Etude de la distribution du nombre de transcrits par gouttelettes

Léonard Gousset
Louis Allain
Julien Heurtin

Projet statistique - Groupe 37

Mai 2023

Tuteur : Emmanuel Curris
Coach : Julie Serieys

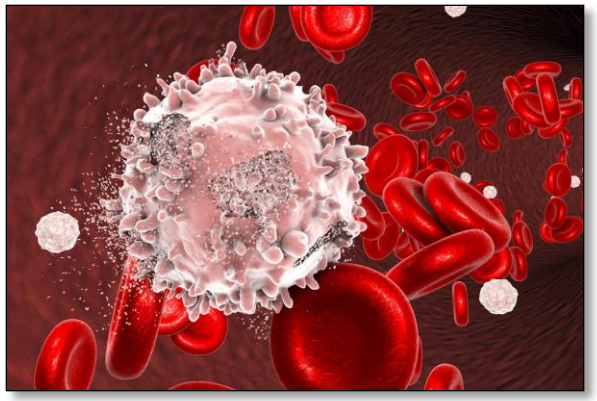
SOMMAIRE

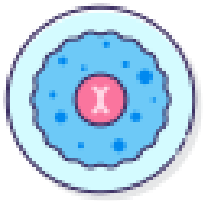
- **Introduction**
- **1) Approche exploratoire**
  - 1.1) Origine/structure des données
  - 1.2) Analyse du nombre de transcrits
  - 1.3) Choix des lois
- **2) Une première classification**
  - 2.1) La construction du modèle
  - 2.2) Les résultats du modèle
  - 2.3) Les limites de ce modèle
- **3) Améliorations du modèle**
  - 3.1) Modélisation de la bosse
  - 3.3) Modélisation sans la loi de Poisson
  - 3.4) Changement des lois normales
- **Conclusion**

INTRODUCTION

Leucémie :

- Cancer du sang
- Production anormale de cellule sanguine
- Greffe de moelle osseuse

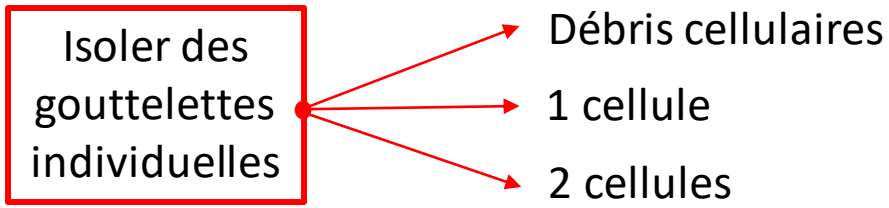




Etudier la régénération des cellules souches après une greffe de moelle osseuse.

INTRODUCTION

➢ "single cell RNA-seq".



Objectif du projet

- **Filtrer les gouttelettes ayant une seule cellule**

➢ Loi de mélange

1) Approche exploratoire

1.1) Origine/structure des données

Gouttelettes à 0 transcrits

	Transcrit n°1	Transcrit n°2	…	Transcrit n°33 538
Gouttelette n°1	1	0	…	4
Gouttelette n°2	0	0	…	5
…	…	…	…	…
Gouttelette n°734 472	12	58	…	0

	Somme des transcrits
Gouttelette n°1	5
Gouttelette n°2	24
…	…
Gouttelette n°327 395	5008

1 seule variable discrète à valeur dans N*

1.2) Analyse du nombre de transcrits

Echelle logarithmique

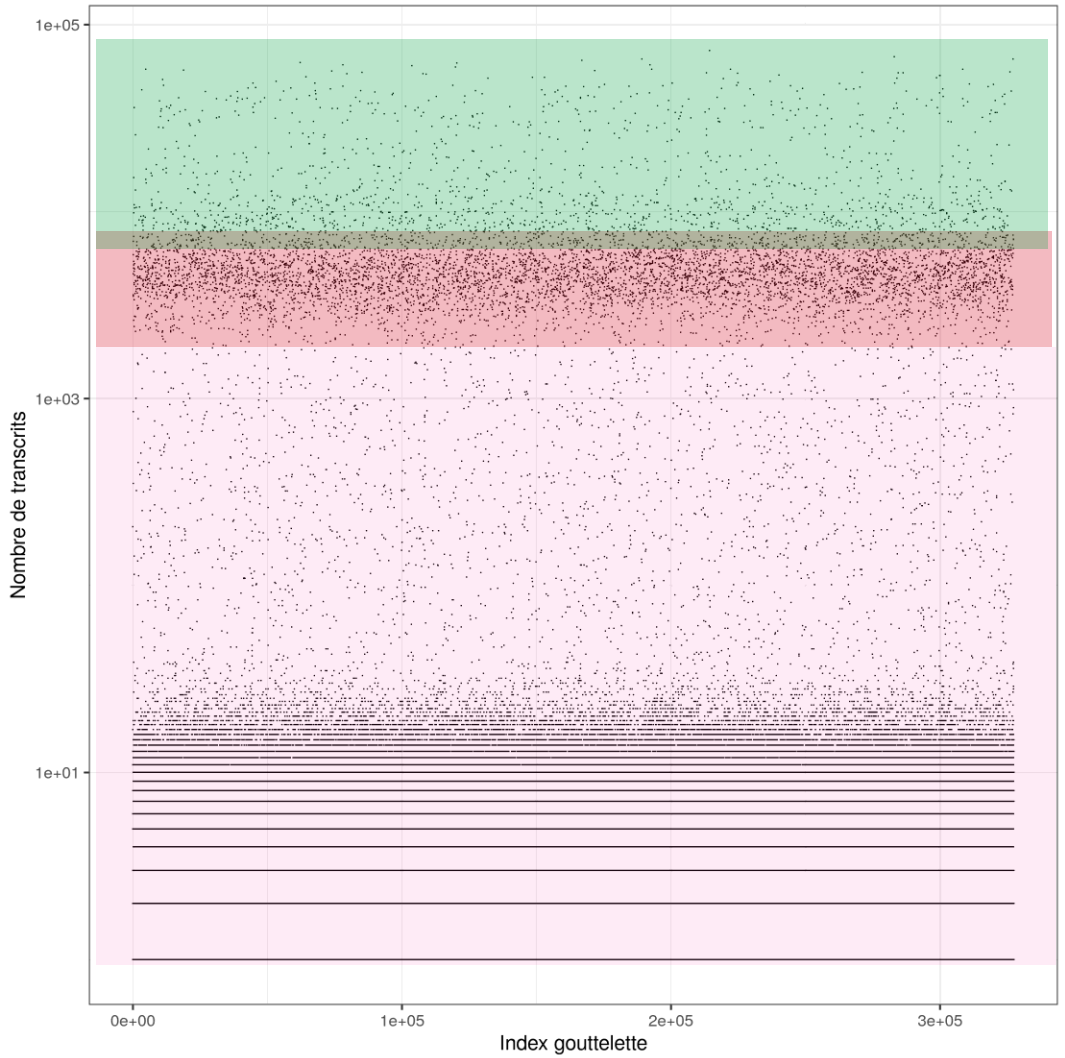


Figure 1 : Nombre de transcrits par gouttelettes

Groupe 3 : **2 cellules** (6000 ~ 73000 transcrits)

~1% des données

Groupe 2 : **Cellule unique** (2000 ~ 7000 transcrits)

~9% des données

Groupe 1 : **Débris cellulaire** (1 ~ 2000 transcrits)

~90% des données

Répartition de nos données en différents groupes.

1.3) Choix des lois

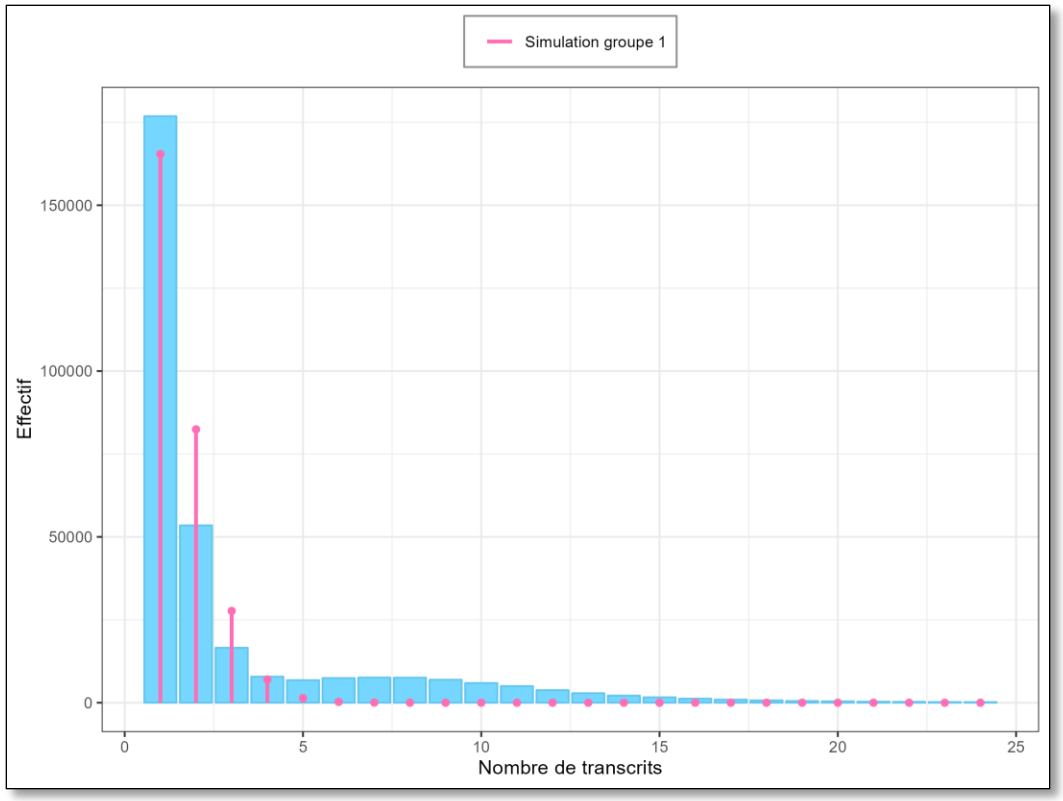


Figure 3 : Superposition données réelles et fonction de masse d'une loi de Poisson (λ =1)

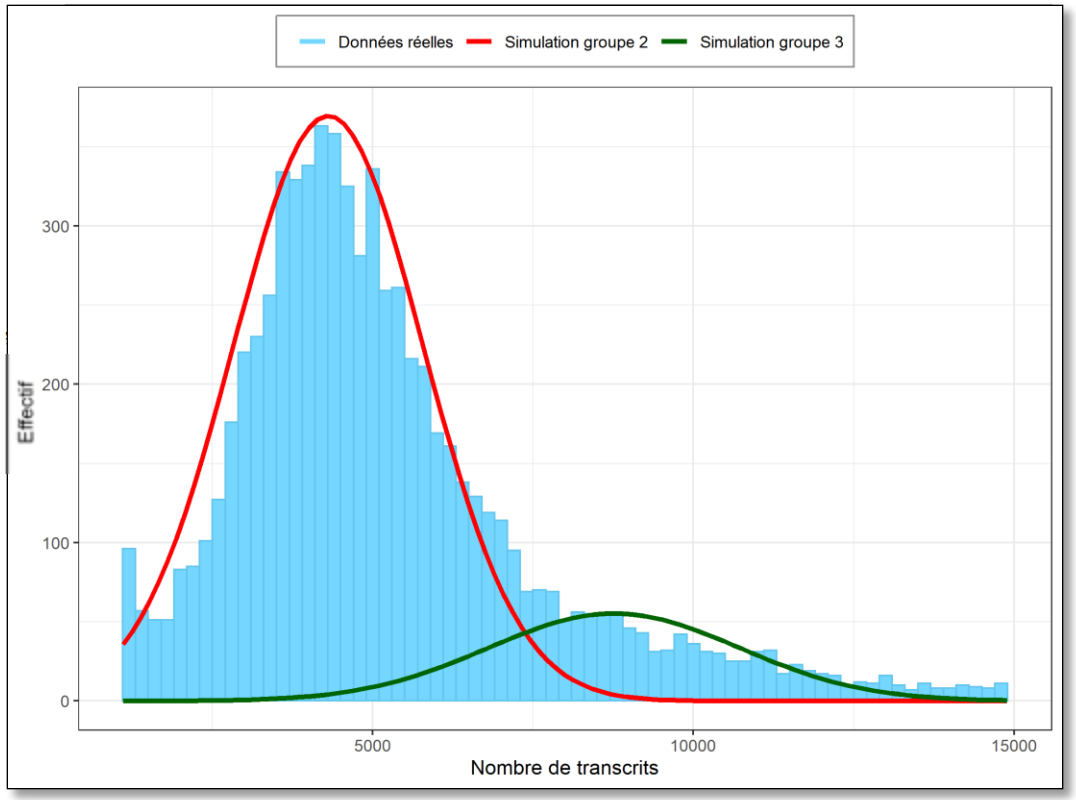


Figure 4 : Superposition données réelles et densités des lois normales

La loi de Poisson et les 2 lois normales semblent bien représenter nos données, hormis la bosse entre 5 et 15 transcrits.

2) Une première classification

2.1) La construction du modèle

$$p(x_i, \theta) = \sum_{k=1}^{K} \pi_k p_k(x_i, \alpha_k)$$

Avec θ = {{πk, αk} : k = 1, ..., K} l'ensemble des paramètres du modèle et αk les paramètres de la loi pk.

$$p_1(x_i, \alpha_1) = p_1(x_i, \lambda) = \frac{\lambda^{x_i}}{x_i!\,(e^{\lambda} - 1)}$$

$$p_2(x_i, \alpha_2) = p_2(x_i, \mu, \sigma) = \begin{cases} \Phi\left(\frac{\left(x_i + \frac{1}{2}\right) - \mu}{\sigma}\right) - \Phi\left(\frac{\left(x_i - \frac{1}{2}\right) - \mu}{\sigma}\right), & \text{si } x_i \neq 1 \\ \Phi\left(\frac{\left(x_i + \frac{1}{2}\right) - \mu}{\sigma}\right), & \text{sinon} \end{cases}$$

$$p_3(x_i, \alpha_3) = p_3(x_i, \mu, \sigma) = \begin{cases} \Phi\left(\frac{\left(x_i + \frac{1}{2}\right) - 2\mu}{\sqrt{2}\sigma}\right) - \Phi\left(\frac{\left(x_i - \frac{1}{2}\right) - 2\mu}{\sqrt{2}\sigma}\right), & \text{si } x_i \neq 1 \\ \Phi\left(\frac{\left(x_i + \frac{1}{2}\right) - 2\mu}{\sqrt{2}\sigma}\right), & \text{sinon} \end{cases}$$

2.2) Les résultats du modèle

Statistiques \ Groupes	Groupe 1	Groupe 2	Groupe 3
Minimum	1	12	15558
Moyenne	2, 4 (±2, 5)	1550 (±2714)	31313 (±12039)
Maximum	11	15696	72639
Population	301921	24963	511
Proportion	92,2%	7,6%	0,2%

2.2) Les résultats du modèle

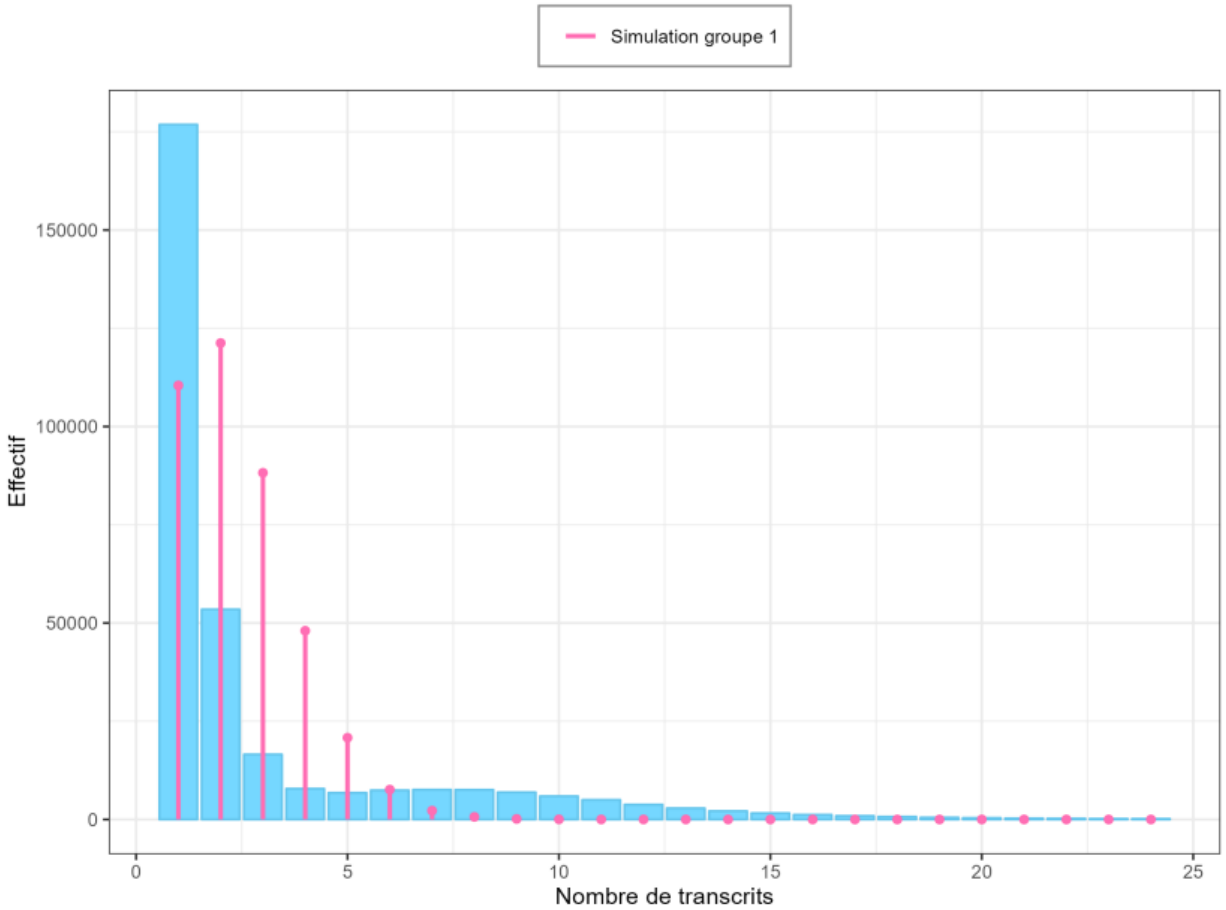


Figure 5 : Superposition données réelles et fonction de masse d'une loi de Poisson avec les paramètres de notre modèle (λ =2,4)

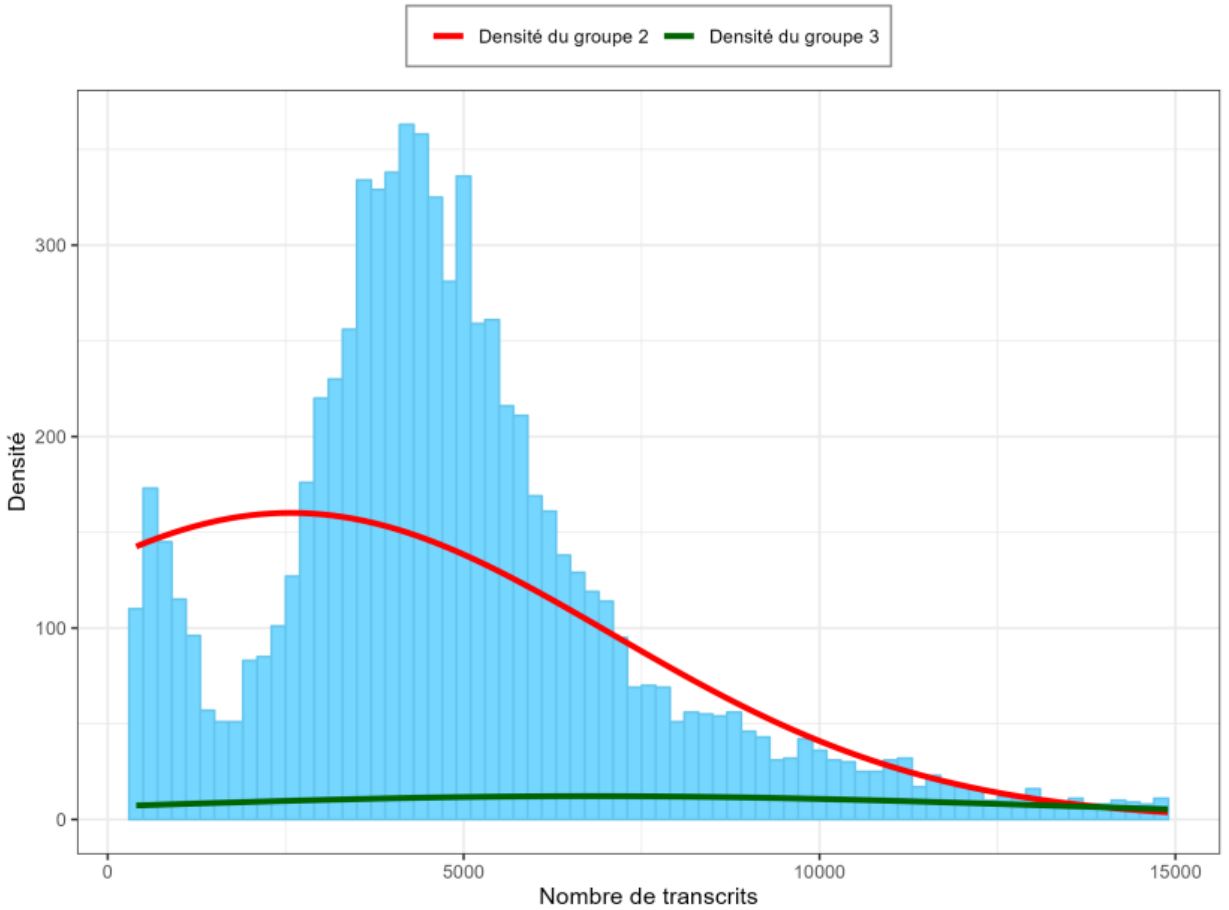


Figure 6: Superposition données réelles et densités des lois normales de notre modèle

2.3) Les limites de nos résultats

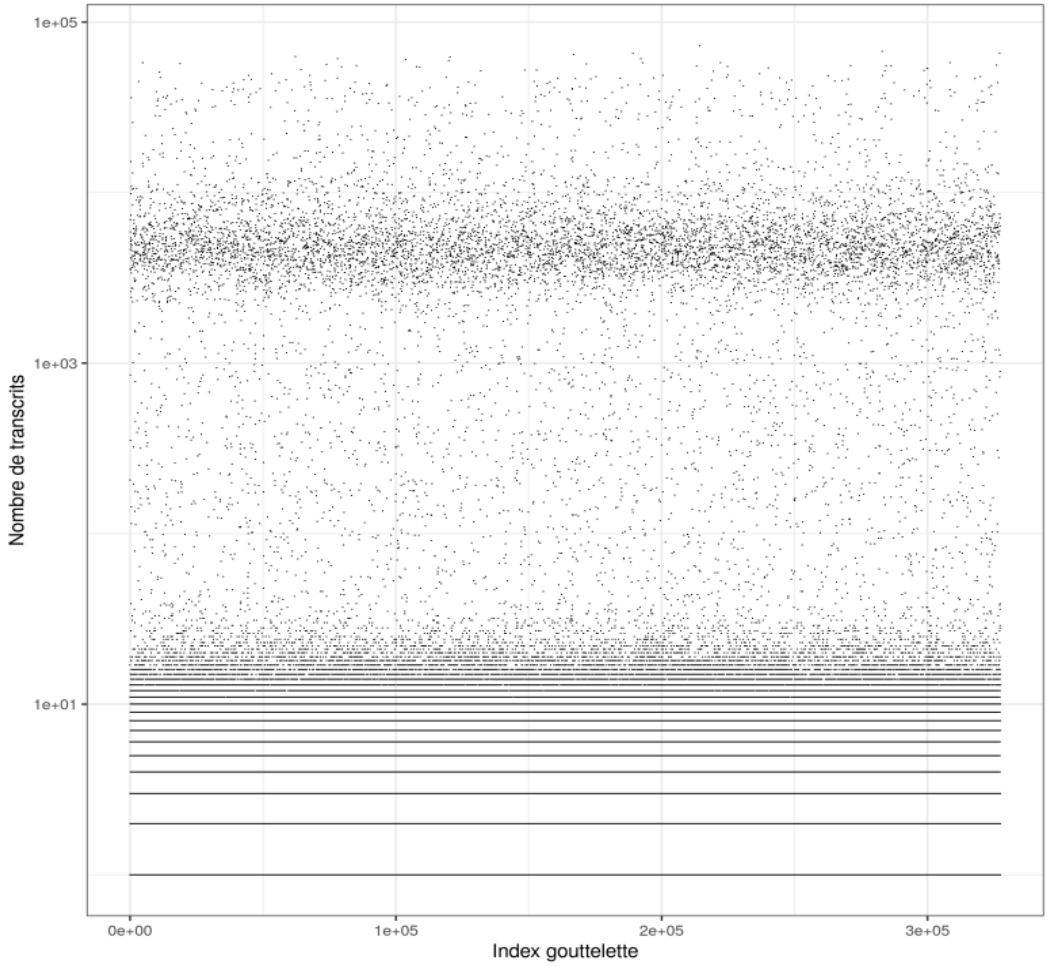


Figure 7 : Distribution des données réelles avec échelle logarithmique

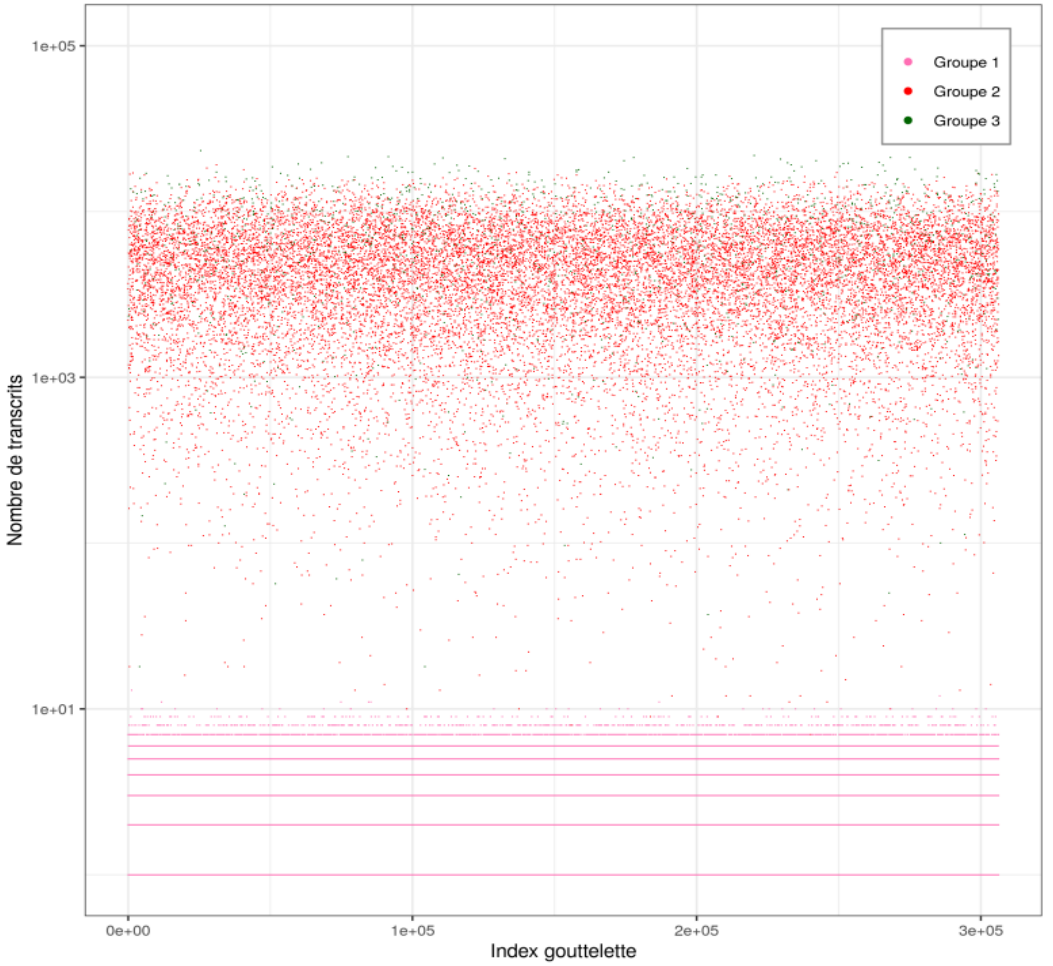


Figure 8: Distribution des données échantillonnées à partir des lois obtenues dans notre modèle

3) Améliorations de la modélisation

3.0) Rappels

Pour un modèle m, le critère BIC s'écrit : $$BIC(m) = l(x, \hat{\theta}_m) - \frac{\nu_m}{2} \ln(n)$$

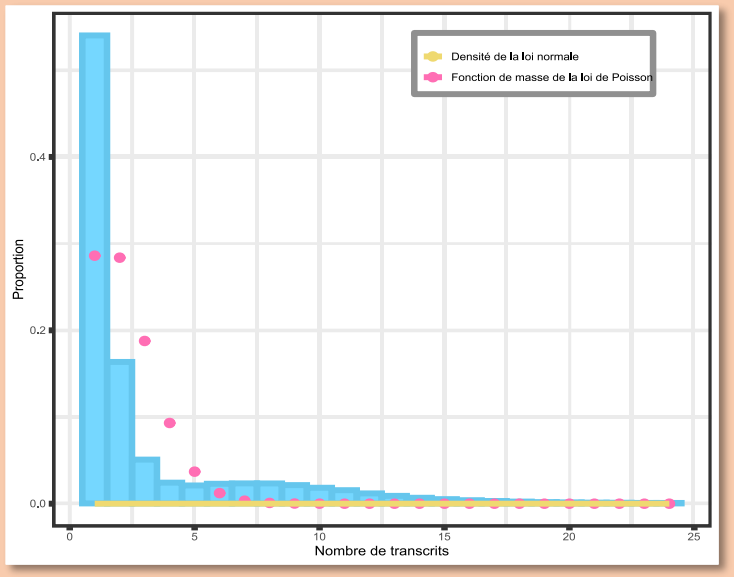
Avec :

- ν_m ➢ Le nombre de paramètres du modèle m
- $l(x, \hat{\theta}_m)$ ➢ La log-vraisemblance complétée du modèle
- n ➢ Le nombre d'individu sur lesquels est entrainé le modèle

➢ **BIC = -982 853**

3.1) Modélisation de la bosse

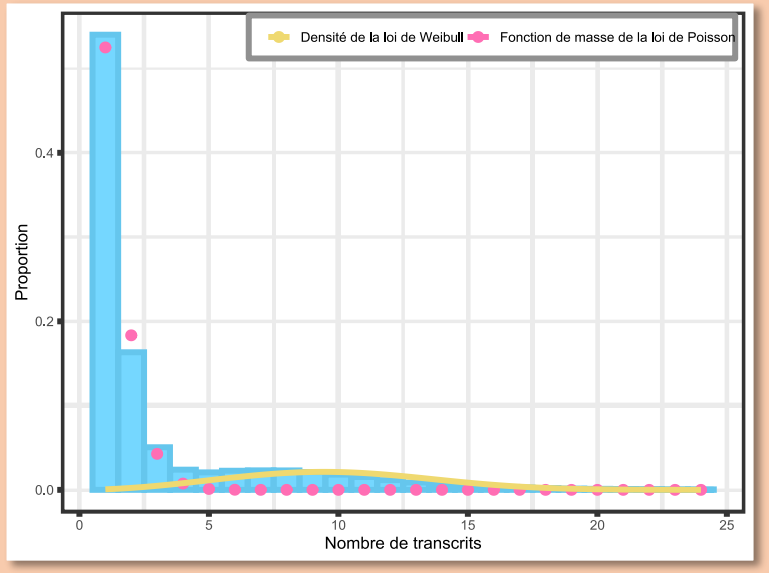
Loi normale



Superposition d'une loi de Poisson et d'une loi normale sur les données

➢ **BIC = -966 903**

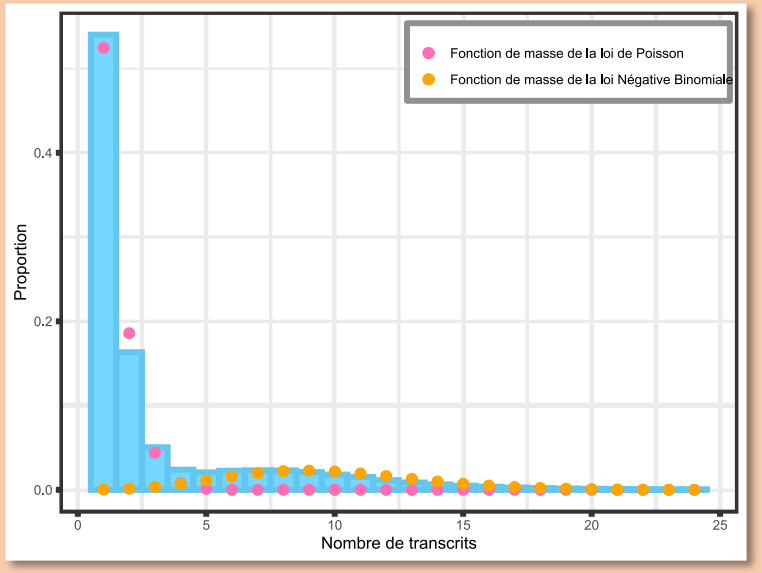
Loi de Weibull



Superposition d'une loi de Poisson et d'une loi de Weibull sur les données

➢ **BIC = -715 670**

Loi binomiale négative

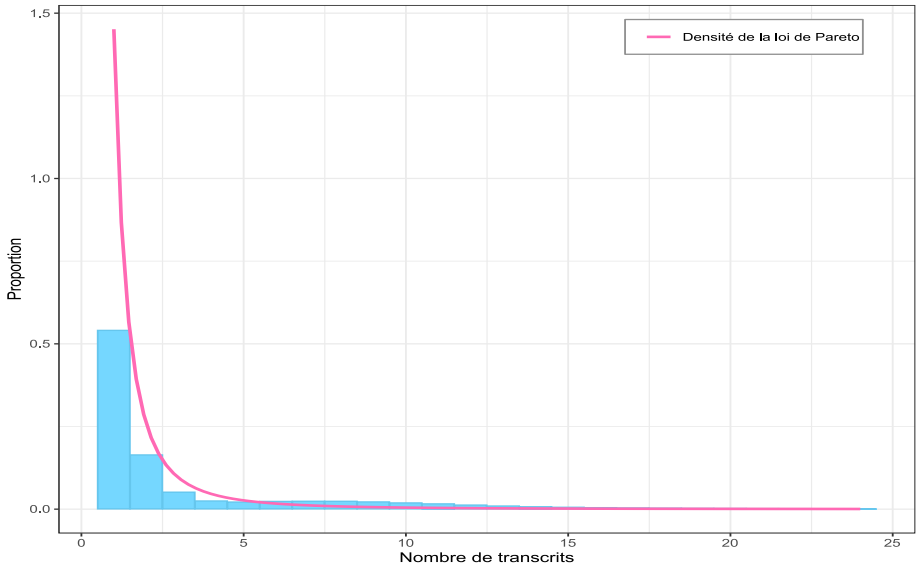


Superposition d'une loi de Poisson et d'une loi binomiale négative sur les données

➢ **BIC = -708 161**

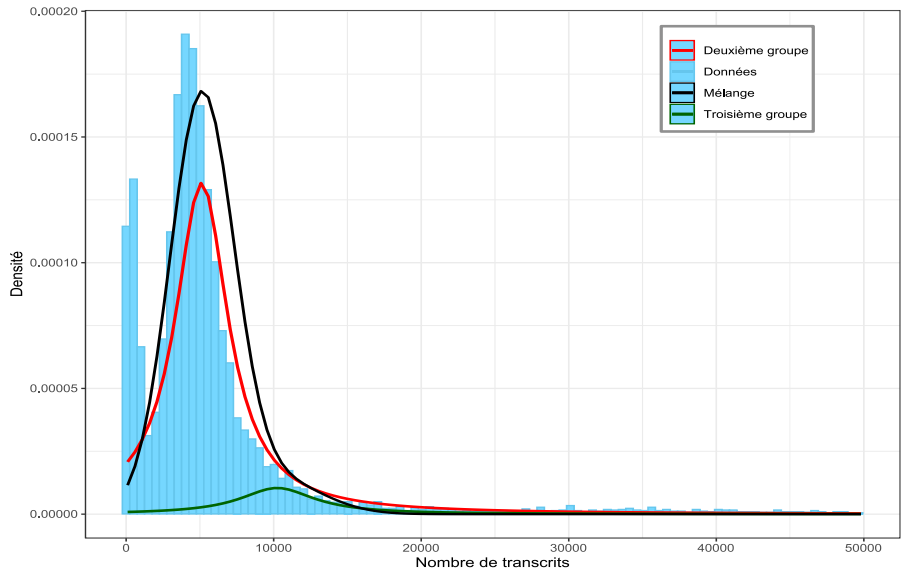
3.3) Modélisation sans la loi de Poisson

Modèle de mélange avec une loi de Pareto et deux lois normales



Superposition de la loi de Pareto sur les données

- **BIC = -520 564**



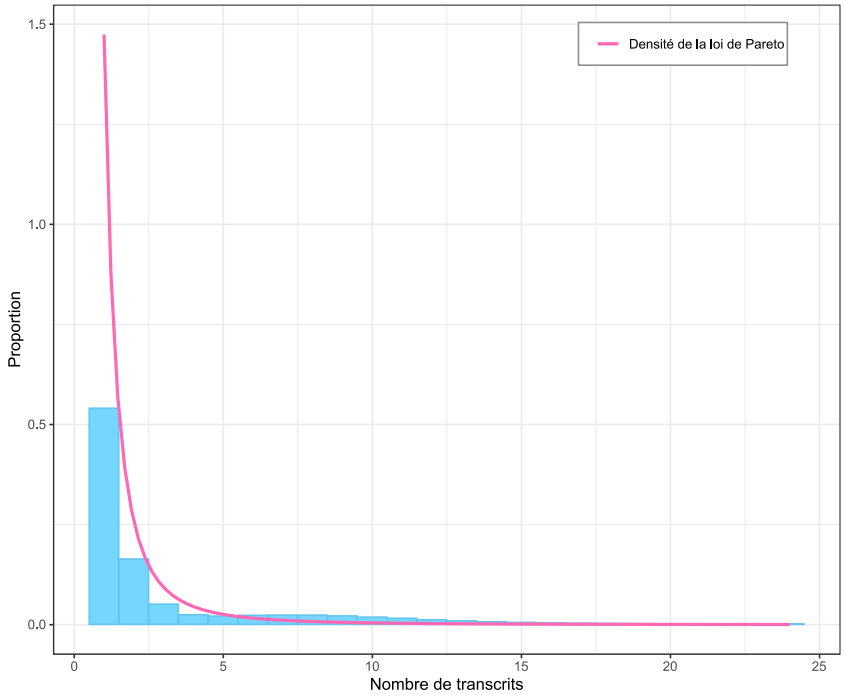
Superposition des deux lois normales sur les données

La loi de Pareto modélise **mieux** les gouttelettes contenant des débris.

Les lois normales sont encore un peu décalées.

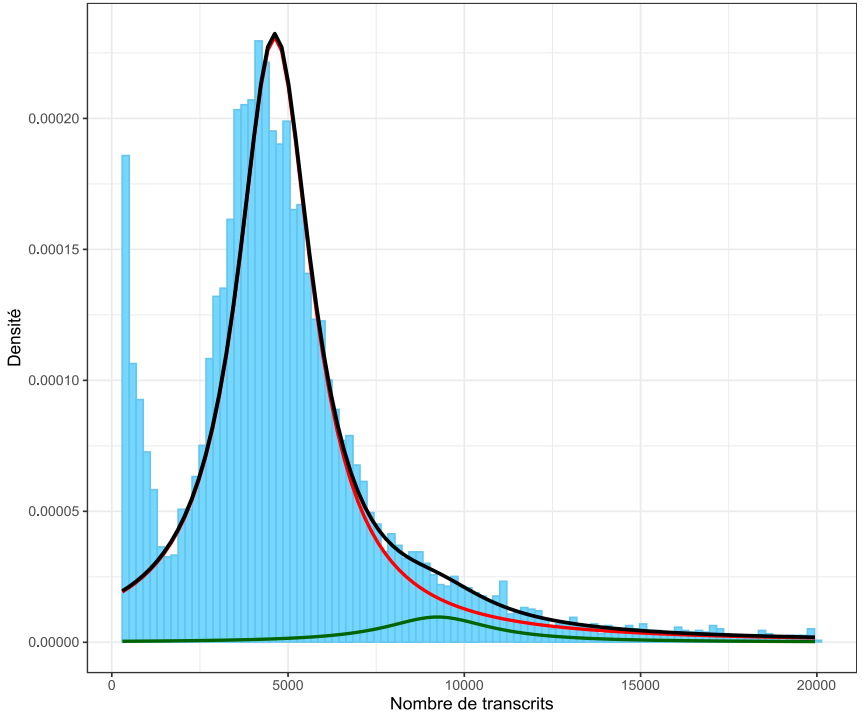
3.4) Changement des lois normales

Modèle de mélange avec une loi de Pareto et deux lois de Cauchy



Superposition de la loi de Pareto sur les données

- **BIC = -518 337**



Superposition des deux lois de Cauchy sur les données

Les lois de Cauchy semblent **mieux** correspondre aux données que les lois normales.

Bien qu'encore imparfait, regardons si nous arrivons à isoler les gouttelettes à une seule cellule.

3.4) Changement des lois normales

Modèle de mélange avec une loi de Pareto et deux lois de Cauchy

➢ **BIC = -518 337**

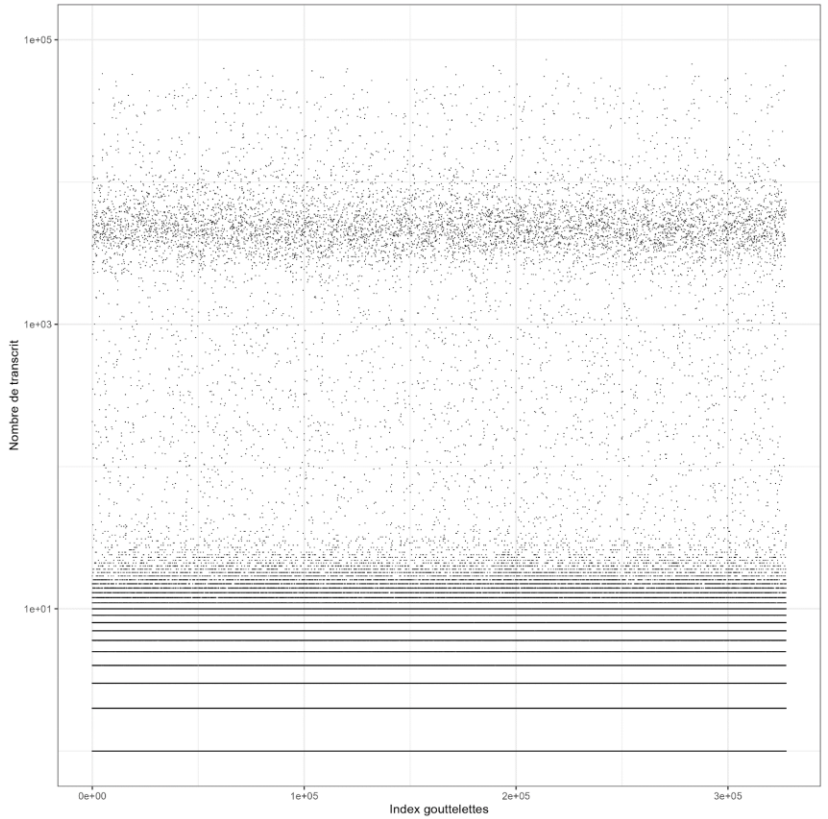
La règle choisie pour décider de l'appartenance d'une gouttelette au groupe 2 ou non est la suivante : $\pi_2 > 0.99 \quad \Rightarrow \quad G.2$

	Min	Moyenne	Max
Gouttelettes choisies	2 933	4 179 (±657)	5 341

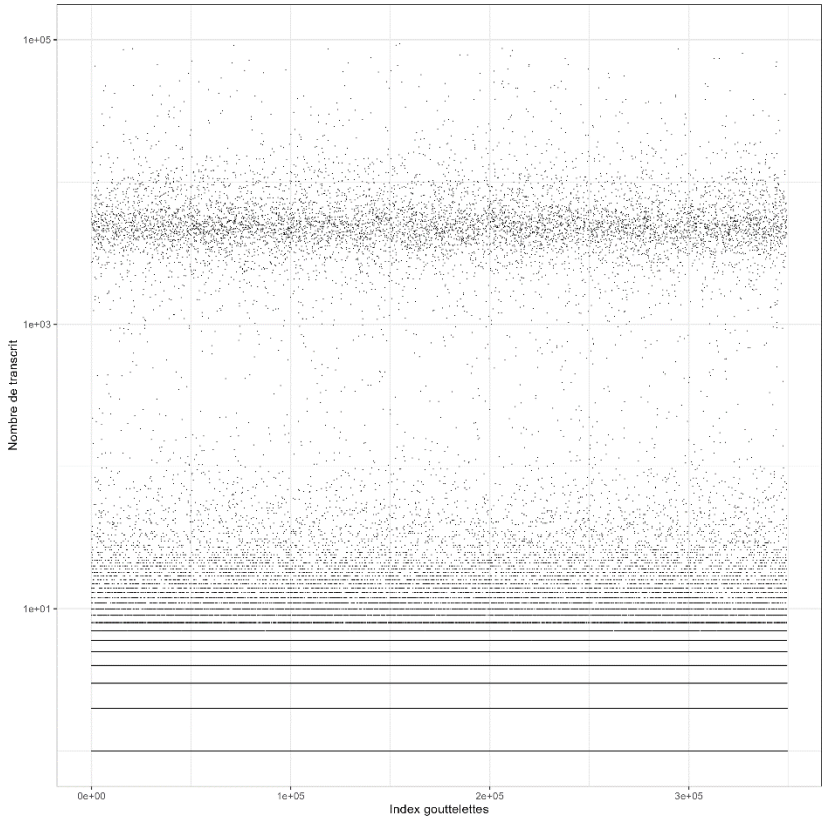
Nous avons donc isolé 3 677 gouttelettes, qui sont censées ne contenir qu'une unique cellule.

3.4) Changement des lois normales

Modèle de mélange avec une loi de Pareto et deux lois de Cauchy



Distribution des données réelles avec échelle logarithmique



Distribution des données échantillonnées à partir des lois obtenues dans notre modèle

CONCLUSION

Objectif

- Filtrer les gouttelettes ayant une seule cellule par une loi de mélange.

Résultats

- Le modèle de mélange permet de **retrouver** ces gouttelettes, avec une probabilité élevée de **réellement contenir** une seule cellule.

Limites

- Sens biologique des lois ?
- Absence de données de réference.
- Importance de certains transcrits

Merci pour votre attention