Étude de la distribution des longueurs des lectures produites par les technologies de 3^{e}génération

Soutenance de stage

Amina Ghoul

INRA - Unité MaIAGE
Encadrant : Jean-François Gibrat

M1 Mathématiques et interactions - UEVE

Présentation du lieu de stage

INRA (Institut national de la recherche agronomique)

- Organisme français de recherche en agronomie fondé en 1946
- Premier institut de recherche agronomique en Europe et deuxième dans le monde
- 8417 chercheurs, ingénieurs et techniciens
- Plus de 200 unités de recherche implantées dans 17 centres en région
- Ambition : assurer une alimentation saine et de qualité, une agriculture compétitive et durable ainsi qu'un environnement préservé et valorisé.

MaIAGE

Unité MaIAGE (Mathématiques et Informatique Appliquées du Génome à l'Environnement)

- Centre Ile-de-France-Jouy-en-Josas qui comprend 29 unités dont l'unité MaIAGE.
- Regroupe des mathématiciens, des informaticiens, des bioinformaticiens et des biologistes autour de questions de biologie et agro-écologie
- Compétences de l'unité : inférence statistique, modélisation dynamique, bioinformatique, l'automatique et l'algorithmique
- les disciplines destinatrices : écologie, environnement, biologie moléculaire et biologie des systèmes.

StatInfOmics

Équipe StatInfOmics (Statistique et Bioinformatique des données Omiques)

- Développer et mettre en oeuvre des méthodes statistiques et bioinformatiques dédiées à l'analyse de données "omiques"
- Essentiellement d'ordre statistique : estimation de distributions, inférence de modèles à variables latentes, prédiction de relations entre jeux de variables, segmentation, visualisation et classification

Objectif du stage

Étude de la distribution des longueurs des lectures produites par les technologies de séquençage de 3^{e} génération.

- 1re et 2^{e} générations : lectures de longueur constante et petite d'au plus 400 bp
- 3^{e} génération : une distribution de longueurs : médiane d'environ 12 kbp et jusqu'à 100 kbp pour les lectures les plus longues

Trouver une expression analytique de la distribution des longueurs des lectures issues du séquençage de troisième génération

Séquençage ADN

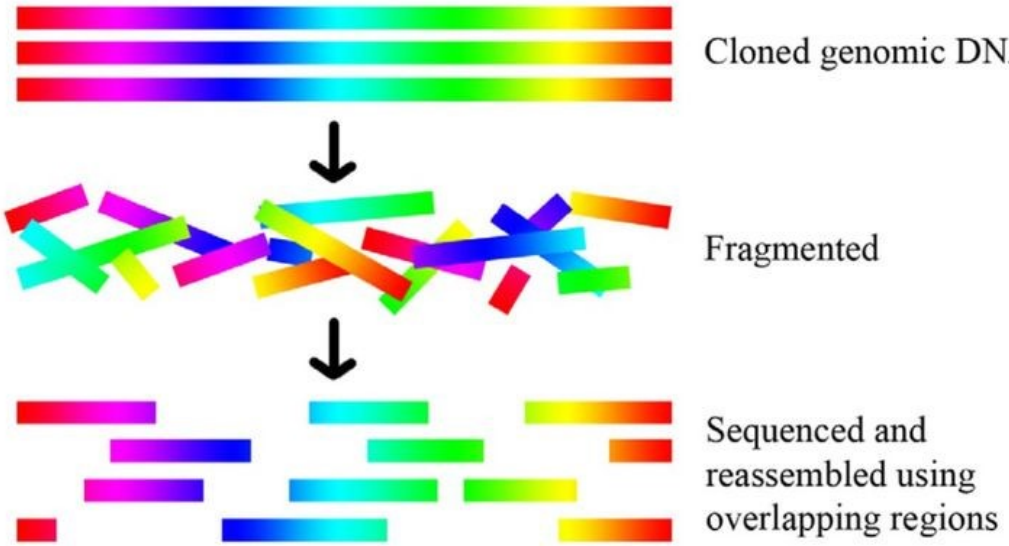
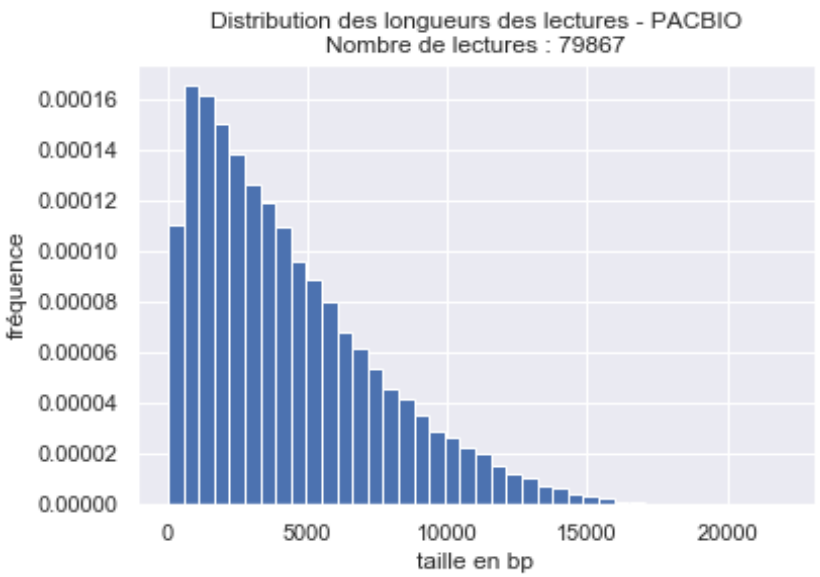
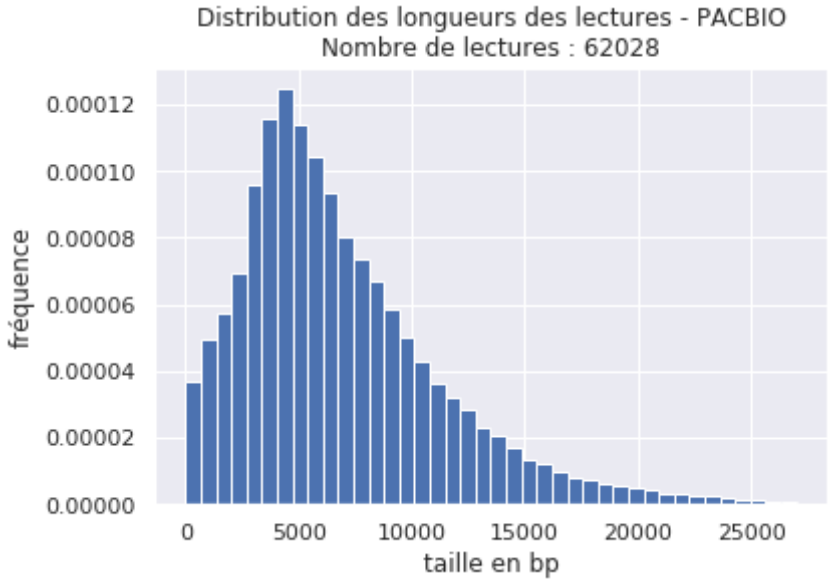


Fig: Shotgun sequencing technique

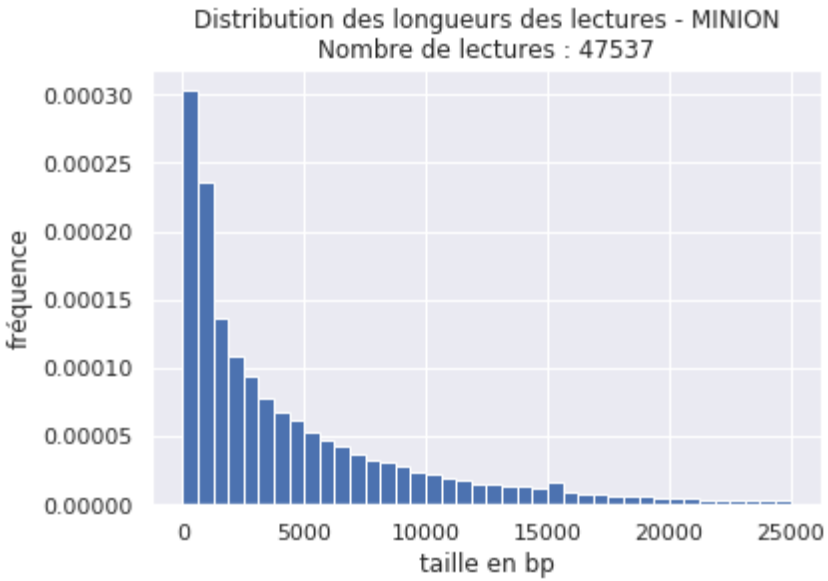
- ADN est cloné, puis fragmenté aléatoirement(shotgun sequencing)
- Fragments introduits dans le séquenceur (PacBio, MinION)

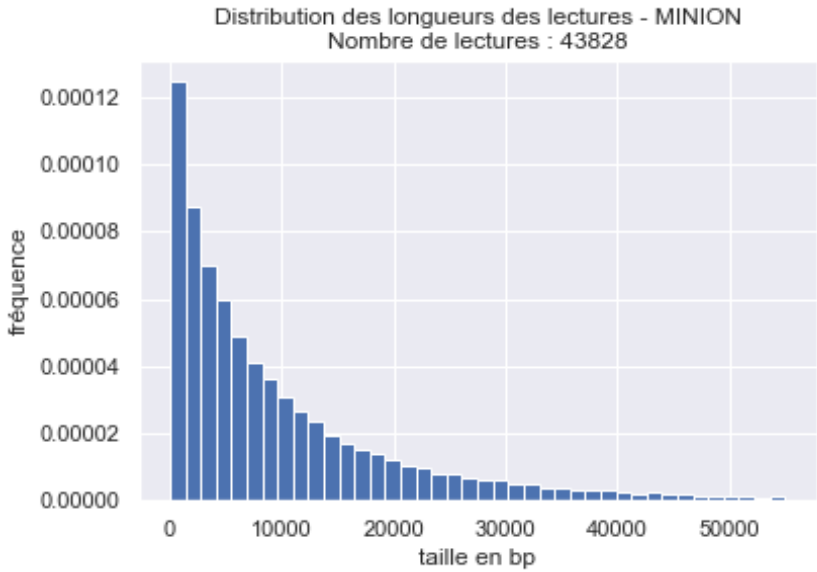
Pacific Biosciences (PacBio)





Oxford Nanopore (MinION)





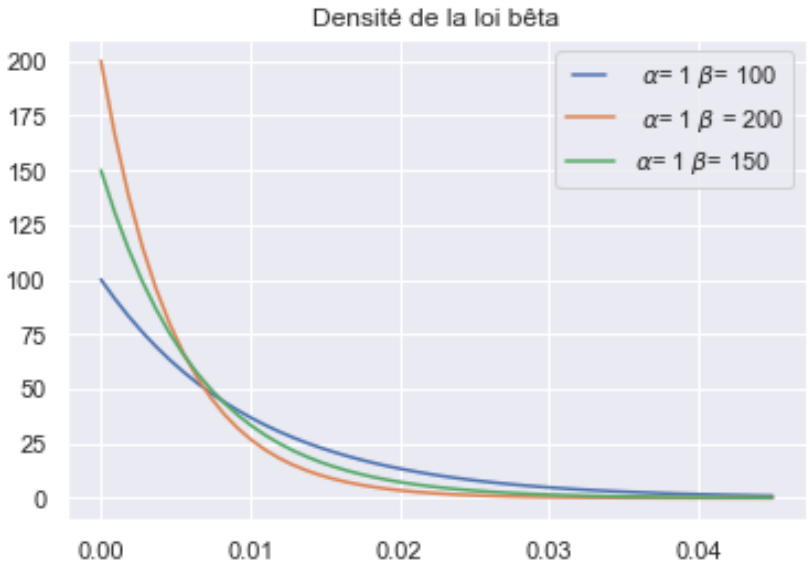
Loi bêta

ADN est fragmenté aléatoirement en N morceaux

Cas d'un bâton coupé en N morceaux aléatoirement

Loi bêta : $B(1, N)$

$p_F(x) = N(1-x)^{N-1}$, $x \in [0, 1]$
$x = \frac{y}{G}$ où y : taille en (bp), G : taille du génome



Loi de N

N est une variable aléatoire

Dépend de la personne qui fragmente l'ADN

Choix de la loi de N :

- Loi normale : $N(\mu, \sigma^2)$, $p_G(n) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(n-\mu)^2}{2\sigma^2}}$ $\mu, \sigma > 0$
- Loi uniforme : $U(a, b)$, $p_G(n) = \frac{1}{b-a}$ $a, b > 0$
- Loi de poisson : $P(\lambda)$, $p_G(n) = \frac{\lambda^n e^{-\lambda}}{n!}$ $\lambda > 0$

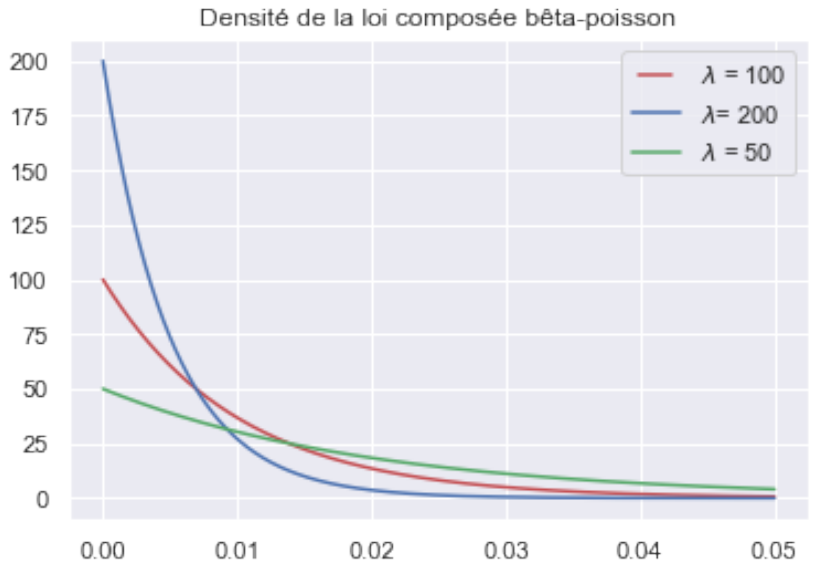
Loi bêta-poisson

Loi composée

$$p_H(x) = \int p_F(x|n)p_G(n)dn$$

Loi bêta-poisson

$$p_H(x) = \sum_{i=0}^{+\infty} n(1-x)^{n-1}\frac{\lambda^n e^{-\lambda}}{n!}$$
$$p_H(x) = \lambda e^{-\lambda x}, \text{ avec x} \in [0,1]$$



Loi bêta-uniforme

Loi composée

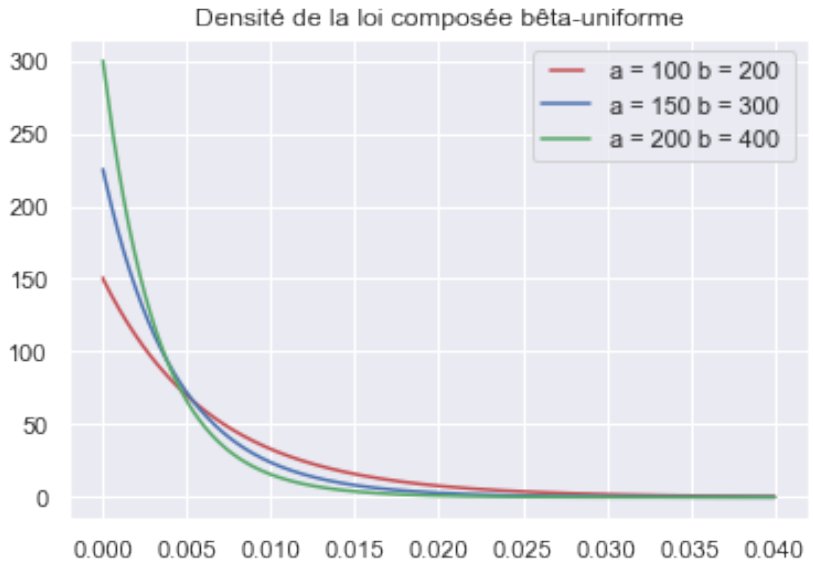
$$p_H(x) = \int p_F(x|n)p_G(n)dn$$

Loi bêta-uniforme

$$p_H(x) = \int_a^b n(1-x)^{n-1}\frac{1}{b-a}dn$$

$$p_H(x) = \frac{e^{-log(1-x)}}{b-a}\left(\frac{be^{blog(1-x)} - ae^{alog(1-x)}}{log(1-x)} + \frac{e^{alog(1-x)} - e^{blog(1-x)}}{(log(1-x))^2}\right)$$

avec x $\in]0,1[$

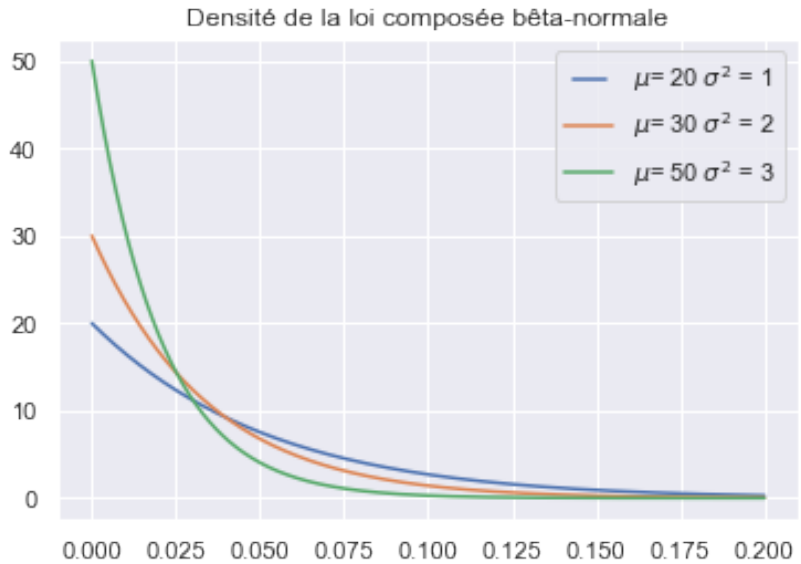


Loi bêta-normale

$$p_H(x) = \int_0^{+\infty} n(1-x)^{n-1} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(n-\mu)^2}{2\sigma^2}} dn$$

$$= \frac{\sigma\pi(\sigma^2 log(1-x)+\mu)e^{\sigma^2 log(1-x)(2(\mu-1)+\sigma^2 log(1-x))+2\mu^2}}{2\sqrt{\pi}} (erf(\frac{\sqrt{2}(\sigma^2 log(1-x)+\mu)}{2\sigma}) + 1)$$

avec x $\in]0,1[$ et $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du$



Loi séquenceur

Influence du séquenceur sur la distribution des tailles des lectures

Dans certains cas, le MinION et le PacBio interrompent en cours de route le séquençage d'un fragment d'ADN

Loi de la distribution finale G(x)

$G(x) = \int_x^G f(x) p_H(l) dl$

G : taille du génome, f(x) : fonction de densité de la loi du séquenceur, $p_H(l) = \lambda e^{-\lambda l}$
Après passage dans le séquenceur, la lecture de taille x, ne peut pas être plus longue que le fragment initial de taille l

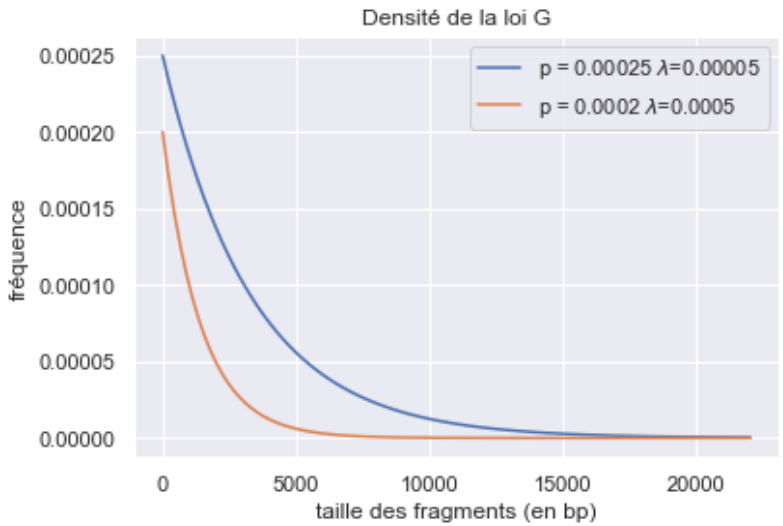
Loi géométrique

$f(x) = p(1-p)^{x-1}$

x : taille en bp, p$\in [0,1]$: probabilité d'interrompre le séquençage. Interruption indépendante de la longueur du fragment introduit dans le séquenceur

Fonction G(x)

$G(x) = p(1-p)^{x-1}(e^{-\lambda x} - e^{-\lambda G})$



Estimation des paramètres-MinION

Log-vraisemblance

$L(x) = \sum_{i=1}^{n} log(p(1-p)^{x_i-1}(e^{-\lambda x_i}) - e^{-\lambda G})$

Expressions des paramètres

- $p = \frac{n}{\sum_{i=1}^{n} x_i}$
- $\sum_{i=1}^{n} \frac{-x_i e^{-\lambda x_i} + G e^{-\lambda G}}{e^{-\lambda x_i} - e^{-\lambda G}} = 0$

  on peut réécrire : $\sum_{i=1}^{n} (G - x_i) \frac{e^{\lambda(G-x_i)}}{e^{\lambda(G-x_i)}-1} = nG$

Optimisation des paramètres-MinION

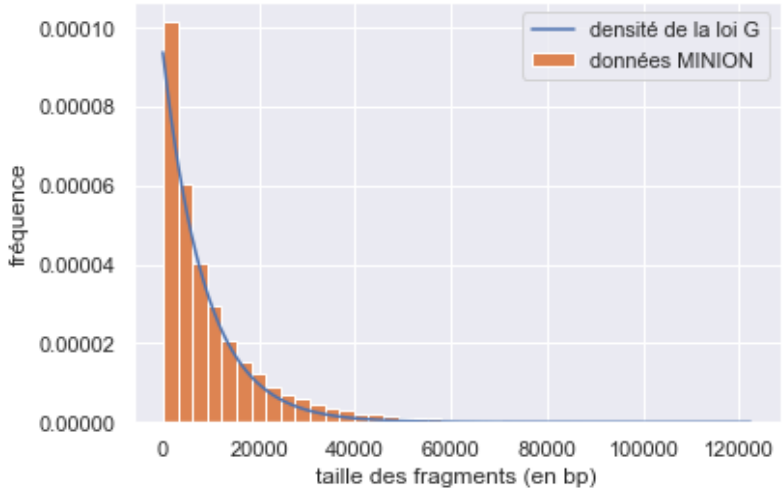


Figure – Représentation de la fonction G(x) et données MinION

Paramètres optimisés

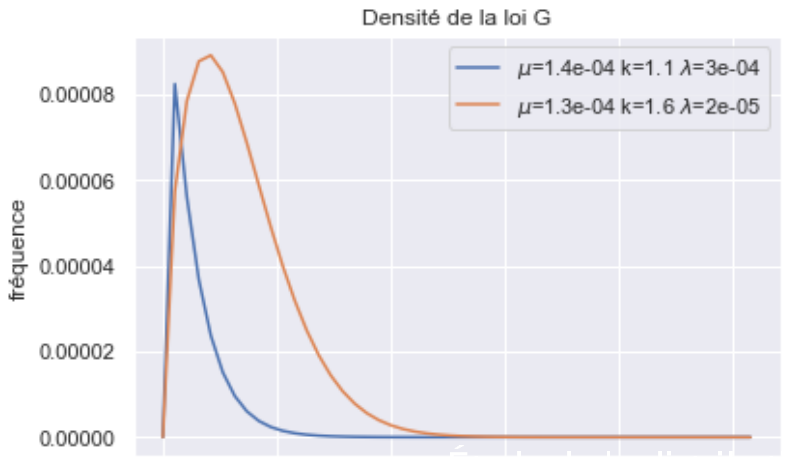
$p = 9.37 * 10^{-5}$ et $\lambda = 2 * 10^{-5}$

Loi de Weibull

$f(x) = \frac{k}{\alpha}(\frac{x}{\alpha})^{k-1}e^{-(\frac{x}{\alpha})^k}$, où k>0, x : taille en bp et $\alpha > 0$
On posera $\mu = \frac{1}{\alpha}$
La probabilité d'interruption du séquençage du fragment introduit croît avec la taille du fragment (k>1).

Loi de G(x)

$G(x) = k\mu(x\mu)^{k-1}e^{-(x\mu)^k}(e^{-\lambda x} - e^{-\lambda G})$
x : taille en bp



Estimation des paramètres-PacBio

Log-vraisemblance

$$L(x) = \sum_{i=1}^{n} log(k\mu(x_i\mu)^{k-1}e^{-(x_i\mu)^k}(e^{-\lambda x_i} - e^{-\lambda G}))$$

Expressions des paramètres

- $\mu = \sqrt[k]{\frac{n}{\sum_{i=0}^{n} x_i^k}}$
- $-\frac{n}{k} + \sum_{i=0}^{n} log(\mu x_i)e^{klog(\mu x_i)} = nlog(\mu) + \sum_{i=0}^{n} log(x_i)$
- $\sum_{i=0}^{n} \frac{-x_i e^{-\lambda x_i} + ge^{-\lambda g}}{e^{-\lambda x_i} - e^{-\lambda g}}$

  on peut réécrire : $\sum_{i=1}^{n}(g - x_i)\frac{e^{\lambda(g-x_i)}}{e^{\lambda(g-x_i)}-1} = nG$

Optimisation des paramètres-PacBio

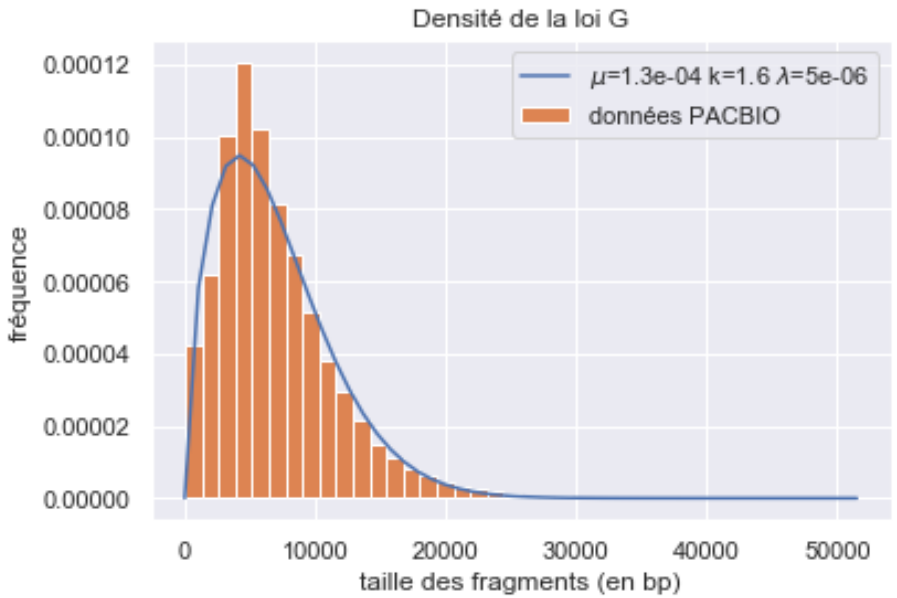


Figure – Représentation de la loi G et données PACBIO

Paramètres optimisés

$\mu = 1.3e-04$, k=1.6 , $\lambda = 5e-06$

Conclusion

MinION

$G(x) = p(1-p)^{x-1}(e^{-\lambda x} - e^{-\lambda G})$
avec $\lambda > 0$, $p \in [0,1]$, x : taille en bp et G : taille du génome

PacBio

$G(x) = k\mu(x\mu)^{k-1}e^{-(x\mu)^k}(e^{-\lambda x} - e^{-\lambda G})$
avec $\lambda > 0$, k>1, $\mu > 0$, x : taille en bp et G : taille du génome

Ouverture

Revisiter les travaux Lander-Watermann(1988) :
- propriétés statistiques du séquençage shotgun
- combien de lectures sont nécessaires pour recouvrir tous le génome
- la taille des lectures n'est plus fixe mais suit une loi de fonction de densité G(x)

Merci pour votre attention