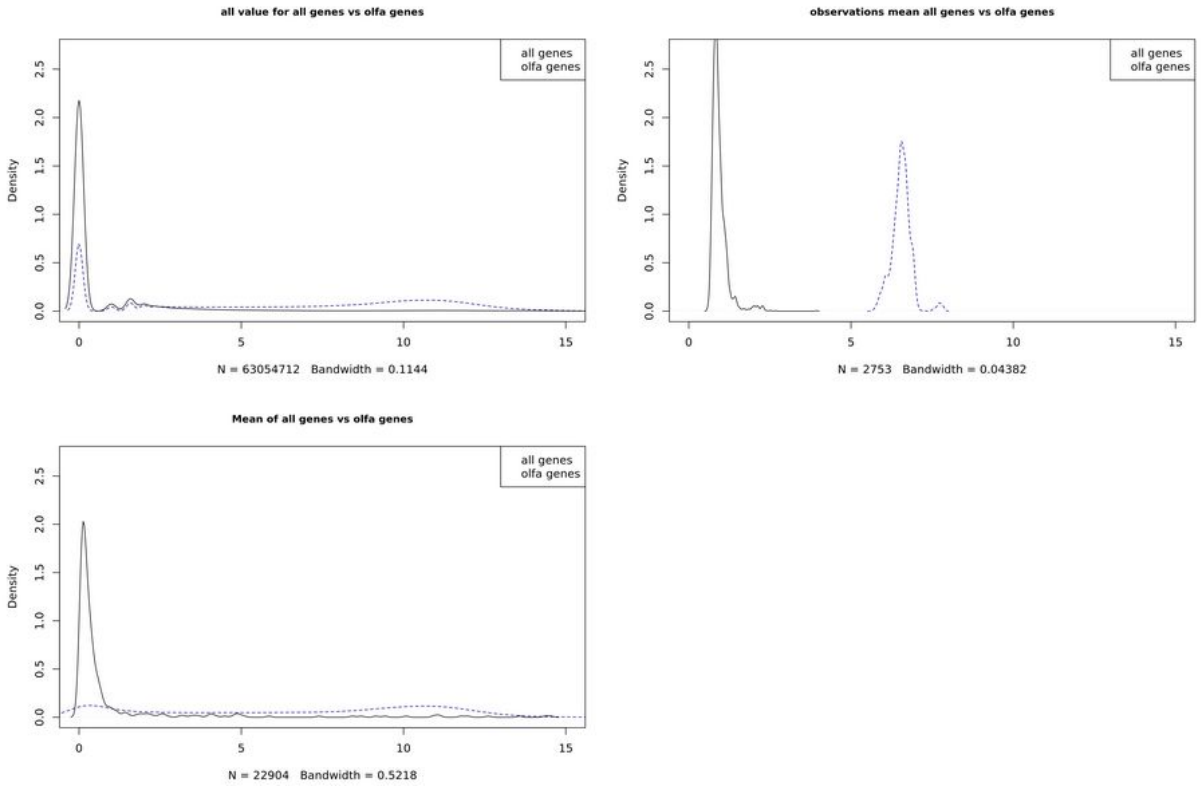
Le second graphique, provient des moyennes effectuées par échantillon, ainsi on peut voir que la distribution prend une forme gaussienne centrée autour de 6.5. Les échantillons, sont donc équivalents et tous exploitables.

Le troisième graphique, affiche la densité des moyennes par gènes. On observe qu'il y a deux valeurs autour desquelles les données se concentrent. Un premier pic en zéro et un second autour de 10. Il y a donc une très forte dispersion dans l'expression des gènes. Certains ont une valeur moyenne très faible alors que d'autres très élevés.

1.5. Analyses des gènes *OLFA*

Les gènes *OLFA* sont les 384 gènes qui nous intéressent et qui ont la particularité de s'exprimer uniquement dans certains tissus. De part, leur caractéristique on/off on s'attend à observer de nombreuses valeurs nulles.

Les trois graphiques ci-dessous, permettent de comparer les distributions globales et marginales (par ligne et par colonne) de tous les gènes et des gènes *OLFA*.



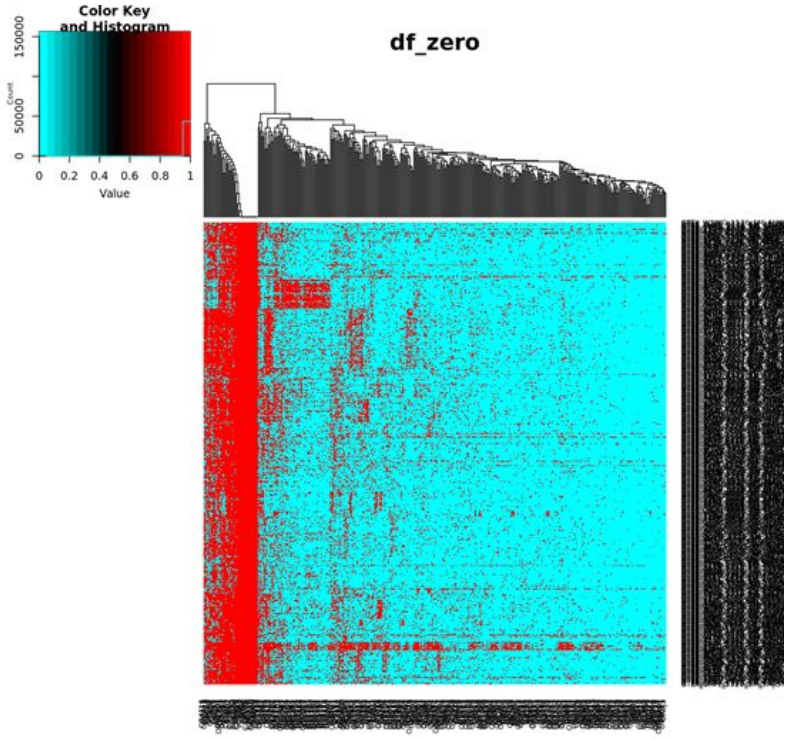
Le premier graphique, représente l'ensemble des valeurs des deux matrices. On peut voir qu'en zéro il y a un pic d'observations deux fois plus important pour les gènes *olfa*. De plus, on peut voir que l'ensemble des gènes présentent des observations autour de 11 alors que les gènes *olfa* semblent plutôt proches de zéro sur ces valeurs.

Le second graphique, affiche la densité des moyennes par individu. On peut voir que les valeurs des gènes olfa sont plus faibles que pour l'ensemble des gènes. Caractérisés par une expression dans quelques tissus seulement, il n'est donc pas surprenant de constater que la moyenne par échantillon est plus faible lorsque l'on s'intéresse uniquement aux gènes *olfa.*

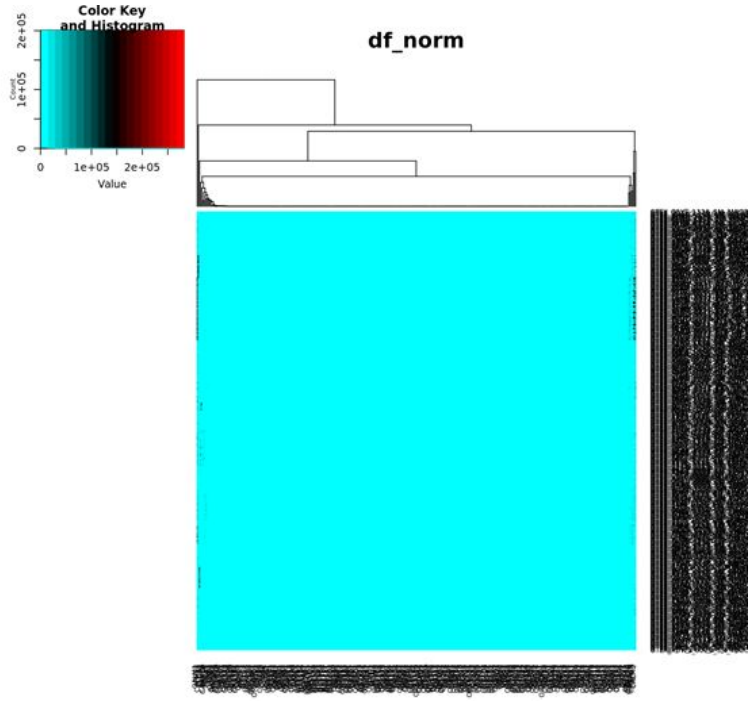
Le troisième graphique affiche la densité des moyennes par gènes. On peut voir qu'il y a un pic autour des valeurs qui sont proches de zéro pour les gènes *olfa* alors que ce pic n'existe pas pour l'ensemble des gènes. Influencés par la forte présence de zéros, il n'est pas étonnant de constater que la moyenne de ces gènes soit très faible. Néanmoins, on observe une queue de distribution qui laisse penser que quelques gènes *olfa* pourraient avoir un comportement similaire aux gènes quelconques. Ainsi au sein du groupe des gènes *olfa*, il y a peut être des gènes qui ne portent pas cette particularité tissu-spécifique.

▸Heatmap

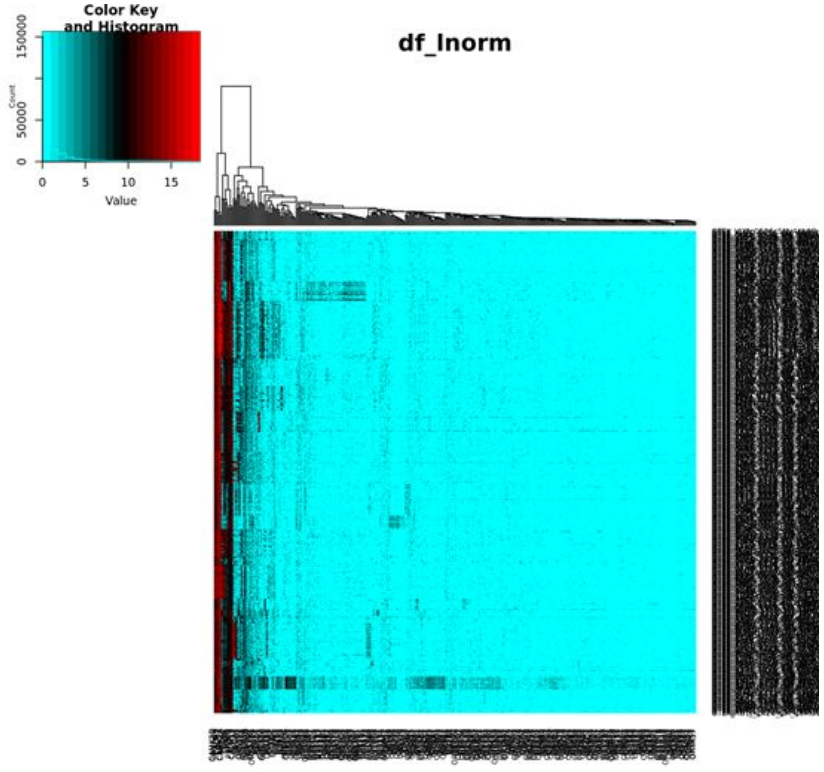
Les trois heatmap présentées ci-après utilisent uniquement les données concernant les gènes *olfa*.



Cette représentation, issue du jeu de données normalisées zéro/pas zéros permet de bien visualiser la forte présence de zéro sur l'ensemble du jeu de données. Les colonnes qui sont colorées en bleu représentent les gènes qui sont riches en zéros. Il y a aussi des gènes pour lesquels il y a très peu de zéros, ce sont les colonnes qui sont colorées de rouge majoritaire. Comme nous le supposions précédemment, ces gènes ne semblent pas être tissu-spécifique comme l'ensemble des autres gènes de cette famille.



La normalisation classique est difficile à représenter (sur-dispersion), il convient généralement d'appliquer un pseudo log2 les données.



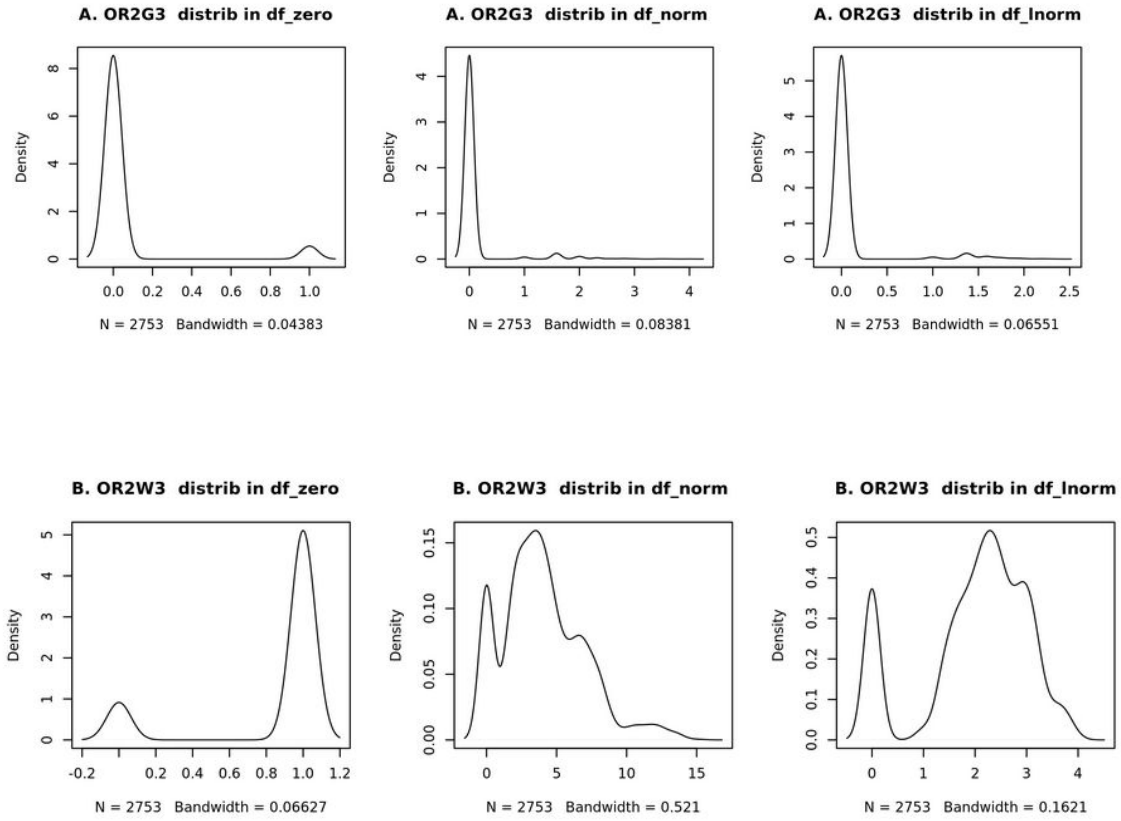
La heatmap ci-dessus est élaborée à partir des données normalisées log transformées et on peut voir que l'on obtient les même résultats qu'avec la normalisation zéro/pas zéro.

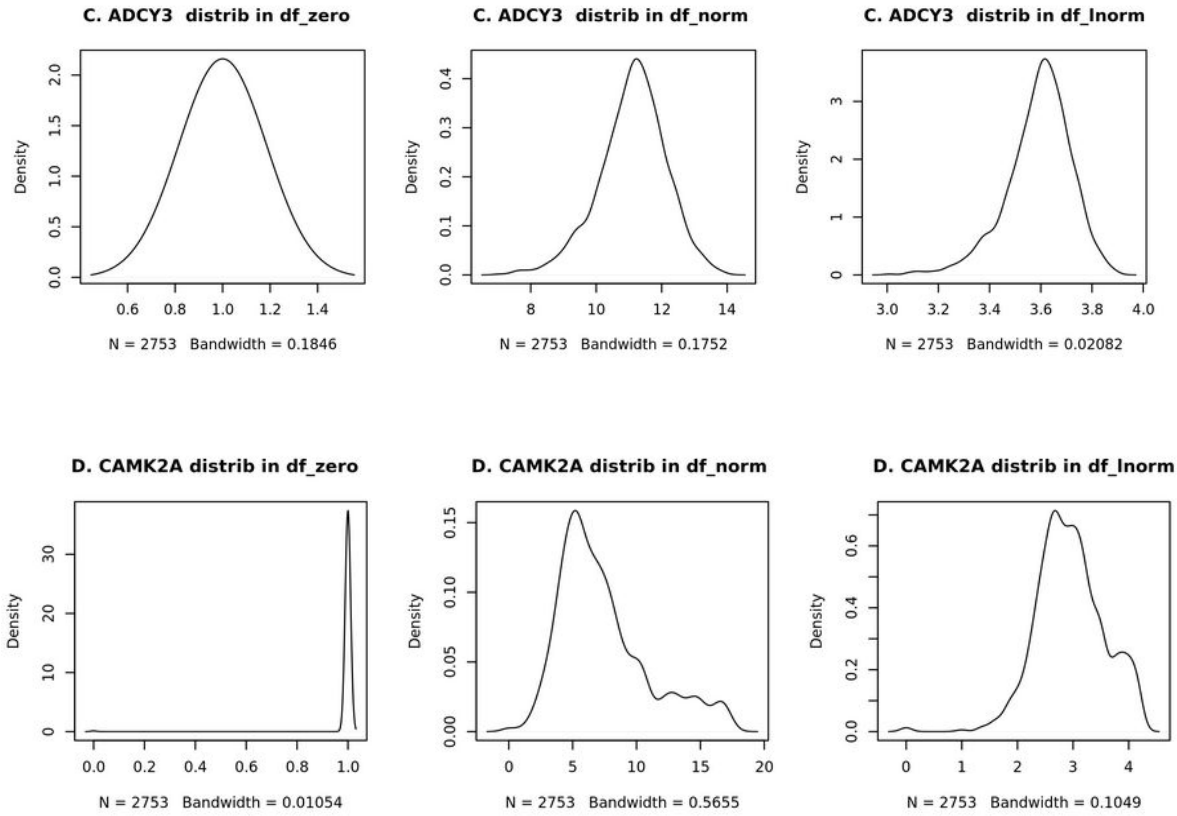
1.6. Quatre classes de gènes

Lors de l'observation des distributions des 384 gènes *olfa*, nous avons pu déceler quatre types de distributions. Nous avons sélectionné quatre gènes représentatifs de chacunes de ces distributions:

- **OR2G3** de la classe des gènes riches en zéro avec peu de dispersion sur le reste des données (ci-dessous figure A). Ce type de distribution est majoritaire dans le jeu de données restreint aux gènes o*lfa* .
- **OR2W3** de la classe des gènes riches en zéro avec beaucoup de dispersions ensuite (figure B)
- **ADCY3** de la classe des gènes avec absence de zéro et peu de dispersion (figure C)
- **CAMK2A** de la classe des gènes avec très peu de zéro et beaucoup de dispersion (figure D)

Les graphiques suivants, affichent la représentation des distributions de ces quatre gènes réalisés à l'aide des trois jeu de données suivants: df_zero, df_norm, df_lnorm. Les données brutes ne sont pas utilisées car il y a trop de dispersion.





Ces quatre types de distributions nous font penser aux lois suivantes :

- Bernoulli
- Normale
- Poisson
- Négative Binomiale
- ZIP : Zero Inflated Poisson
- ZINB : Zero inflated Negative Binomiale

D'après cette analyse, nous avons conclu qu'il ne nous serait pas possible d'utiliser un seul modèle pour l'ensemble de nos données.

1.7. Études des cofacteurs

A l'aide de l'Annexe 1, on peut voir que la répartition des données n'est pas égalitaire.

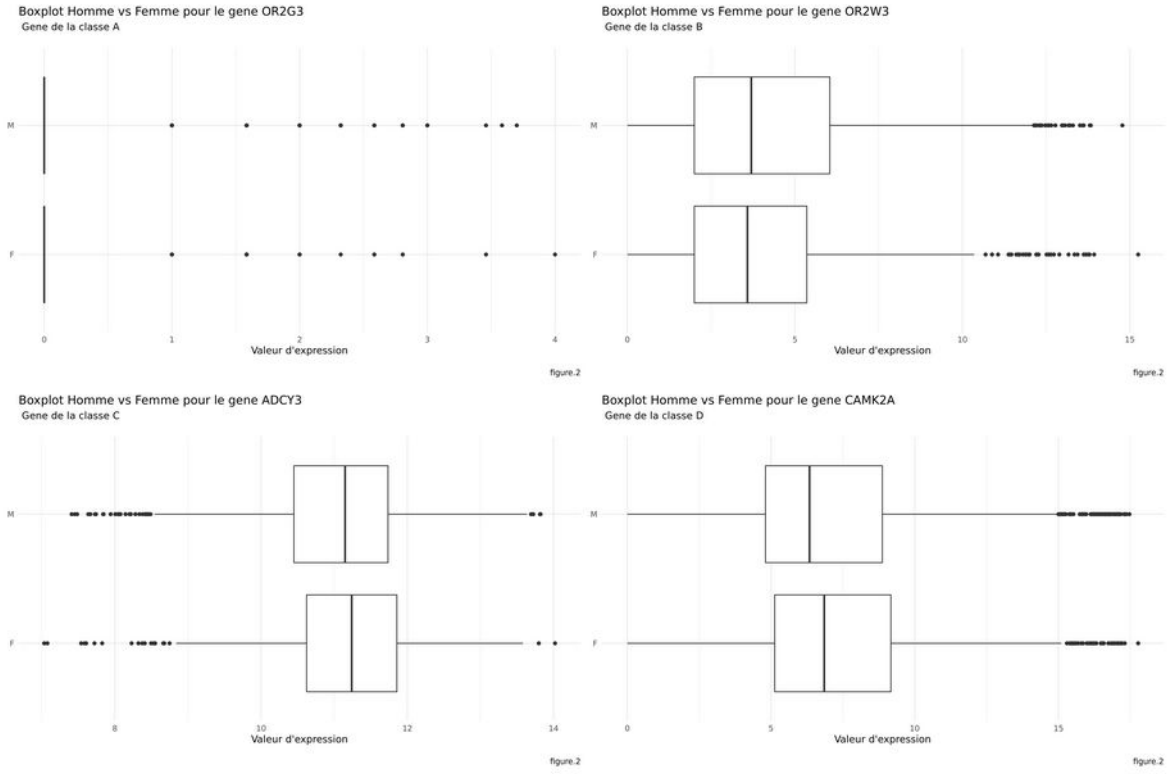
Pour le sexe des individus, on dénombre 1014 Femmes pour 1739 Hommes.

La répartition des échantillons par type de tissu est très inégalitaire également, il y a par exemple 5 échantillons provenant de la bouche, 63 du coeur ou encore 318 du cerveau.

Si l'on croise les deux variables ensemble, la répartition est également non égalitaire. Il y a bien sûr des organes qui sont uniquement féminins. Mais par exemple le cerveau, il y a 202 hommes pour 116 femmes

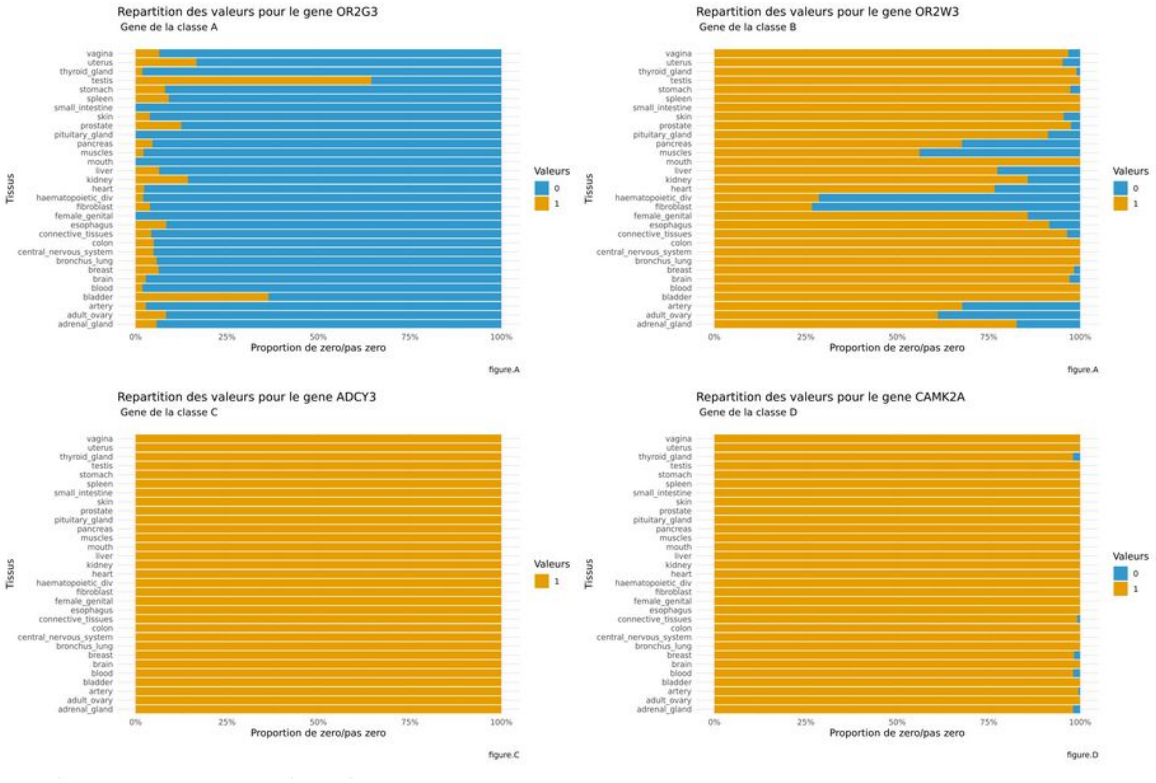
▸Répartition en fonction du sexe

Voici une représentation sous forme de boxplot pour chacun des gènes identifiés précédemment, en fonction du sexe.

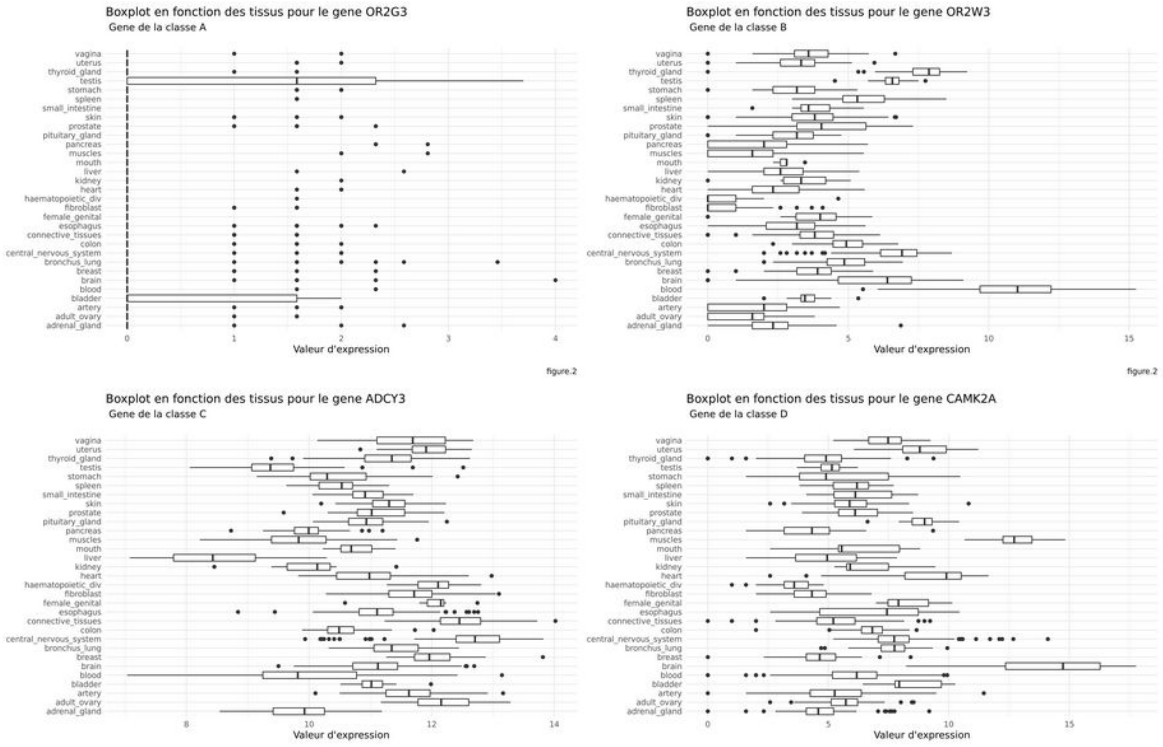


D'après ces graphiques, on peut voir que le sexe ne semble, à priori, pas impacter les valeurs d'expression.

▸Répartition en fonction du tissu



Sur le premier graphique, on voit bien que le gène *OR2G3* est fortement enrichi de zéro dans chacun des tissus sauf pour *testis* et *bladder*. Le gène de la classe B, semble exprimé dans une forte majorité de tissu, sauf pour *haematopoietic_div* et *fibroplast*. Les deux derniers graphiques montrent bien que ces types de gènes sont peu ou pas riche en zéro.



Ces boxplot construit à partir des données normalisées log transformées, montrent qu'en fonction du tissu l'expression du gène varie fortement

D'après cette exploration, nous avons vu que les gènes olfa étaient caractérisés par une forte présence de zéro. Ceci est directement lié à leur caractéristique de gènes tissu-spécifiques on/off. La représentation visuelle des données à l'aide d'une normalisation zéro/pas zéro, nous apporte déjà beaucoup d'information puisqu'elle permet de mettre en avant cette caractéristique de tissu-spécifique.

C'est pourquoi la modélisation par une loi de bernoulli sur le jeu de données normalisées zéro/non zéro semble être adéquat. Bien que le modèle bernoulli, avec les données normalisées zéro ou pas zéro, est un modèle réducteur d'information, il semble néanmoins favorable à ce type de modélisation.

Cependant, l'exploration a aussi montré qu'il existe des gènes totalement dépourvus de zéro c'est à dire qu'ils sont exprimés dans l'ensemble des tissus. Ainsi la modélisation par une loi de bernoulli sur ces derniers, ne peut être une bonne méthode. C'est pourquoi, la recherche d'un autre modèle est envisagé. Un autre facteur important à souligner est la dispersion des données.

Ainsi, il apparaît que plusieurs lois de probabilités sont nécessaires pour la modélisation. Nous allons donc explorer la modélisation de ces lois sur les gènes *olfa* et essayer de voir s'il existe un modèle plus adapté pour une classe de gènes. Il se pose une question importante qui est de savoir comment

évaluer la pertinence d'un modèle ? On peut se demander également, quel modèle peut nous permettre d'expliquer ces co-cofacteurs?

Cette approche est une première tentative pour trouver une modélisation des gènes tissu-spécifiques et ainsi pouvoir déceler dans d'autres dataset des gènes qui auraient ce même profil.

Pour répondre à cette problématique, nous vous présentons les lois qui vont être utilisées pour la modélisations dans la section qui suit.

2. Présentation des modèles

2.1. Loi Normale

Soit une variable réponse $Y = (Y_i, i = 1, ...n)$. Si la composante $Y_i \sim N(\mu_i, \sigma)$ alors sa densité est donnée par

$f(y_i, \mu_i, \sigma)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}}$

L'espérance est alors $E(Y_i) = \mu_i$ et la variance est $Var(Y_i) = \sigma$.

Le log-vraisemblance vaut :

$L = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2$.

Grâce à la relation linéaire qui existe entre la réponse Y_i et les co-variables X_i, le log vraisemblance devient :

$L(\beta) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i\beta)^2$, β désigne le vecteur contenant les coefficients, et x_i le vecteur représentant l'i ème observation des co-variables.

2.2. Loi de Poisson

La loi de Poisson est très privilégiée dans un modèle de comptage [1]. Cette loi est discrète et décrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé. Ainsi on dit que une variable aléatoire Y_i suit une loi de Poisson de paramètre λ_i si la probabilité d'obtenir y_i occurrences est donnée par :

$P(Y_i = y_i/\lambda_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$

La probabilité d'obtenir 0 occurrence est $P(Y_i = 0) = e^{-\lambda}$

L'espérance est alors $E(Y_i) = \lambda_i$ et la variance est $Var(Y_i) = \lambda_i$.

Le log de vraisemblance est donnée par : $L = \sum_{i=1}^{n} [y_i ln(\lambda_i) - \lambda_i - ln(y_i!)]$ ou soit

$L(a) = \sum_{i=1}^{n} [y_i ln(F(x_i a)) - F(x_i a) - ln(y_i!)]$, a désigne le vecteur des coefficients de la régression de Poisson et F est la fonction exponentielle.

2.3. Loi Négative Binomiale

Il existe des populations dispersées dans lesquelles la loi de Poisson change de paramètre de manière aléatoire .On note ainsi une surdispersion c'est-à-dire une variance plus grande que la moyenne (en violation de l'hypothèse de la loi de Poisson). [2]Une distribution de probabilité tenant compte de cette dispersion est la Négative Binomiale (mélange Poisson-Gamma) définie par :

$$P(Y_i = y_i/\lambda_i, \alpha) = \frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})y_i!}(\frac{\alpha\lambda_i}{1+\alpha\lambda_i})^{y_i}(\frac{1}{1+\alpha\lambda_i})^{\frac{1}{\alpha}}$$

Γ désigne la fonction Gamma vérifiant $\Gamma(n+1) = n!\,,\ \forall\ n \in N$

L'espérance est alors $E(Y_i) = \lambda_i$ et la variance est $Var(Y_i) = \lambda_i + \alpha\lambda_i^2$. α est un paramètre auxiliaire mesurant le degré de sur-dispersion . La loi Négative binomiale tend vers la loi de Poisson quand α tend vers zéro.

Le log de vraisemblance est donnée par :

$$L = \sum_{i=1}^{n}\{ln[\Gamma(y_i + \alpha^{-1})] - ln[\Gamma(\alpha^{-1})] - \alpha^{-1}ln(1+\alpha\lambda_i) - y_i ln(1+\alpha\lambda_i) + y_i ln(\alpha) + y_i ln(\lambda_i)\}$$

Si nous supposons que la variable réponse n'est influencée par aucune covariable alors $\lambda_i = \lambda\ \ \forall\ i, i = 1, ..., n$.

L'estimateur par la méthode des moments est donné :

$\hat{\lambda}^{MM} = \overline{X}$ et $\hat{\alpha}^{MM} = \frac{s^2-\overline{X}}{\overline{X}^2}$.

$\overline{X}$ désigne la moyenne empirique et s^2 la variance calculées à partir des observations.

Nous proposons une correction de l'estimation de α comme suit

$$\hat{\alpha}^{MC} = \begin{cases} \hat{\alpha}^{MM} & si \quad s^2 > \bar{X}^2 \\ 0 & sinon \end{cases}$$

2.4. Loi Zero Inflated de Poisson

La distribution Négative Binomiale bien que tenant compte des sur-dispersions, n'arrive pas à capter les sur-représentations des observations nulles.

Une alternative est alors de combiner deux lois de probabilités pour la modélisation. Il existe plusieurs modélisations possibles dites à gonflement zéro dont les plus connues sont zero inflated Poisson (ZIP) et Zero Inflated Negative Binomial(ZINB). Le ZIP est constitué d'une partie binomiale(Bernoulli) et d'une partie Poisson .

Une variable réponse Y_i est modélisée par un ZIP si sa distribution s'exprime comme:

$$P(Y_i = y_i/\pi_i, \lambda_i) = \{\pi_i\ si\ y_i = 0 \quad , \quad (1-\pi_i)\frac{\lambda_i^{y_i}e^{-\lambda_i}}{y_i!}\ si\ y_i \geq 0$$

π_i est la probabilité d'inflation de zéro. En regroupant les observations $y_i = 0$, nous avons finalement

$$P(Y_i = y_i/\pi_i, \lambda_i) = \{\pi_i + (1-\pi_i)e^{-\lambda_i}\ si\ y_i = 0 \quad , (1-\pi_i)\frac{\lambda_i^{y_i}e^{-\lambda_i}}{y_i!}\ si\ y_i > 0$$

Avec $E(Y_i) = (1-\pi_i)\lambda_i$ et $Var(Y_i) = (1-\pi_i)\lambda_i(1+\pi_i\lambda_i)$

Cette modélisation suppose donc deux processus générateurs de zéro.

Le log de vraisemblance est donnée par la formule :

$L = L1 + L2 - L3$ avec

$$L1 = \sum_{\{i,y_i=0\}} ln[\frac{\pi_i}{1-\pi_i} + e^{-\lambda_i}]$$

$$L2 = \sum_{\{i,yi>0\}} [y_i ln(\lambda_i) - \lambda_i - ln(y_i!)$$

$$L3 = \sum_{i=1}^{n} ln(1 + \frac{\pi_i}{1-\pi_i})$$

Pour le modèle sans covariables ,l'estimation par méthode des moments donne :

$\hat{\lambda}^{MM} = \frac{s^2+\overline{X}^2}{\overline{X}} - 1$ et $\hat{\pi}_i^{MM} = \frac{s^2-\overline{X}}{s^2+\overline{X}^2-\overline{X}}$.

Nous corrigeons ces estimateurs comme suit :

$\hat{\lambda}^{MC} = \{\hat{\lambda}^{MM}$ *si* $\overline{X}\neq 0$, 0 *sinon*

$\hat{\pi}_i^{MC} = \{\hat{\pi}^{MM}$ *si* $s^2 > \overline{X}^2$, 0 *sinon*

2.5. Loi Zero Inflated Negative Binomiale

Généralement, une variable réponse Y_i est modélisée par un ZINB s'il y a une surreprésentation de zéro et aussi une dispersion causée par valeurs autres que zéro. Le ZINB est une combinaison d'une loi Binomiale(Bernoulli) et d'une Négative Binomiale. La définition est la suivante:

$$P(Y_i = y_i/\pi_i, \lambda_i, \alpha) = \{\pi_i + (1-\pi_i)(\frac{1}{1+\alpha\lambda_i})^{\frac{1}{\alpha}} si\ y_i = 0,\ (1-\pi_i)\frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})y_i!}(\frac{\alpha\lambda_i}{1+\alpha\lambda_i})^{y_i}(\frac{1}{1+\alpha\lambda_i})^{\frac{1}{\alpha}} si\ y_i > 0$$

Avec $E(Y_i = y_i) = (1-\pi_i)\lambda_i$ et $Var(Y_i = y_i) = (1-\pi_i)\lambda_i(1+(\alpha+\pi_i)\lambda_i)$

Le ZINB tend vers le ZIP quand α tend vers zéro.

Le log de vraisemblance est donnée par la formule :

$L = L1 + L2 + L3 - L4$ avec

$$L1 = \sum_{\{i,y_i=0\}} ln[\frac{\pi_i}{1-\pi_i} + (1+\alpha\lambda_i)^{-\alpha^{-1}}]$$

$$L2 = \sum_{\{i,yi>0\}} \sum_{j=0}^{y_i-1} ln(j + \alpha^{-1})$$

$$L3 = \sum_{\{i,yi>0\}} [-ln(y_i!) - (y_i + \alpha^{-1})ln(1+\alpha\lambda_i) + y_i ln(\alpha) + y_i ln(\lambda_i)]$$

$$L4 = \sum_{i=1}^{n} ln(1 + \frac{\pi_i}{1-\pi_i})$$

Nous avons considéré la relation $ln(\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})}) = \sum_{j=0}^{y_i-1} ln(j + \alpha^{-1})$

2.6. Loi de Bernoulli

Considérons une variable réponse $Y = (Y_i, i = 1, ...n)$. La composante Y_i suit une loi de Bernoulli de paramètre π_i si sa loi de probabilité est donnée par :

$$P(Y_i = y_i/\pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, y_i \in \{0, 1\}$$

L'espérance mathématique est $E(Y_i) = \pi_i$ et la variable $Var(Y_i) = \pi_i(1 - \pi_i)$.

Le log de vraisemblance est donnée par :

$L = \sum_{i=1}^{n}[y_i ln(\pi_i) + (1 - y_i)ln(1 - \pi_i)]$. De plus s'il existe des co-variables alors , le log de vraisemble peut s'écrire

$L(\beta) = \sum_{i=1}^{n}[y_i ln(F(x_i\beta)) + (1 - y_i)ln(1 - F(x_i\beta))]$, F est la bijection réciproque de la fonction logistique définie par : $F(t) = \frac{e^{(t)}}{1+e^{(t)}}$

2.7. Qualité d'ajustement, tests et choix entre différents modèles

Une fois l'estimation des paramètres de chaque loi effectuée, il est fondamental d'étudier la qualité de l'ajustement et de vérifier les hypothèses concernant les coefficients du modèle. Dans le cas où plusieurs modèles concurrents sont maintenant en compétition, des critères de choix du modèle le plus adéquat sont proposés.

2.7.1. Qualité d'ajustement

Les statistiques souvent utilisées pour juger l'adéquation du modèle aux données:

- **la déviance :**

$D = 2(L_{sat} - L)$ où L_{sat} est la log-vraisemblance du modèle saturé, c'est-à-dire le modèle possédant autant de paramètres que de variables ou observations, L est la log-vraisemblance du modèle estimé

- **la statistique du Khi-deux de Pearson**

$$\chi^2 = \sum_{i}^{n}(y_i - \hat{\mu}_i)/var(\hat{\mu}_i)$$

Les deux statistiques suivent approximativement une loi de khi-deux

Lorsque les données sont binaires (loi de Bernoulli) le test d'ajustement de Hosmer et Lemeshow est plus adapté. Il s'agit d'un test d'adéquation de l'ajustement qui utilise la statistique du Khi-deux de Pearson.

2.7.2. Test du rapport de vraisemblance

Le rapport de vraisemblance est le critère habituel utilisé pour tester la significativité des effets du modèle.

C'est un test qui permet de faire des comparaisons entre deux modèles M_1 avec p paramètres et M_2 avec q paramètres emboîtés ($q \leq p$)

L'hypothèse nulle est la nullité des coefficients n'appartenant pas simultanément aux deux modèles. Ce test a pour alternative l'existence d'au moins un coefficient non nul dans l'ensemble supposé nul sous l'hypothèse nulle.

L'hypothèse nulle peut être testée au moyen de la statistique :

$\chi^2_L = -2(L(q) - L(p))$, $L(q)$ désigne le log-vraisemblance du modèle avec q paramètre.

Cette statistique sous l'hypothèse nulle, suit asymptotiquement une loi de $\chi^2(p-q)$ pour les lois (Bernoulli, Poisson) et une loi de Fisher pour les lois à deux paramètres (exemple la loi normale).

Une autre façon de voir la statistique du test est: $\chi^2_L = (D_q - D_p)$, où D_q et D_p les déviances respectives dans M_1 et M_2 .

2.7.3. Autres critères

L'AIC (Akaike Information Criterion) ou le BIC (Bayesian Information Criterion) sont souvent utilisés pour comparer des modèles qui ne sont pas forcément emboîtés.

- $AIC = -2L + 2p$, pour un modèle avec p paramètres.
- $BIC = -2L + p * ln(n)$, pour un modèle avec p paramètres.

3. Modélisation

3.1. “mise en contexte”

L’analyse différentielle de l’expression des gènes repose sur 3 principales étapes : i) la normalisation des données ii) la modélisation de l’expression des gènes iii) le test statistique de l’influence des paramètres biologiques sur l’expression des gènes [3].

La méthode la plus fréquente repose sur la moyenne géométrique du log des données de comptage [7, 8]. Cependant, cette méthode est souvent mal comprise, elle est souvent délaissée pour d’autres méthodes plus simples et les résultats obtenus diffèrent de manière drastique [3,4,5]. Il n’y a donc pas de consensus sur le sujet.

Etant donné la surdispersion des données de comptage dans le contexte de l’expression des gènes, le modèle le plus fréquemment utilisé est un modèle linéaire généralisé fondé sur une loi négative binomiale [7, 8]. Si ce modèle donne de bons résultats sur la plupart des analyses, il se révèle insuffisant quand les données présentent un trop grand nombre de zéro [9,10,11].

Nous proposons dans cette étude d’une part de considérer une nouvelle méthode, “normalisation free” fondé sur la loi de Bernoulli et d’autre part de considérer des versions “zero inflated” des modèles classiquement utilisés.

Nous évaluerons la performance de ces différents modèles sur le jeu de données d’expression de gènes dans des tissus humains sains [12].
Nous nous restreindrons au sous ensemble de gènes particulièrement riche en zéro, les gènes Olfa [13].
Nous évaluerons la capacité de chaque modèle à caractériser l’origine biologique de l'échantillon (organe).
Les résultats obtenus pour les gènes *olfa* seront comparés avec ceux de 384 gènes pris au hasard ainsi que les résultats que l’on obtient pour l’ensemble des gènes.

Comme support visuel de ces résultats nous utiliserons les quatre distributions que nous avons identifiées dans la partie concernant les statistiques descriptives pour lequel nous avions utilisé quatre gènes représentatifs.

Nous allons utiliser ces quatre gènes pour effectuer des comparaisons graphiques de leur distribution avec la loi théorique modélisée. Cette procédure sera effectuée pour chacune des modélisations que nous allons réaliser par la suite.

3.2. Outils utilisés

Nous avons besoins d'accès au luke (cluster des noeuds) puis au noeud du projet Epimed pour récupérer les données et lancer le logiciel R sur ce cluster, les calculs se font ensuite par les machines de calcul de CIMENT.

RStudio est aussi utilisé pour travailler en local, quelques packages importants utilisés sont: pscl, epimedtools, lmtest, MASS, nonnest2.

Google Docs (document partagé) pour travailler en groupe à distance sur le même rapport.

3.3. Modélisation à l'aide d'une loi Normale

Dans un premier temps, les gènes sont modélisés par la loi Normale pour voir si celle-ci peut être adaptée. On utilise le jeu de données normalisées et log transformées (normalisation DESeq2 auxquelles on ajoute 1 et on applique le log).

◆ Procédure utilisée

Pour l'estimation des paramètres du modèle nul, nous utilisons la méthode des moments. On procède de la façon suivante:

- l'estimation de μ est faite par la moyenne de chaque gène
- l'estimation de σ est faite par l'écart-type de chaque gène
- génération des données issues de la loi normale théorique à l'aide de la fonction *dnorm*, avec les paramètres estimés
- calcul de la vraisemblance à l'aide des paramètres estimés et application de la somme des log pour obtenir le log vraisemblance.

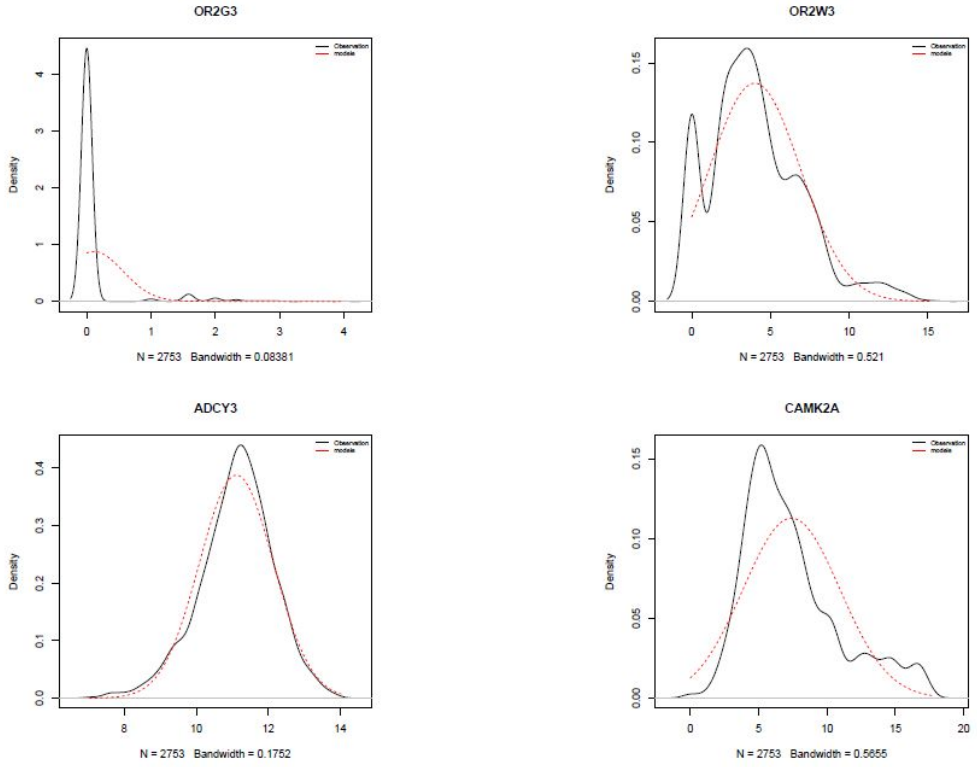
Ci-dessous l'affichage des premières lignes de cette fonction appliquée sur les gènes *OLFA*.

Modelisation de la loi normale sur les genes olfa

	mu	sigma	logL
OR4F5	1.386	1.225	-4465.521
OR4F16	1.035	1.034	-3998.616
CALML6	5.073	1.576	-5158.947
PRKACB	10.924	1.461	-4950.032
CLCA2	4.849	3.435	-7303.163
CLCA1	2.179	2.637	-6575.673

◆ Modélisation de la loi normale sur les quatre classes de gènes

Voici la comparaison graphique de la densité des quatre classe de gènes avec la densité de la loi normale:



Modelisation de la loi Normale

Graphiquement, on peut voir que la loi normale s'adapte mal aux gènes qui sont riches en zéro comme le gène *OR2G3* et *OR2W3*.

Pour le gène *ADCY3*, l'adéquation semble meilleure que pour les autres gènes. Enfin pour le gène *CAMK2A* la loi normale s'adapte mal dû à la forte dispersion des données.

La loi Normale ne semble donc pas être une loi adéquate pour la modélisation de nos gènes.

Le tableau suivant affiche les résultats de la modélisation par la loi Normale sur ces 4 gènes:

	mu	sigma	logL
OR2G3	0.109	0.454	-1731.760
OR2W3	4.004	2.905	-6841.903
ADCY3	11.114	1.028	-3982.460
CAMK2A	7.399	3.537	-7383.905

◆ Modèle avec tissu : anova

Malgré les résultats qui nous semblent peu en adéquation avec une modélisation par la loi Normale, nous avons cherché à connaître l'influence des tissus avec cette modélisation. Ainsi nous avons fait une exploration à l'aide d'une ANOVA.

Voici les premières lignes de résultats sur les gènes *olfa*:

Anova genes olfa

	r2	f	pvaov
OR4F5	0.212	24.413	0
OR4F16	0.185	20.568	0
CALML6	0.539	105.960	0
PRKACB	0.796	353.443	0
CLCA2	0.579	124.608	0
CLCA1	0.674	187.782	0

Le seuil de significativité de test multiples est ajusté par la méthode de Bonferroni. Initialement fixé à 0.05 pour chaque test, ce seuil est ajusté par le nombre de tests effectués. Ce nombre correspond au nombre total de gènes soit 22904.

La valeur de bonferroni est donc de: bonf = 0.05/22904

Lorsque l'on calcule la proportion de test significatif pour les gènes *olfa*, on obtient 86.5%.

On applique l'ANOVA sur les 384 gènes pris au hasard, dont le tableau ci-dessous donne un aperçu des premières lignes.

Anova genes random

	r2	f	pvaov
FCHSD2	0.738	255.463	0
MRPS18B	0.713	225.939	0
UBE3B	0.512	95.026	0
GVQW3	0.552	111.881	0
UPK2	0.487	86.178	0
MIR3176	0.073	7.095	0

La modélisation par ANOVA pour l'ensemble des gènes est significatif à 95% et pour les gènes pris au hasard la significativité est de 97%. Pour les gènes olfa, le pourcentage de tests significatifs est bien inférieur puisque l'on obtient 86%.

Le modèle linéaire classique donne des résultats plutôt convaincants à propos de l'influence du tissu sur les gènes *olfa.* Compte tenu de la nature discrète de nos données on est obligé d'effectuer une log transformation avant l'application de l'anova , et donc la puissance statistique du modèle n'est plus la même. Pour référence on peut suivre cette lecture qui apporte davantage d'information (warton et al)[14].

Ainsi pour la suite, nous allons nous intéresser aux modèles linéaires généralisés.

3.3. Modélisation par la loi de Poisson

Pour la modélisation par la loi de poisson, on utilise le jeu de données normalisées. Deux méthodes d'estimation sont appliquées afin d'en comparer les résultats.

◆ Procédure utilisée

▸La méthode des moments

Sous R nous créons une fonction qui estime le paramètre lambda par la méthode des moments.

La procédure est la suivante:

- estimation de λ par la moyenne de chaque gène
- estimation des probabilités théoriques en appliquant la fonction *dpois* avec pour valeur de lambda l'estimation précédemment faite
- calcul de la vraisemblance par la somme des log de proba

▸La méthode du maximum de vraisemblance

La modélisation de Poisson peut être également réalisée à l'aide de la fonction *glm.* Elle est basée sur une estimation du paramètre lambda par la méthode du maximum de vraisemblance. L'utilisation successive des deux méthodes nous permet de comparer leurs estimations. La fonction *glm* permet également de faire une modélisation en prenant en compte les tissus afin de détecter s'ils ont une influence.

La procédure utilisée est la suivante:

- modélisation par la fonction *glm*, famille de poisson pour un modèle nul
- estimation de lambda

- simulation aléatoire des données selon une loi de poisson à l'aide de la fonction *rpois* avec pour paramètres n = nombre d'observations du jeu de données et lambda = lambda estimé précédemment
- calcul de la proportion de zéro obtenue par cette simulation. On utilise ce résultat pour évaluer la capacité du modèle à estimer la proportion de zéros.
- seconde modélisation par la fonction *glm*, loi de poisson en ajoutant les tissus comme variable explicative
- calcul du rapport de vraisemblance du modèle nul et du modèle avec les tissus

◆ Résultats de la modélisation de poisson

Ci dessous les premiers résultats de la modélisation sur les gènes *olfa* par la méthode des moments:

Modelisation de la loi de poisson sur les genes olfa MM

	lambda_poiss	logL_poiss
CLCA1	911.224	-Inf
CLCA4	3042.898	-Inf
OR10T2	0.105	-1127.225
OR10K2	0.132	-1449.115
OR10K1	0.124	-1288.666

Ci-dessous les premiers résultats de la modélisation par le maximum de vraisemblance (fonction *glm*) pour le modèle nul et avec tissu

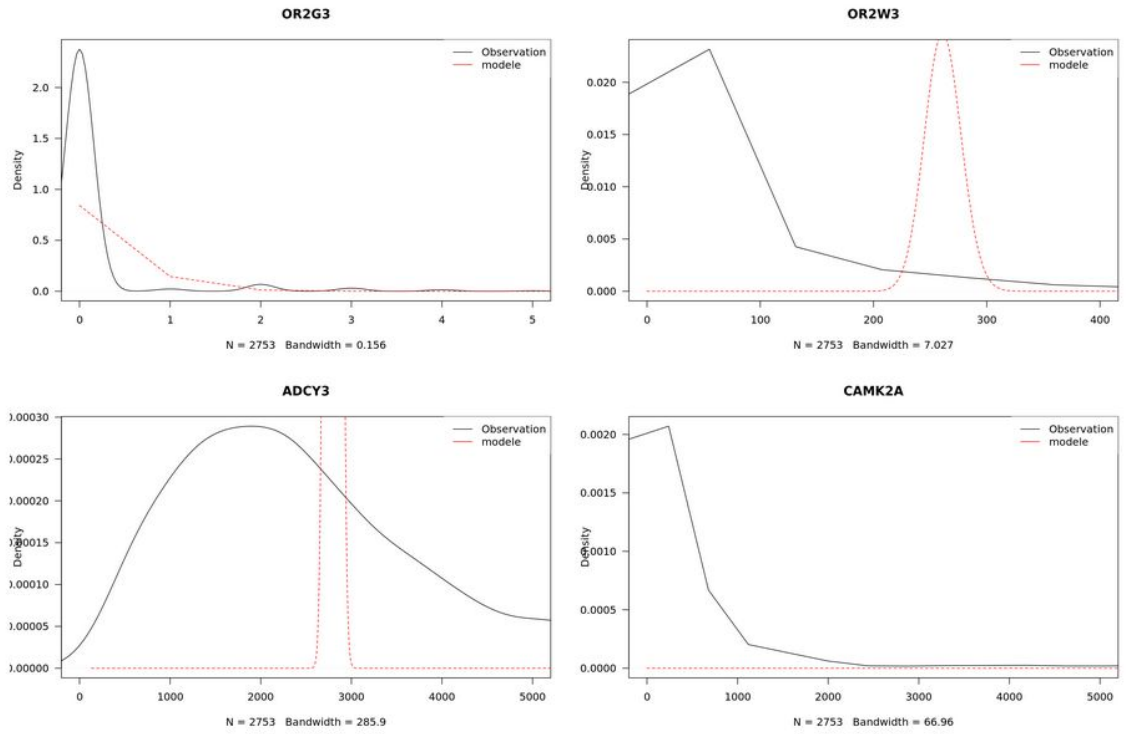
Modelisation de la loi de poisson sur les genes olfa MLE

	Lambda	aic_null	aic_tissue	pseudoR2	pvaleur	prop_zero	prop_predicte
CLCA1	911.224	19093441.183	2549977.754	0.867	0	0.327	0.000
CLCA4	3042.898	46423370.918	14647770.418	0.685	0	0.100	0.000
OR10T2	0.105	2256.450	2089.177	0.117	0	0.958	0.908
OR10K2	0.132	2900.230	2366.969	0.230	0	0.957	0.879
OR10K1	0.124	2579.332	2493.355	0.066	0	0.951	0.888

Le paramètre de poisson estimé par la méthode des moments et du maximum de vraisemblance sont exactement les mêmes.

◆ Modélisation de poisson sur les quatre classes de gènes

Ci-dessous la comparaison graphique de la modélisation de poisson sur les quatre type de gènes:



Graphiquement, on peut voir que la loi de poisson n'est pas adaptée pour ces gènes.

Voici les résultats obtenus pour ces quatre gènes:

Resultats de la loi de poisson sur les quatre genes MM

	lambda_poiss	logL_poiss
OR2G3	0.172	-1651.282
OR2W3	261.452	-Inf
ADCY3	2798.111	-Inf
CAMK2A	5904.487	-Inf

Resultats de la loi de poisson sur les quatre genes MLE

	Lambda	aic_null	aic_tissue	pseudoR2	pvaleur	prop_zero	prop_predicte
OR2G3	0.172	3304.564	2600.482	0.269	0	0.940	0.848
OR2W3	261.452	3845490.798	788919.921	0.797	0	0.152	0.000
ADCY3	2798.111	3386825.142	1271546.754	0.629	0	0.000	0.000
CAMK2A	5904.487	74500164.960	16891827.308	0.773	0	0.003	0.000

On peut voir que la vraisemblance du modèle est infini pour *OR2W3*, *ADCY3* et *CAMK2A*, l'AIC pour ces gènes est aussi très grande. Si on regarde la proportion de zéro estimée par le modèle, on peut voir que cette loi n'est pas très adaptée pour cette estimation. Pour l'ensemble des gènes l'AIC avec tissu est meilleur que l'AIC du modèle nul.

Remarque: dans notre approche, nous nous basons sur des représentations graphiques mais il faudrait faire un test d'ajustement du Khi-deux afin de conclure sur l'adéquation des modèles à nos données.

◆ Influence du tissus

Le modèle de Poisson détecte de façon très significative l'influence du tissu pour 99% les gènes *olfa* à un seuil de Bonferroni 2.10-06. Pour les 384 gènes pris au hasard, ce test est significative pour 97% des gènes.

Malgré que le modèle de Poisson arrive à détecter l'influence du tissue, les vraisemblances ou les AIC obtenus sont très grands voir infini réduisant ainsi la vraisemblance du modèle.

3.4. Modélisation par la loi Négative Binomiale

Une autre loi qui peut être utilisée est la négative binomiale cf de Jong & Heller (2008), sections 6.2 et 6.3, Hilbe (2011) [15] . Nous allons procéder à la modélisation de celle-ci sur nos données normalisées.

◆ Procédure utilisée

Le modèle Négative Binomiale est adapté lorsqu'il y a une surdispersion, c'est à dire quand la variance est plus grande que la moyenne.

Pour les gènes où il n' y a pas de surdispersion, la modélisation par la loi Négative Binomiale n'est pas possible et il sera donc appliqué un modèle de poisson ordinaire.

Dans notre notre approche , nous avons réalisé une première fonction dont les calculs sont basés sur la méthode des moments afin d'estimer les paramètres suivants: lambda la moyenne des observations et alpha le paramètre de dispersion.

Une seconde estimation est réalisée en utilisant la fonction *glm.nb* de R qui se base sur la méthode du maximum de vraisemblance pour estimer les paramètres. Elle utilise l'algorithme Fisher scoring. Une comparaison du modèle nul et du modèle qui prend en compte les tissus est également effectuée à l'aide de cette fonction.

La procédure est la suivante:

- calcul de la proportion de zéro pour chaque gène
- modélisation par une loi négative binomiale si les données sont surdispersées sinon application du modèle de poisson
- estimation de lambda par le modèle nul
- estimation d'alpha pour la négative binomiale, définit à nul pour poisson
- modélisation avec les tissus
- simulation des données selon la loi Négative Binomiale ou Poisson

- estimation de la proportion de zéro obtenue suite à cette simulation
- calcul le rapport de vraisemblance

◆ Comparaison des deux méthodes

▸ Résultats de la méthode des moments

Voici les premiers résultats des estimations:

Modelisation de la loi NB sur les genes olfa MM

	Lambda	Alpha	p_zero	p_predict
OR4F5	3.314	15.231	0.311	0.772
OR4F16	1.765	2.151	0.409	0.448
CALML6	92.072	22.455	0.009	0.701
PRKACB	3675.566	2.377	0.000	0.021
CLCA2	1191.049	11.543	0.057	0.448
CLCA1	911.224	54.423	0.327	0.810

Remarque: pour les gènes olfa, tous ont pu être modélisés par une loi Négative Binomiale, aucun ne présentant un paramètre de surdispersion nul.

▸ Résultats par la méthode du maximum de vraisemblance

Modelisation de la loi NB sur les genes olfa MLE

	Lambda	Alpha	aic_null	aic_tissue	pvaleur	p_zero	p_predict
OR4F5	3.314	1.886	12500.163	11717.17	0	0.311	0.328
OR4F16	1.765	1.437	9902.076	9330.14	0	0.409	0.399
CALML6	92.072	1.723	29791.745	25564.32	0	0.009	0.052
PRKACB	3675.566	1.093	50697.292	45552.79	0	0.000	0.000
CLCA2	1191.049	5.586	33011.549	28320.44	0	0.057	0.209
CLCA1	911.224	11.391	20693.574	16038.65	0	0.327	0.457

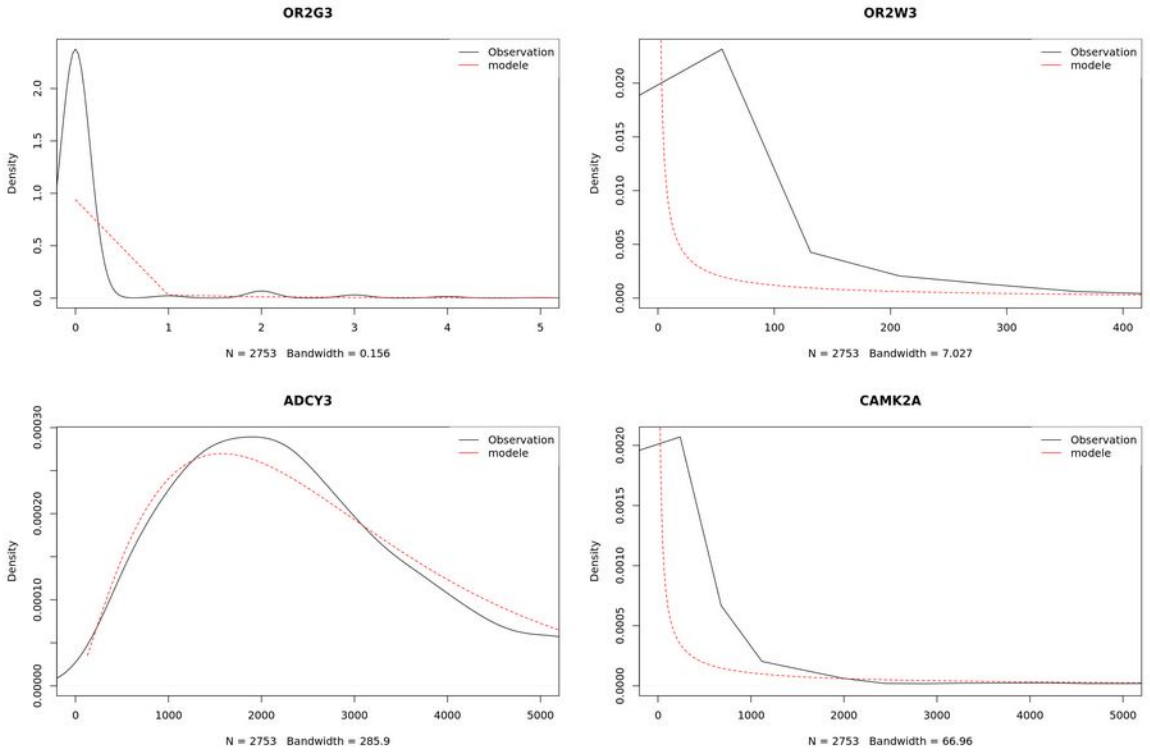
En comparant les deux résultats, nous voyons que l'estimation du paramètre lambda est le même par les deux méthodes et il est égal à celui estimé par la loi de Poisson.

Cependant, le paramètre de dispersion estimé par les deux méthodes n'est pas le même. C'est plutôt l'estimation par la méthode des moments qui est remis en cause. En effet, les estimations faites par les

modèles linéaires généralisés sont asymptotiquement sans biais et consistants (*Ahrmeir and Kaufmann ,1985*) et (*Antoniadis et al. ,1992*) [16] .

Remarque: Pour les gènes olfa, il a été possible de modéliser l'ensemble par la Négative Binomiale. Pour les gènes random, seuls trois d'entre eux ne sont pas modélisés par cette loi.

◆ Modélisation NB sur les quatre classes de gènes



D'après ces graphiques, on peut voir que la loi Négative Binomiale s'ajuste mieux à nos données que la loi de Poisson. Pour le gène *OR2G3*, visuellement on voit que le modèle a du mal à s'ajuster sur le pic en zéro.

Visuellement entre ces quatre gènes, il semble que la modélisation soit la meilleure pour le gène *ADCY3*.Ceci est un point d'attention sur ce que l'on peut émettre comme conclusion à partir de graphiques. En effet, dans le tableau des résultats présenté ci -après on peut voir que l'AIC de ce gène n'est pas le meilleur parmis les 4. De plus, le gène *ADCY3* est de la classe des gènes avec peu de dispersion. Or la négative Binomiale prend en compte le paramètre de dispersion ainsi on peut s'attendre à une meilleure modélisation sur des données présentant davantage de dispersion.

Le tableau ci-après présente les résultats de la modélisation pour ces 4 gènes.

Resultats de la modelisation NB sur les 4 classes genes MLE

	Lambda	Alpha	aic_null	aic_tissue	pvaleur	prop_true	prop_predicte
OR2G3	0.172	28.380	1906.088	1860.815	0	0.940	0.947
OR2W3	261.452	5.056	27992.327	22672.430	0	0.152	0.239
ADCY3	2798.111	0.436	48343.200	45668.798	0	0.000	0.000
CAMK2A	5904.487	4.918	43279.506	36844.523	0	0.003	0.125

Si on s'intéresse aux paramètres estimés par le modèle. On peut voir que cette loi a une bonne capacité de modélisation pour la proportion de zéro. Pour le gène *OR2G3*, qui est très riche en zéro, le modèle estime 94,7% de zéros pour une proportion réelle de 94%.

On peut voir que l'AIC est assez faible pour le gène OR2G3 comparé à l'AIC des trois autres gènes.

◆ Influence du tissu

La modélisation par une loi Binomiale Négative détecte l'effet de tissu pour 74% des gènes avec le même seuil que celui fixé pour le modèle poisson.

Pour les 384 gènes pris au hasard, seulement 97 % d'entre eux sont significatifs pour le modèle Négative Binomiale.

La particularité des gènes *olfa* est qu'ils s'expriment dans peu de tissus. Ainsi, pour mettre en évidence cette expression différentielle, nous considérons des modèles dits à gonflement de zéro proposés par (Mullahy 1986) [17] .

3.5.Modélisation par une loi Zero Inflated de Poisson

Intéressons nous maintenant à une modélisation par un modèle à inflation de zéro qui semble pouvoir être adéquat pour les gènes qui présentent beaucoup de zéro. Dans un premier temps, nous allons utiliser la modélisation Zero Inflated de Poisson (ZIP) avant de tester la Zero Inflated Négative Binomiale (ZINB).

◆ Procédure utilisée

Comme pour la Négative Binomiale, nous faisons les estimations des paramètres de cette loi par la méthode des moments en créant notre propre fonction pour un modèle nul. Puis nous ferons une seconde estimation grâce à la fonction *zeroinfl* du package pscl enrichit du modèle avec les tissus.

▸ Estimation par la méthode des moments

On procède de la manière suivante:

- estimation de pi et lambda par la méthode des moments
- calcul de proportion de zéro de chaque gène
- simulation de données aléatoires grâce aux paramètres estimés pour chaque gène
- calcul de la proportion de zéro obtenue par cette simulation

▸ Estimation par la méthode du maximum de vraisemblance

On procède de la manière suivante:

- calcul de la proportion de zéro de chaque gène
- estimation des paramètres pi et lambda pour le modèle nul
- nouvelle modélisation avec prise en compte du tissu comme variable explicative, (pour la partie Poisson, expliquant les count)
- calcul de l'AIC pour les deux modèles
- test du rapport de vraisemblance pour l'influence du tissu sur la partie Poisson(sur les count)
- simulation aléatoire des données avec les paramètres estimés pour le modèle nul
- estimation de la proportion de zéro grâce aux données simulées
- remarque: si le gène ne peut être modélisé par un ZIP (absence de zéro) tous les paramètres sont fixés à zéro

Il est à noter que pour la modélisation ZIP avec prise en compte du tissu, il faut choisir si le tissu est une variable explicative pour la partie count ou pour la partie d'inflation en zéro. Il a été choisi que celui-ci soit explicatif pour la partie des count en se basant sur l'AIC et aussi dans l'intention de faire une comparaison avec le modèle de Poisson par la suite.

Remarque: À noter que les types de tissus peuvent aussi être explicatifs pour la partie d'inflation en zéro, mais dans notre étude ce modèle ne permet pas de faire une comparaison pertinente avec les autres modèles, notamment avec Poisson.

◆ Résultats

Estimation par la méthode des moments:

Modelisation de la loi zip sur les genes olfa MM

	pi	lambda	p_zero	p_pred
OR4F5	0.938	53.794	0.311	0.941
OR4F16	0.683	5.561	0.409	0.677
CALML6	0.957	2159.545	0.009	0.961
PRKACB	0.704	12412.615	0.000	0.696
CLCA2	0.920	14939.087	0.057	0.926
CLCA1	0.982	50502.506	0.327	0.983

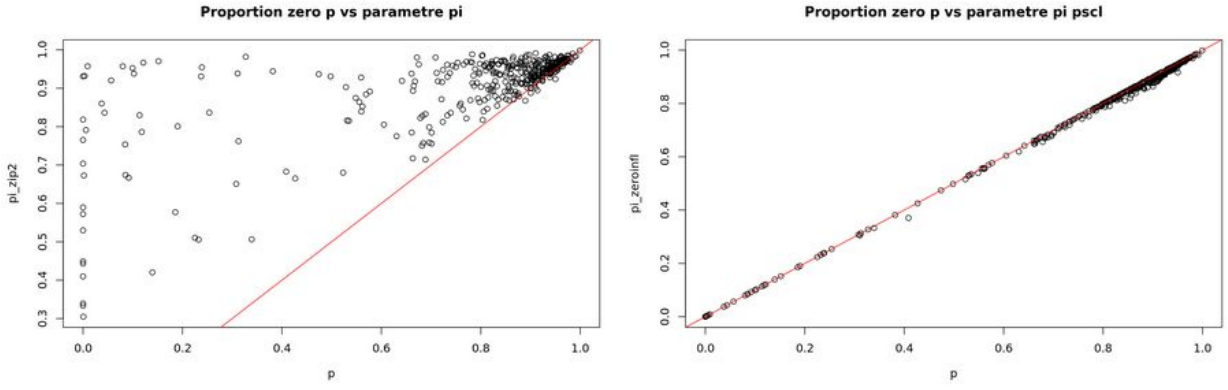
Estimation par la méthode du maximum de vraisemblance (package *pscl*):

Modelisation de la loi zip sur les genes olfa MLE

	pi	Lambda	AIC	AIC_tissue	pvaleur	p_zero	p_pred
OR4F5	0.305	4.769	24291.84	20742.08	0	0.311	0.317
OR4F16	0.370	2.804	11437.19	10100.91	0	0.409	0.407
CALML6	0.009	92.916	848883.28	283518.02	0	0.009	0.009
PRKACB	0.000	0.000	0.00	0.00	0	0.000	0.000
CLCA2	0.057	1262.595	14451528.05	3890088.29	0	0.057	0.055
CLCA1	0.327	1354.535	17107981.10	2491391.94	0	0.327	0.333

Pour le modèle nul (sans variable explicative) nous constatons que les estimations faites avec la méthode des moments sont significativement très différentes de celles données par la fonction *zeroinf* du package *pscl*.

Ci-après la représentation graphique de la proportion de zéro réel du jeu de données en fonction du paramètre pi estimé par le modèle. Le premier graphique est celui de l'estimation de pi donné par la méthode des moments, le second ceux donnés par la fonction *zeroinfl* du package *pscl*.

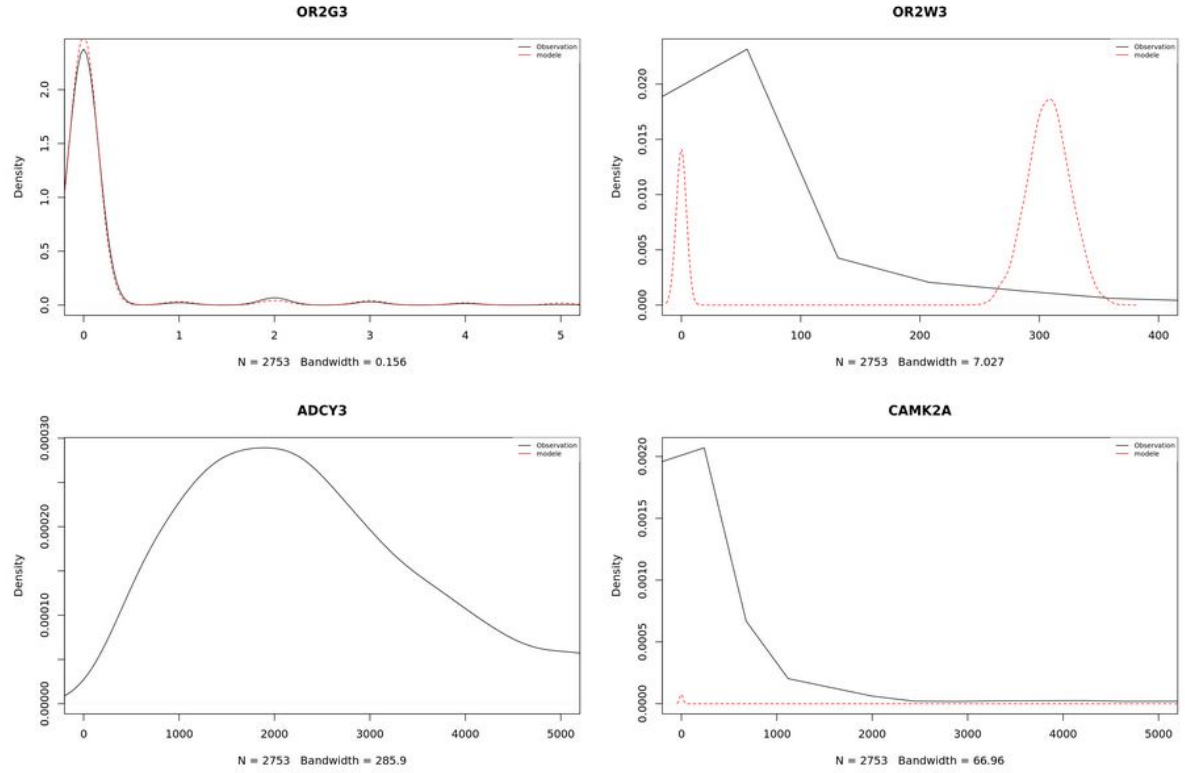


Sur le premier graphique, on observe que l'ensemble des points est très éloigné de la première bissectrice. Le modèle surestime le nombre de zéro par rapport au nombre réel.

Sur le second graphique, on peut voir que les points sont sur la première bissectrice ainsi le modèle fournit un bon ajustement du nombre de zéro.

Comme on l'a vu précédemment les estimations faites pour les modèles linéaires généralisés sont asymptotiquement normales, sans biais et consistants. C'est pourquoi nous allons exploiter les résultats donnés par la fonction *zeroinfl* pour la suite.

◆ Modélisation ZIP sur les quatre classes de gènes



D'après les graphiques ci dessus, on peut voir que le modèle ZIP est bien adapté pour le gène *OR2G3* qui est riche en zéro. Comme on peut le voir, la courbe théorique vient s'ajuster très proche sur le pic en zéro ainsi que sur les petites variations qui suivent.

Par contre, ce modèle ne semble pas adapté pour les trois autres gènes. Pour le gène *OR2W3*, qui présente 15% de zéro et une forte dispersion ensuite, le modèle ZIP semble mal s'adapter. Cela est surement dû à la limite du modèle lorsqu'il y a une forte dispersion.

Remarque: Dans la modélisation, il y a certains gènes sur lesquels on ne peut pas appliquer cette modélisation car ils sont dépourvus de zéros. Ce sont les 12 gènes identifiés lors de l'analyse descriptive (cf. 2.2). C'est le cas pour le gène *ADCY3* dépourvus de zéro.

Enfin pour le gène *CAMK2A* qui a une très faible proportion de zéro (moins de 1%) et une forte dispersion des données, on peut voir que le modèle n'est pas adapté.

Resultats de la modelisation zip sur les 4 genes

	pi	Lambda	AIC	AIC_tissue	pvaleur	p_zero	p_pred
OR2G3	0.936	2.668	1871.388	1847.657	0	0.940	0.945
OR2W3	0.152	308.256	3610772.556	784703.554	0	0.152	0.146
ADCY3	0.000	0.000	0.000	0.000	0	0.000	0.000
CAMK2A	0.003	5923.853	74393833.076	16890524.280	0	0.003	0.003

En regardant les résultats chiffrés, on peut voir que l'estimation de la proportion de zéro est plutôt bonne pour les quatre gènes. L'AIC est très élevée pour les gènes *OR2W3* et *CAMK2A* .

En comparant ces résultats avec ceux que l'on obtient pour la modélisation avec la Négative Binomiale, on peut noter que les AIC avec tissu obtenus pour la Négative Binomiale ont des valeurs moins dispersées que pour ceux du ZIP. Pour le gène *OR2G3,* les deux modèles donnent un AIC de l'ordre de 2.10^3 alors que pour le gène *OR2W3* le ZIP donne 4.10^5 pour 2.10^4 pour la Négative Binomiale, et pour le gène *CAMK2A* le ZIP donne 2.10^7 pour 10^4 pour NB. Ces résultats mènent à penser que la modélisation Négative Binomiale est plus adaptée pour la partie des count que la modélisation de Poisson.

◆ Influence du tissu

Pour les gènes *olfa*, lorsque l'on ajoute le tissu comme variable explicative sur la partie des count, on observe que 85% des tests sont significatifs (avec la correction de Bonferroni).

Pour la même modélisation sur le groupe des 384 gènes que l'on a pris au hasard, le modèle ZIP est applicable pour 212 de ces gènes. Le test de l'influence des tissus est significatif dans 95% des cas au même seuil.

Il serait intéressant de réaliser la modélisation pour l'ensemble des gènes afin de comparer la significativité de l'ensemble mais cela demande trop de ressources pour le calcul.

3.6. Modélisation par une loi Zero Inflated Négative Binomiale

Nous venons de voir que modèle de zip est bien adapté pour les gènes ayant une forte proportion de zéro, mais pas vraiment lorsqu'il y a une forte dispersion.

Ainsi nous considérons le modèle Zero Inflated Negative Binomial (ZINB) qui ajoute un paramètre de dispersion au modèle zip pour contrôler la surdispersion.

◆Procédure utilisée

Pour ce modèle, nous avons utilisé une fonction programmée par le package *pscl* et nous procédons comme suit :

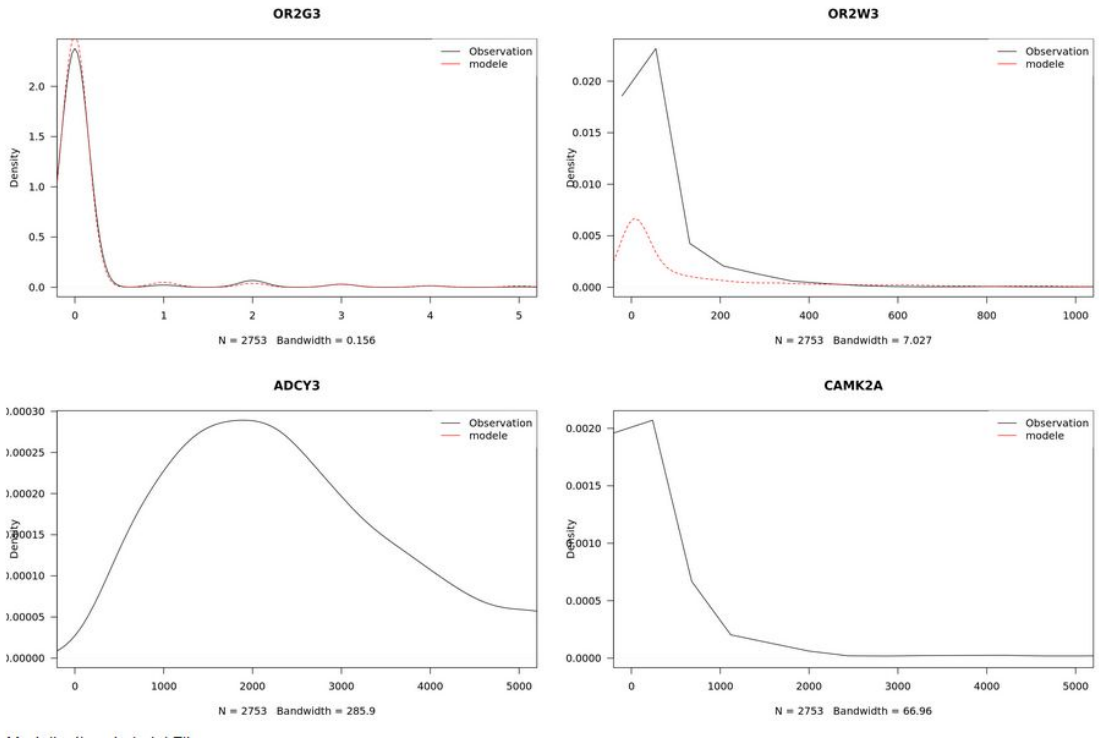
- calcul de la proportion de zéro pour chaque gène
- estimation des paramètres pour le modèle nul de ZINB
- estimation des paramètres pour le modèle avec tissu comme variable explicative (pour la partie des count)
- calcul de l'AIC des deux modèles (nul et avec tissu)
- test du rapport rapport de vraisemblance pour estimer l'effet du tissu
- simulation aléatoire de données selon les paramètres estimés pour le modèle nul
- estimation de la proportion de zéro dans les données simulées
- Remarque: les paramètres et estimations sont fixés à zéro si le gène ne peut pas être modélisé par cette loi (s'il ne possède pas de zéro)

◆ Résultats

Modelisation de la loi zibn sur les genes olfa

	pi	Lambda	Alpha	AIC	AIC_tissue	pvaleur	p_zero	p_predict
OR4F5	0	3.314	1.886	12502.165	11719.168	0	0.311	0.337
OR4F16	0	1.765	1.437	9904.077	9316.741	0	0.409	0.408
CALML6	0	92.073	1.723	29793.745	25534.374	0	0.009	0.052
PRKACB	0	0.000	0.000	0.000	0.000	0	0.000	0.000
CLCA2	0	1191.091	5.586	33013.561	28321.626	0	0.057	0.201
CLCA1	0	911.159	11.391	20695.575	16018.958	0	0.327	0.447

◆ Modélisation ZINB sur les quatre classes de gènes



D'après ces graphiques, on peut voir que le modèle ZINB est bien adapté pour le gène *OR2G3* qui est riche en zéro.

Contrairement au ZIP qui ne s'adapte pas très bien au gène *OR2W3,* ce modèle semble avoir une meilleure adéquation aux données.

Remarque: ce modèle, comme pour le ZIP ne peut pas être utilisé si les données ne comportent pas de zéros. Aussi, la modélisation ne s'est pas faite sur les 12 gènes dépourvus de zéro.

Resultats de la modelisation zinb sur les 4 genes

	pi	Lambda	Alpha	AIC	AIC_tissue	pvaleur	p_zero	p_predict
OR2G3	0.929	2.437	0.253	1851.839	1841.043	0	0.940	0.947
OR2W3	0.000	261.452	5.056	27994.345	22656.357	0	0.152	0.237
ADCY3	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000
CAMK2A	0.000	5904.442	4.919	43281.510	36846.523	0	0.003	0.119

On remarque que pour le gène *OR2G3* la proportion de zéro sur les données simulées est approximativement équivalente à celle des données réelles. Pour le gène *OR2W3,* l'estimation de pi est de zéro et la proportion de zéro des données simulées est plus élevée que la proportion de zéro

réelle. Pour le gène *CAMK2A*, la proportion de zéro des données simulée est plus élevée que les données réelles. Concernant les AIC de ces modèles, on remarque qu'ils sont plus faibles que ceux que l'on obtient avec le modèle ZIP. La comparaison pour ces gènes semblent montrer que la modélisation ZIBN est plus adaptée que ZIP quand il y a beaucoup de zéro.

◆ Influence du tissu

Pour les gènes *olfa*, les 12 gènes qui n'ont pas de zéro n'ont pas été modélisés par cette loi alors que pour les gènes random, il y a 172 gènes dans ce cas.

A l'aide de cette modélisation, lorsque l'on ajoute le tissu comme variable explicative on obtient pour les gènes *olfa* 78% de significativité. Alors que pour les gènes random l'influence du tissu est significatif pour 95% .

3.7 Modélisation par une loi de Bernoulli

L'objectif de notre projet est de mettre en évidence la particularité des gènes *olfa* qui sont tissu-spécifiques. Comme on l'a vu cela s'exprime par une forte proportion de zéro.

Ainsi après avoir utilisé des modèles classiques, nous utilisons dans cette section une modélisation par une loi de Bernoulli. Nous supposons pour ce modèle, que nous avons un succès si le comptage observé est nul et un échec sinon. L'intérêt de cette modélisation est qu'elle se passe de toute normalisation.

On travaille sur les données brutes que l'on modifie de la façon suivante: un si c'est un un zéro et zéro si c est une autre valeur.

◆Procédure utilisée

▸Méthode des moments

Nous créons une fonction qui nous permet d'estimer le paramètre p (la proportion de zéro) par la méthode des moments.

La démarche est la suivante:

- estimation du paramètre p par la fréquence de zéro pour chaque gène, modèle nul
- calcul du log-vraisemblance du modèle nul
- estimation du paramètre p pour chaque gène et chaque tissu
- calcul du log-vraisemblance pour chaque gène et chaque tissu
- pour un gène donné, calcul de la somme des log-vraisemblance obtenue pour chaque tissu

- réalisation du test du rapport de vraisemblance
- calcul du ratio du log vraisemblance du modèle nul sur le modèle avec tissu

▸Méthode du maximum de vraisemblance

La modélisation de Bernoulli peut être également réalisée à l'aide de la fonction *glm* qui comme on l'a vu se base sur une estimation par le maximum de vraisemblance. Nous avons voulu faire la comparaison des deux méthodes.

La procédure est la suivante:

- modélisation par la fonction *glm* de famille binomial("logit") pour un modèle nul
- estimation de pi la proportion de zéros à l'aide des coefficients
- simulation aléatoire des données selon une loi de bernoulli avec les paramètres que l'on a estimé (*rbinom*: n = nombre d'observations de notre jeu de données, size = 1, et prob = pi,le nombre de zéro estimé par le modèle)
- calcul de la proportion de zéro obtenue par cette simulation. Cela va nous aider à voir la capacité du modèle à estimer la proportion de zéros
- modélisation par la fonction *glm* de la famille binomial avec les tissus
- calcul du rapport de vraisemblance du modèle nul et du modèle avec les tissus
- calcul du pseudo R2

◆ Résultats

Voici les observations des six premières lignes de résultats pour les deux méthodes de bernoulli:

Modelisation de la loi bernouilli sur les genes olfa MM

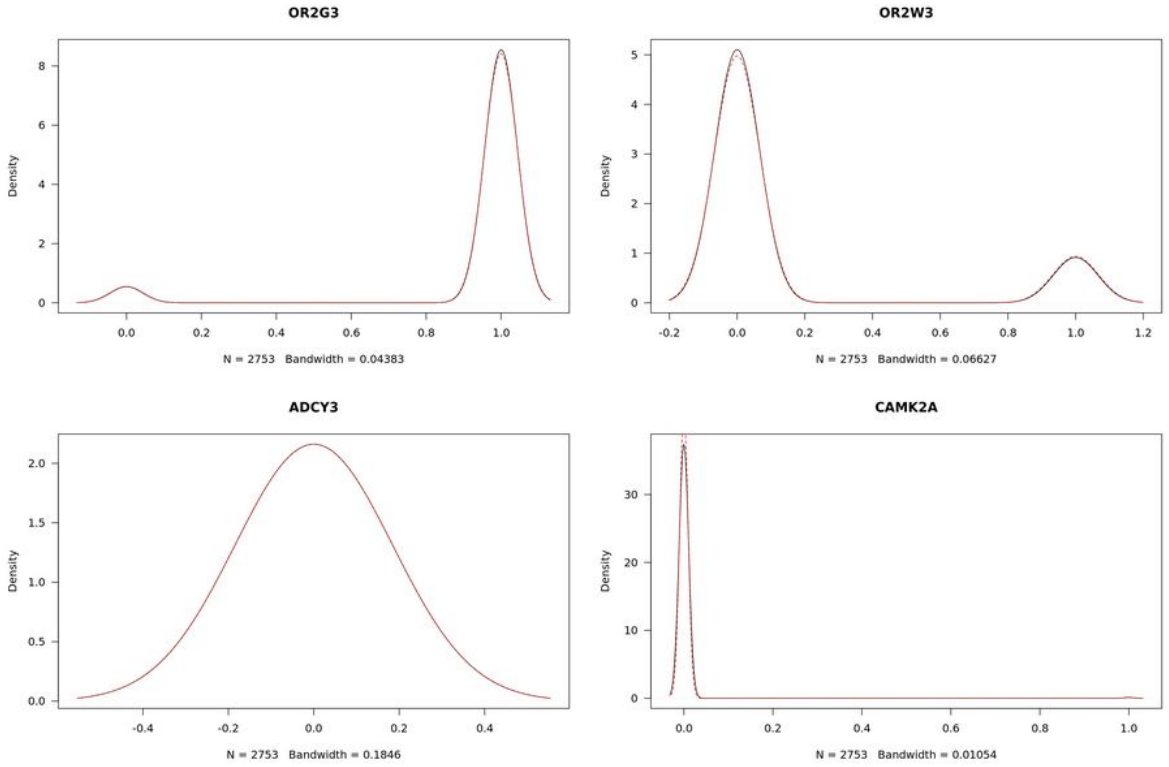
	p	logL	logL_tissue	lr	pval
OR4F5	0.311	-1706.435	-1546.956	318.958	0
OR4F16	0.409	-1862.024	-1731.557	260.934	0
CALML6	0.009	-142.425	-93.873	97.104	0
PRKACB	0.000	0.000	0.000	0.000	1
CLCA2	0.057	-599.306	-502.597	193.419	0
CLCA1	0.327	-1740.543	-1508.571	463.945	0

Modelisation de la loi bernouilli sur les genes olfa MLE

	p	aic_nul	aic_tissue	R2	pvaleur	p_zero	p_predict
OR4F5	0.311	3414.870	3155.911	0.093	0	0.311	0.315
OR4F16	0.409	3726.047	3525.113	0.070	0	0.409	0.406
CALML6	0.009	286.851	249.746	0.341	0	0.009	0.009
PRKACB	0.000	2.000	62.000	0.000	1	0.000	0.000
CLCA2	0.057	1200.613	1067.194	0.161	0	0.057	0.055
CLCA1	0.327	3483.087	3079.142	0.133	0	0.327	0.333

Les résultats obtenus par la méthode des moments et par la méthode du maximum de vraisemblance sont les mêmes.

◆ Modélisation de Bernoulli sur les quatre classes de gènes



Sur ces graphiques, nous avons représenté les données brutes transformées (zéro ou un) ainsi que la courbe théorique de loi de bernoulli avec les paramètres estimés de chaque gène.

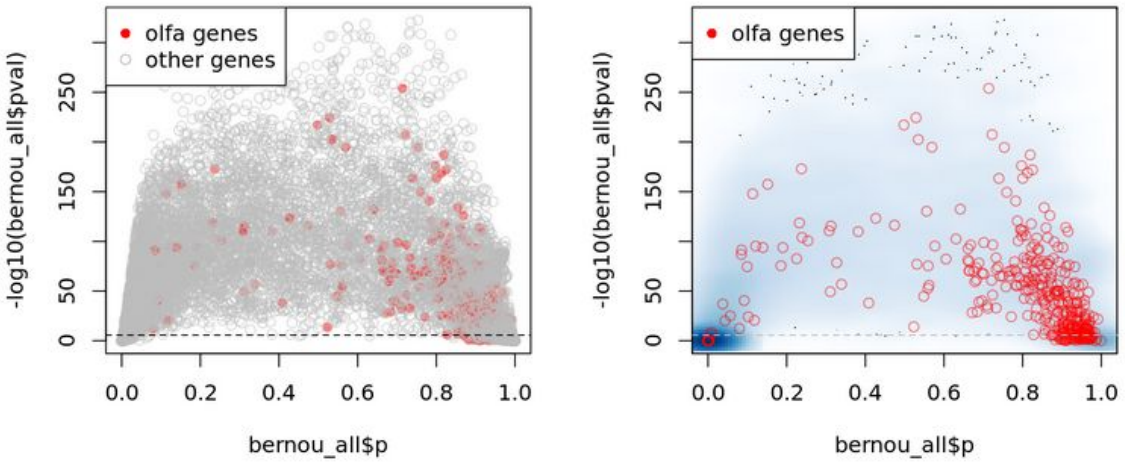
On peut voir qu'avec cette transformation des données, la loi de Bernoulli est bien adaptée pour tous les gènes.

Rsultats de la modelisation bernouilli sur les 4 genes

	p	aic_nul	aic_tissue	R2	pvaleur	p_zero	p_predict
OR2G3	0.940	1250.692	1097.465	0.171	0.000	0.940	0.942
OR2W3	0.152	2346.883	1563.187	0.360	0.000	0.152	0.146
ADCY3	0.000	2.000	62.000	0.000	1.000	0.000	0.000
CAMK2A	0.003	122.989	156.478	0.219	0.649	0.003	0.003

◆ Caractéristiques des gènes OLFA

Afin de savoir si les gènes olfa dégagent une particularité, on effectue la comparaison avec les gènes pris au hasard.



On peut voir que la famille des gènes *olfa* est composée majoritairement par des gènes avec une forte proportion de zéro.

Ces deux graphiques sont particulièrement intéressants. Ils montrent d'une part que les gènes *olfa* ont une probabilité élevée et qu'ils sont en majorité significatif pour le test (ceux qui sont au dessus du trait en pointillé). Ainsi le modèle avec le tissu est meilleur que celui sans. La plupart des gènes est certe significatif mais les valeurs de p sont beaucoup plus dispersées.

Les gènes *olfa* sont significatifs à 80 % avec le modèle de bernoulli. Alors que pour les 384 gènes pris au hasard, le taux de significativité est d'environ 42%.

4. Comparaison des modélisations

Dans la partie précédente, nous avons utilisé plusieurs lois pour modéliser les gènes. Il convient maintenant d'effectuer une comparaison de ces modélisations.

Selon (*MacDonald and Lattimore* ,2010) [18], il n'existe pas de meilleur modèle à proprement parlé, tout dépend du type de données utilisées ainsi que de leur capacité prédictive.

De plus, il n'y a pas critère clair concernant la proportion de zéros acceptable ou non pour justifier l'utilisation de la loi binomiale négative ou les modèles modifiés en zéro. L'article de Yau, Wang, & Lee (2003) propose alors de comparer l'indice d'ajustement, le BIC.

D'après nos lectures, nous avons décidé de juger la qualité de nos modèles sous trois angles.

Dans un premier temps, nous allons nous intéresser à leur capacité à modéliser la proportion de zéro.

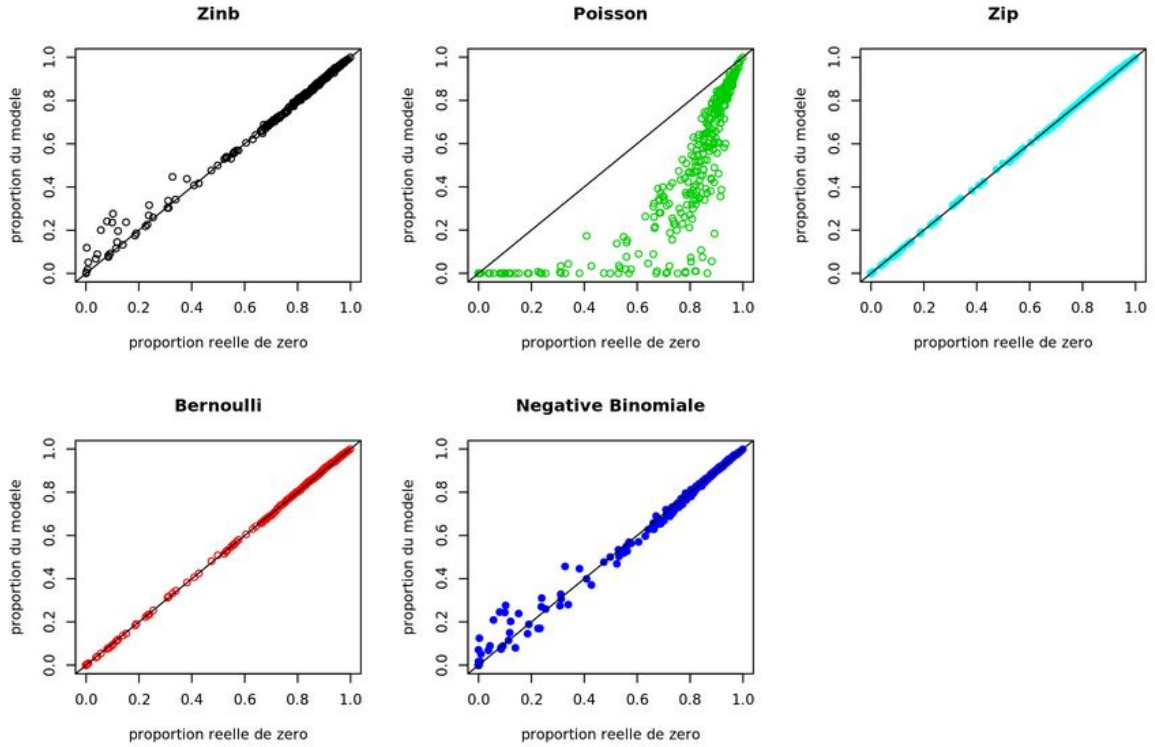
Dans un second temps , nous allons regarder les critères d'ajustement: le Critère d'information d'Akaike (AIC) .Ce critère, plus il est faible meilleur est le modèle.

Malheureusement, il n'est pas aisé de comparer différentes distributions entre elles. Selon (Atkins & Gallop, 2010)[19], le modèle de Poisson est niché dans le modèle Négative Binomiale alors que le Zero Inflated de Poisson est niché dans le Zero Inflated Negative Binomial. Il est donc possible de comparer le modèle de Poisson avec le modèle négative binomiale et le Zero Inflated de Poisson avec le Zero Inflated Negative Binomial grâce au test du rapport de vraisemblance. C'est ce que nous effectuons dans un premier temps.

Puis nous utiliserons le test de Vuong pour comparer les modèles non imbriqués, c'est à dire le modèle de Poisson avec ZIP et Négative Binomiale avec ZINB.(Vuong 1989) [20].

4.1. Evaluation de la modélisation de la proportion de zéro

Nous allons analyser graphiquement cette modélisation pour chacun des modèles en traçant le graphique de la proportion de zéro réelle du jeu de données en fonction de la proportion donnée par le modèle. Cette évaluation ne concerne que les gènes *OLFA*.



Comme on peut le voir, le modèle de poisson n'apporte pas une bonne modélisation des zéros. Le modèle de Bernoulli, le ZIP , la Négative Binomiale, et le ZINB parviennent à approcher une estimation de zéro très proche des observations réelles. C'est notamment les modèles de Bernoulli et ZIP qui semblent être les meilleurs pour estimer les zéros.

Les tableaux ci-après affichent l'erreur quadratique moyenne des proportions de zéro estimées par rapport à la réalité, pour les gènes olfa puis pour les 384 gènes pris au hasard.

EQM Nombre de zero, genes olfa

Modele	EQM
Poisson	9.23e-02
NB	5.85e-04
ZIP	2.21e-05
ZINB	4.34e-04
Bernoulli	1.83e-05

EQM Nombre de zero, genes random

Modele	EQM_rnd
Poisson	5.82e-02
NB	3.69e-03
ZIP	1.18e-05
ZINB	2.31e-03
Bernoulli	6.16e-06

On peut voir que Zip et Bernoulli ont une erreur très faible de l'ordre de 10^{-5} .

4.2. Évaluation à l'aide du rapport de vraisemblance

Pour chacun des gènes, on réalise le test du rapport de vraisemblance entre deux modélisations. Les modèles utilisés sont ceux pour lesquelles le tissu est considéré comme variable explicative.
On effectue ce test à l'aide de la fonction *lrtest*. C'est un test unilatéral qui a pour hypothèses :

- $H_0 : \alpha = 0$
- $H_1 : \alpha > 0$

Sous l'hypothèse nulle, la statistique du test suit une loi de khi-deux.
Ainsi pour chaque gène, la fonction retourne la valeur de la statistique du test et la p-valeur.
Ce test est possible pour comparer la Négative Binomiale avec la loi de Poisson et la loi ZIP avec la loi ZINB. En effet, la loi de poisson est un cas particulier de la négative binomiale.

◆ Négative Binomiale vs Poisson

La réalisation de ce test sur les gènes olfa est significatif pour 383 gènes sur 384 avec la correction de bonferroni. La loi Négative Binomiale modélise mieux les gènes *olfa* que la loi de Poisson.

Le test du rapport de vraisemblance est également utilisé sur les modélisations faites pour les gènes pris au hasard. Le résultat est identique 383 gènes sur 384 sont significatifs au seuil corrigé de bonferroni.

On peut donc penser que la loi Négative Binomiale est mieux adaptée dans le cadre de la modélisation de gènes qu'une loi de poisson.

◆ ZIP vs ZINB

Le test du rapport de vraisemblance entre le modèle ZIP et le modèle ZINB est significatif pour 292 gènes olfa contre 369 gènes random.

Le modèle ZINB est davantage adéquat pour les gènes random que pour les gènes olfa.

4.3. Évaluation à l'aide du test de Vuong

Le test de Vuong permet de vérifier si les modèles zero inflated sont mieux que les modèles de Poisson ou Négative Binomiale.

Sous l'hypothèse nulle, le rapport de vraisemblance des deux modèles suit une loi normale de moyenne nulle puisque les vraisemblances des deux modèles sont généralement proches.

A l'issue de ce test , l'hypothèse alternative permet de définir si un modèle a significativement plus de vraisemblance sur les données que l'autre.

La réalisation de ce test est faite pour comparer le modèle ZIP avec Poisson et le modèle ZINB avec Négative Binomiale. Pour chacun de ces tests, les modèles sont construit avec le tissu comme variable explicative.

◆ ZIP vs Poisson

Les résultats du test de Vuong pour comparer le modèle ZIP avec le modèle de Poisson sont résumés dans les tableaux suivants:

Vuong: ZIP vs Poisson, genes olfa

	ZIP	Poisson	NoTest	NoInflation
Total	276	0	96	12
Significatif	247	0	96	12

Vuong: ZIP vs Poisson, genes random

	ZIP	Poisson	NoTest	NoInflation
Total	196	1	15	172
Significatif	93	0	15	172

On peut voir que le modèle ZIP est mieux adapté que le modèle de Poisson pour 247 des 384 gènes *olfa,* soit 89%. Pour les gènes pris au hasard, seuls 196 présentent des zéros et peuvent donc être modélisés par un ZIP. Pour ces gènes, seuls 93 tests (soit 47%) permettent de dire que la modélisation ZIP est meilleure que Poisson.

Il semble donc que le modèle ZIP soit plus adapté pour les gènes *olfa* que pour les gènes pris au hasard.

◆ ZINB vs NB

Les résultats du test de Vuong sont résumés dans les tableaux suivants:

Vuong: ZINB vs NB, genes olfa

	ZINB	NB	NoTest	NoInflation
Total	227	22	122	13
Significatif	33	0	122	13

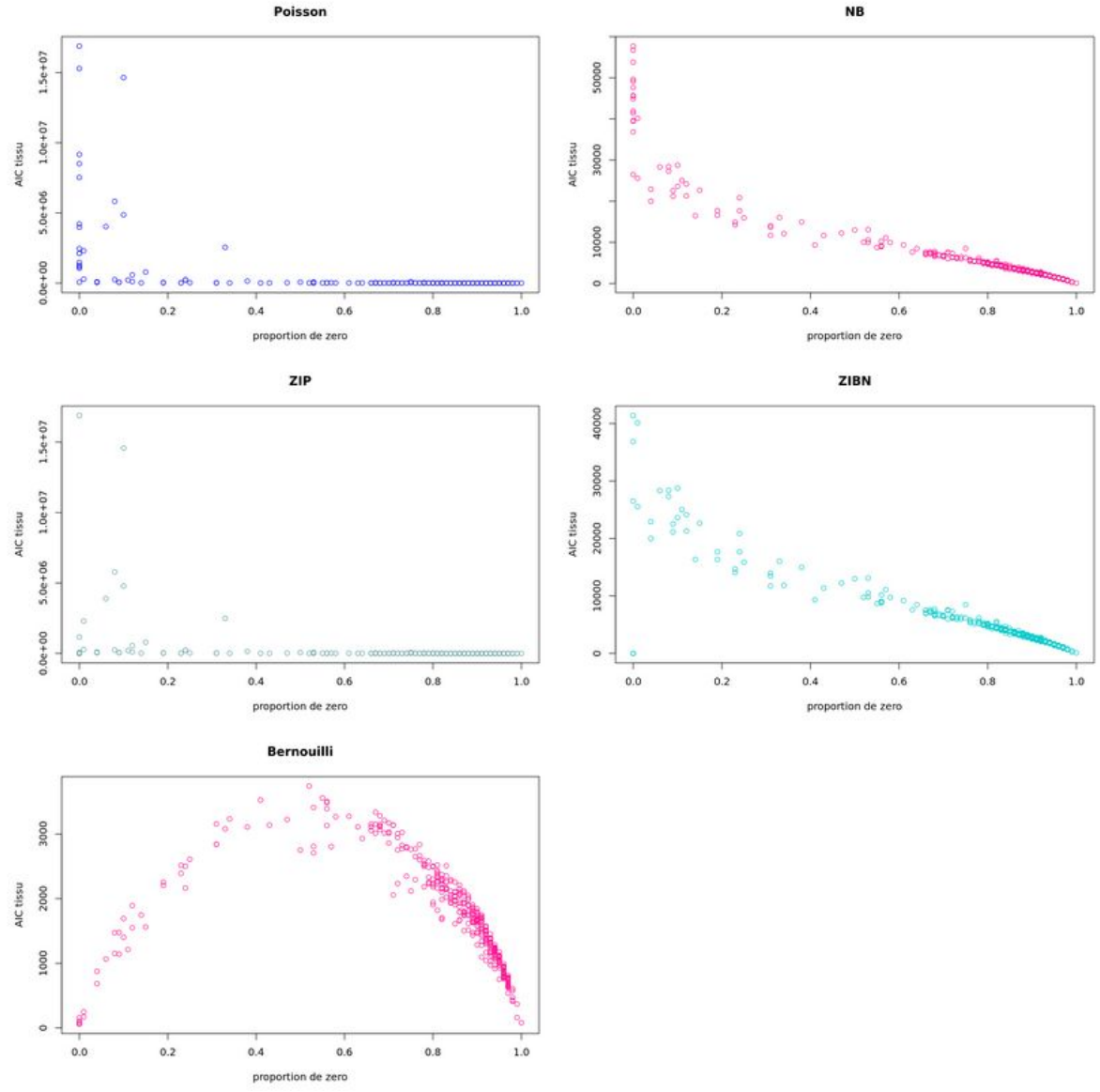
Vuong: ZINB vs NB, genes random

	ZINB	NB	NoTest	NoInflation
Total	142	45	24	173
Significatif	13	1	24	173

On peut voir qu'il y a très peu de tests significatifs. Pour les gènes *olfa*, seulement 33 tests sont significatifs sur 227 qui montrent que le modèle ZINB est significativement meilleur que NB(soit environ 13%). Pour les gènes pris au hasard, il y a 13 tests qui montrent que le modèle ZINB est meilleur que ZINB soit 7%. Ces tests sont effectués avec un seuil très pessimiste expliquant le peu de test significatif.

4.4. AIC et proportion de zéro

Intéressons nous maintenant à l'évolution de l'AIC pour chacun des modèles. Pour chaque modèle (avec prise en compte du tissu comme variable explicative) l'évolution de la vraisemblance en fonction de la proportion de zéro est représentée graphiquement.



On observe dans l'ensemble qu'il y a une tendance: l'AIC diminue quand la proportion de zéro augmente. Ceci est d'autant plus marqué pour le modèle ZINB et NB. Pour le Bernoulli, on observe une diminution de l'AIC en s'éloignant de la proportion de zéro égale à 0.5
Il est important de noter que les AIC les plus faibles sont observés pour les modèles de Bernoulli, ZINB, NB puis ZIP.

En conclusion, on a observé que la modélisation des gènes est plus adaptée avec une loi Négative Binomiale que Poisson compte tenu de la forte dispersion des données. Mais ces deux modèles trouvent leur limite lorsqu'il y a une forte proportion de zéro dans les données.

Pour estimer cette proportion de zéro, le modèle de Bernoulli est retenu comme étant le meilleur. Les modèles ZIP, ZINB, NB permettent d'estimer cette proportion avec une faible erreur. Ce résultat est également vérifié et valable pour des gènes pris au hasard.

En ce qui concerne l'influence du tissu dans la modélisation des données, les modèles anova, NB, ZIP, ZINB et Bernoulli suggèrent l'influence significative sur l'expression des gènes *olfa* dans 80% des cas à une erreur près.
Le modèle de Poisson donne un résultat plutôt élevé(99% de significativité) comparé aux autres modèles , nous laissant croire qu'il est le modèle le moins adapté à nos données .
Il est à noter qu'en cas de dispersion et de sur-représentation de zéros dans les données, les écarts type des coefficients sont sous estimés par la régression de Poisson, faussant les tests de significativité (les coefficients semblent tous significatifs). Cf (Ricco Rakotomalala, Régression de Poisson)

Pour les gènes pris au hasard, toutes les approches classiques, anova, Poisson, NB, ZIP, ZINB détectent une influence du tissu de façon significative pour environ 96 % des gènes. L'influence des tissus pour les gènes pris au hasard est significative dans plus de cas que pour les gènes *olfa*.
Le modèle de Bernoulli, par son approche de catégorisation en deux classes (soit 0 soit 1) discriminent les gènes pris hasard, seulement 42% sont significatifs alors que c'est 80% pour le gènes *olfa* .

Dans l'ensemble des modélisations, il y a donc une différence de significativité de l'influence du tissu entre les gènes *olfa* et les gènes pris au hasard. Cette différence est d'autant plus grande en ce qui concerne le modèle de Bernoulli.

Conclusion du projet

Au terme de ce projet, il a été montré la limite du modèle de Poisson et de la Négative Binomiale dans la prise en compte des masses de zéro dans les données de comptage. L'approche par des modèles à inflation de zéro est plus adaptée.
Cependant, ces modèles à inflation zéro ne peuvent pas être utilisés pour l'ensemble des données compte tenu de la nécessité d'avoir au minimum une observation nulle. Or certains gènes sont exprimés partout même faiblement et ne peuvent donc pas être modélisés par cette loi.

Ainsi, l'utilisation d'une modélisation par un modèle Bernoulli a permis de modéliser la particularité des gènes *olfa* et de leur forte proportion de zéro tout en étant adapté à l'ensemble des données. L'autre avantage majeur de cette modélisation est qu'elle est libre de toute normalisation.

Il est nécessaire de préciser les limites de ce travail. En effet les tests effectués, nécessitent la vérification en amont de certaines conditions de validité (sur les résidus, l'ajustement des modèles utilisés par rapport aux données) qui n'a pas été menée.

Il serait ainsi préférable de vérifier les conditions d'application afin de rendre les résultats de ces tests plus robustes avant de poursuivre la recherche d'autres gènes présentant le même profil que les gènes *olfa*.
Il serait également intéressant d'intégrer d'autres variables explicatives dans la modélisation.

BIBLIOGRAPHIE

[1] Ricco Rakotomalala ZIP regression 2020 March 29

[2] https://fr.wikipedia.org/wiki/Loi_binomiale_n%C3%A9gative Consulté le 10 Avril 2020

[3] SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data
Hugo Varet, Loraine Brillet-Guéguen, Jean-Yves Coppée, Marie-Agnès Dillies
Plos One 2016

[4] Theory Biosci. 2012 Dec;131(4):281-5. doi: 10.1007/s12064-012-0162-3. Epub 2012 Aug 8.
Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.
Wagner GP1, Kin K, Lynch VJ.

[5] BMC Bioinformatics. 2015 Oct 28;16:347. doi: 10.1186/s12859-015-0778-7.
Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data.
Li P1, Piao Y2, Shon HS3, Ryu KH4.

[6] J Biomed Inform. 2018 Sep;85:80-92. doi: 10.1016/j.jbi.2018.07.016. Epub 2018 Jul 21.
How does normalization impact RNA-seq disease diagnosis?
Han H1, Men K2.

[7] Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2
Michael I Love, Wolfgang Huber & Simon Anders
Genome Biology volume 15, Article number: 550 (2014)

[8] Bioinformatics. 2010 Jan 1; 26(1): 139–140.
edgeR: a Bioconductor package for differential expression analysis of digital gene expression data
Mark D. Robinson,1,2,*† Davis J. McCarthy,2,† and Gordon K. Smyth2

[9] Statistical analysis of variability in TnSeq data across conditions using zero-inflated negative binomial regression
Siddharth Subramaniyam, Michael A. DeJesus, Anisha Zaveri, Clare M. Smith, Richard E. Baker, Sabine Ehrt, Dirk Schnappinger, Christopher M. Sassetti & Thomas R. Ioerger
BMC Bioinformatics volume 20, Article number: 603 (2019) Cite this article

[10] zingeR: unlocking RNA-seq tools for zero-inflation and single cell applications
Koen Van den Berge, Charlotte Soneson, Michael I. Love, Mark D. Robinson, Lieven Clement
doi: https://doi.org/10.1101/157982
bioRxiv posted June 30, 2017.

[11] Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq.
Ran D, Daye ZJ
Nucleic Acids Research, 01 Jul 2017, 45(13):e127

[12] The Genotype-Tissue Expression (GTEx) project
John Lonsdale, Jeffrey Thomas, Helen F Moore
Nature Genetics volume 45, pages580–585(2013)

[13] https://www.gsea-msigdb.org/gsea/msigdb/cards/KEGG_OLFACTORY_TRANSDUCTION consulté en mars 2020

[14] David I. Warton & Mitchell Lyons & Jakub Stoklosa1 & Anthony R. Ives3 Three points to consider when choosing a LM or GLM test for count data

[15] de Jong & Heller Generalized Linear Models for Insurance Data sections 6.2 et 6.3,2008

[16] Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear model. Ann. Statist., 13 :342–368, 1985.
[17] Mullahy, J. (1986). Specification and testing of some modified count data. Journal of Econometrics(33), 341-365.
[18] MacDonald, J. M., & Lattimore, P. K. (2010). Count Models in Criminology Handbook of Quantitative Criminology (pp. 683-698): Springer.
[19] Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: a tutorial on count regression and zero-inflated models. Journal of family psychology : JFP : journal of the Division of Family Psychology of the American Psychological Association (Division 43), 21(4), 726-735.
[20] Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica, 57(2), 307-333

ANNEXES

ANNEXE 1: Tables des co-facteurs

Repartition des echantillons par sex

sex	Freq
F	1014
M	1739

Repartition des echantillons par tissu

tissue	Freq
mouth	5
female_genital	7
kidney	7
bladder	11
small_intestine	16
pituitary_gland	23
liver	31
vagina	31
spleen	33
adult_ovary	36
prostate	40
uterus	42
haematopoietic_div	49
adrenal_gland	52
testis	62
breast	63
pancreas	65
stomach	75
colon	80
thyroid_gland	106
bronchus_lung	122
central_nervous_system	122
muscles	139
connective_tissues	140
fibroblast	150
skin	155
blood	162
heart	171
artery	214
esophagus	226
brain	318

Repartition des echantillons par tissu et par sexe

	F	M
adrenal_gland	20	32
adult_ovary	36	0
artery	72	142
bladder	5	6
blood	52	110
brain	116	202
breast	33	30
bronchus_lung	43	79
central_nervous_system	47	75
colon	33	47
connective_tissues	44	96
esophagus	84	142
female_genital	7	0
fibroblast	54	96
haematopoietic_div	11	38
heart	56	115
kidney	0	7
liver	10	21
mouth	1	4
muscles	50	89
pancreas	18	47
pituitary_gland	9	14
prostate	0	40
skin	57	98
small_intestine	2	14
spleen	10	23
stomach	29	46
testis	0	62
thyroid_gland	42	64
uterus	42	0
vagina	31	0