

# Mélange de lois normales et tailles de génomes bactériens

J.R. Lobry

---

Étude de la taille de 279 génomes de bactéries exprimée en kilobases (données de 2002). Mise à jour avec des données plus récentes à partir de GOLD (Genomes Online Database).

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Les données de 2002</b>	<b>2</b>
<b>3</b>	<b>Estimation d'un mélange de lois normales</b>	<b>4</b>
<b>4</b>	<b>Interprétation biologique</b>	<b>6</b>
<b>5</b>	<b>Exercice</b>	<b>7</b>
	<b>Références</b>	<b>9</b>

## 1 Introduction

Cette fiche s'utilise en complément de tdr221 dont elle reprend toutes les notations.

On reprend en particulier la définition de la fonction `logvraineg()` qui retourne la valeur du logarithme de la fonction du maximum de vraisemblance dans le cas d'un mélange de deux lois normales (en fait l'opposé de cette valeur parce que l'on cherche à maximiser la vraisemblance et que la fonction d'optimisation utilisée ici, `nlm()`, cherche à minimiser la valeur de la fonction passée en argument).

```
logvraineg <- function(param, obs) {
  p <- param[1]
  m1 <- param[2]
  sd1 <- param[3]
  m2 <- param[4]
  sd2 <- param[5]
```

```
-sum(log(p*dnorm(obs,m1,sd1)+(1-p)*dnorm(obs,m2,sd2)))  
}
```

Ainsi que la fonction `simulmixnor()` pour simuler un mélange de deux lois normales :

```
simulmixnor <- function(n, p, m1, sd1, m2, sd2) {  
  n1 <- rbinom(1,n,p)  
  x1 <- rnorm(n1, m1, sd1)  
  x2 <- rnorm(n-n1, m2, sd2)  
  c(x1,x2)  
}
```

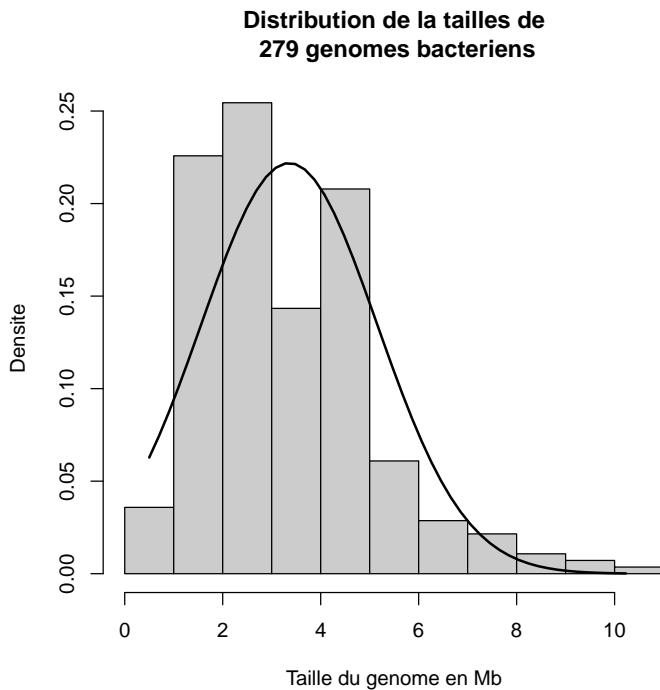
Pour ces deux fonctions, le paramètre `p` représente la fréquence relative de la première population dans le mélange des deux populations, `m1` et `m2` la moyenne pour la première et la deuxième population, respectivement, `sd1` et `sd2` l'écart-type pour la première et la deuxième population, respectivement.

## 2 Les données de 2002

```
data <- read.table(file="http://pbil.univ-lyon1.fr/R/donnees/bactgensize.txt", h=T, sep="\t", quote="")  
names(data)  
[1] "genus"   "species"  "strain"   "sizeKb"   "nORF"
```

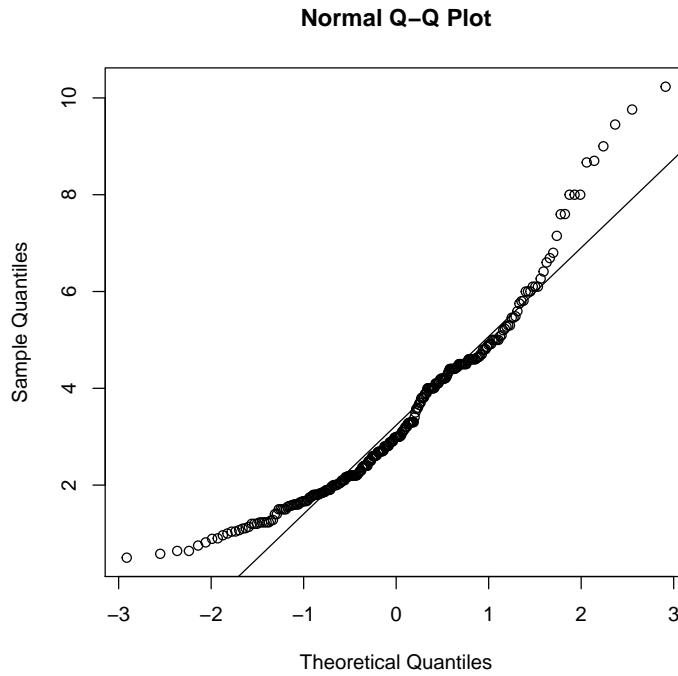
Les données du fichier `bactgensize.txt` ont été extraites de la "Genomes Online Database" (GOLD® <http://www.genomesonline.org/> [4, 2]) le 30 mai 2002. Cette base de données donne la liste des génomes bactériens entièrement séquencés ou en cours de séquençage. La colonne `sizeKb` du fichier donne la taille de 279 génomes de bactéries exprimée en kilobases. Bactérie est pris ici au sens large pour les domaines Archaeal et Bacterial de GOLD®.

```
x <- data$sizeKb/1000  
hist(x, col = grey(0.8), proba = TRUE,  
  main=paste("Distribution de la tailles de\n", length(x), "genomes bacteriens"),  
  xlab="Taille du genome en Mb",  
  ylab="Densite")  
xseq <- seq( from = min(x), to = max(x), length = 50)  
lines( xseq, dnorm(xseq, mean(x), sd(x)), lwd = 2 )
```



La distribution des tailles des génomes ne suit pas une loi normale, ce que l'on visualise bien avec un graphe quantiles-quantiles :

```
qqnorm(x)  
qqline(x)
```



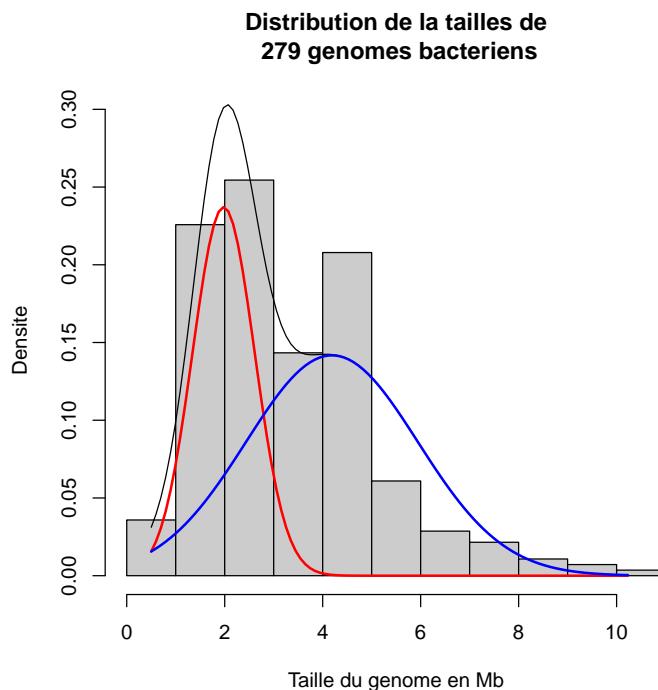
### 3 Estimation d'un mélange de lois normales

On peut suspecter une hétérogénéité dans les données et essayer d'ajuster un mélange de loi normales. L'optimisation de la fonction `logvraineg()` nous donne alors :

```
resnlm <- nlm(f = logvraineg, p = c(0.5, 2, 1, 4.5, 1), obs = x)
resnlm$estimate
[1] 0.3762302 1.9775784 0.6331008 4.1890334 1.7554831
```

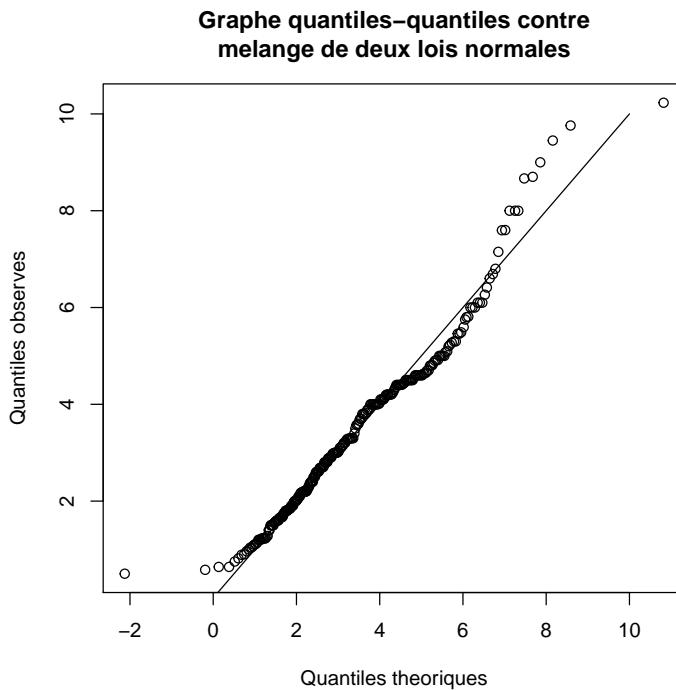
Pour ce jeu de données, la première population représente donc 37.6 % de la population totale, avec une taille de génome voisine de 2 Mb, alors que la taille moyenne des génomes de la seconde population fait plus du double. Graphiquement :

```
xseq <- seq( min(x), max(x), le = 100)
est <- resnlm$estimate
y1 <- est[1]*dnorm(xseq, est[2], est[3])
y2 <- (1-est[1])*dnorm(xseq, est[4], est[5])
hist(x, proba = TRUE, ylim = c(0,max(y1+y2)), col = grey(0.8),
      main=paste("Distribution de la tailles de\n", length(x), "genomes bacteriens"),
      xlab="Taille du genome en Mb",
      ylab="Densite")
lines(xseq, y1+y2)
lines(xseq, y1, col="red", lwd = 2)
lines(xseq, y2, col="blue", lwd = 2)
```



La description de la distribution de la taille des génomes bactériens semble plus satisfaisante avec un mélange de deux lois normales qu'avec une seule loi normale. Si on y regarde de plus près avec un graphe quantiles-quantiles :

```
theo <- simulmixnor(10000, est[1],est[2],est[3],est[4],est[5])
qqplot(theo, x,
       main="Graphe quantiles-quantiles contre\nmelange de deux lois normales",
       xlab="Quantiles theoriques",ylab="Quantiles observes")
lines(c(0,10), c(0,10))
```



On voit que cette description n'est pas bonne aux extrémités de la distribution. La distribution théorique prédit des tailles de génome négatives, ce qui est impossible. Elle ne prévoit pas assez de génomes entre 5000 et 7000 Kb et trop en deçà.

## 4 Interprétation biologique

A quoi correspondent ces deux groupes de génomes bactériens ? Un examen rapide des plus petits génomes :

```
data[order(x),c(1,2,4)][1:15, ]
  genus species sizeKb
7   Nanoarchaeum   equitans    500
239  Mycoplasma genitalium   580
32   Buchnera      aphidicola  640
212  Buchnera      aphidicola  640
223  Ureaplasma    urealyticum 751
237  Mycoplasma    pneumoniae  816
110  Mycoplasma    hypopneumoniae 890
66   Ehrlichia     sennetsu    900
200  Mycoplasma    pulmonis   963
41   Chlamydophila abortus    1000
40   Chlamydia     trachomatis 1038
229  Chlamydia     trachomatis 1042
220  Chlamydia     trachomatis 1069
265  Wolbachia     sp.        1100
228  Rickettsia    prowazekii 1111
```

montre qu'ils correspondent à des endosymbiontes ou à des parasites intracellulaires, alors que les plus grands génomes :

```
data[rev(order(x)),c(1,2,4)][1:15, ]
```

	genus	species	sizeKb
29	Bradyrhizobium	japonicum	10231
116	Nostoc	punctiforme	9760
112	Myxococcus	xanthus	9450
74	Gemmata	obscuriglobus	9000
253	Streptomyces	avermitilis	8700
179	Streptomyces	coelicolor	8667
257	Thermobifida	fusca	8000
252	Streptomyces	ambofaciens	8000
34	Burkholderia	fungorum	8000
33	Burkholderia	cepacia	7600
211	Mesorhizobium	loti	7596
123	Pirellula	sp.	7150
75	Geobacter	metallireducens	6800
197	Sinorhizobium	meliloti	6690
131	Pseudomonas	fluorescens	6600

correspondent à des bactéries capables d'une forme de vie autonome. C'est un phénomène connu, il y a une tendance à la réduction de la taille des génomes des bactéries intracellulaires, tendance qui poussée à l'extrême conduit aux petits génomes des mitochondries et des chloroplastes.

Pour en savoir plus on pourra consulter par exemple les articles des génomes complets de *Rickettsia prowazekii* [1]) et *Mycobacterium leprae* [3].

Plus généralement, il semblerait que les deux groupes opposent les bactéries *spécialistes* aux bactéries *généralistes*. On ne sait pas vraiment pourquoi la distribution est multimodale.

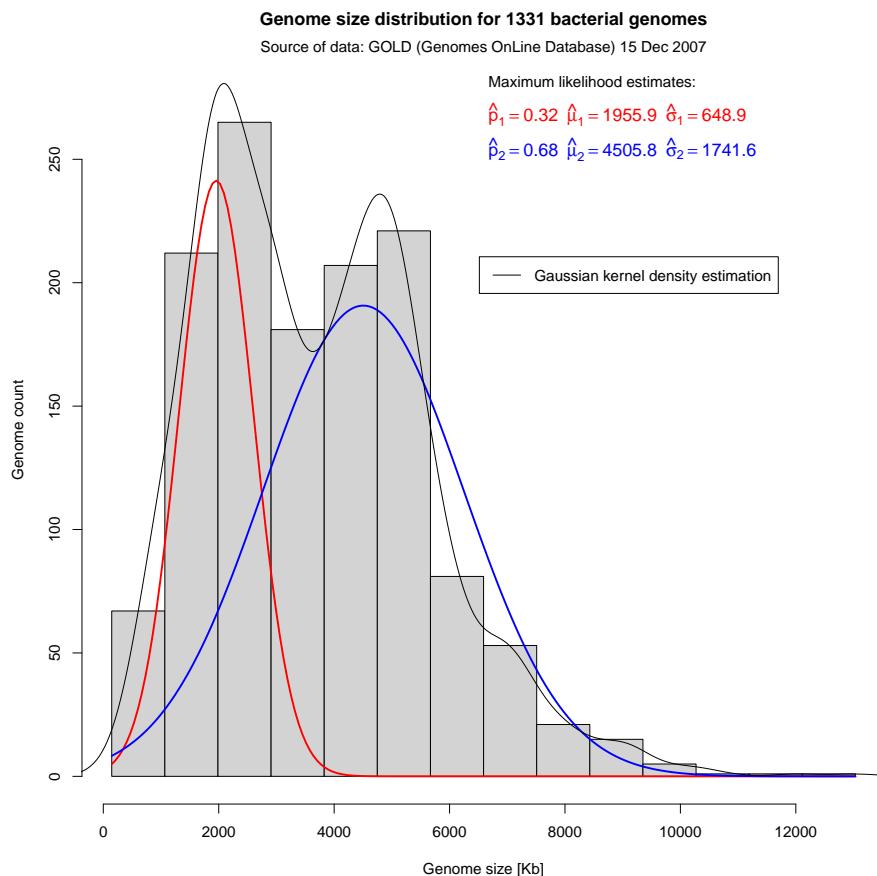
## 5 Exercice

Il y a beaucoup plus de données disponibles maintenant, vous devez essayer de refaire l'analyse sur des données mises à jour. Allez sur le site de GOLD® <http://www.genomesonline.org/> :



Cliquez sur le bouton "Download" pour récupérer le fichier `goldtable.txt` dans votre répertoire de travail puis importez les données dans R pour refaire l'analyse sur des données actualisées. Pour importer les données dans R vous pouvez utiliser la fonction `read.table()` en contrôlant les arguments `file`, `header`, `sep`, `comment.char` et `quote`.

Pour information, lors de la dernière compilation de ce document (17 février 2017) il y avait 1331 données disponibles, ressemblant à ceci :



Les plus petits génomes bactériens étaient :

		genus	species	gs
1226	Candidatus	Sulcia	Candidatus Sulcia muelleri	146
251	Candidatus	Carsonella	Candidatus Carsonella ruddii	159
245	Buchnera		Buchnera aphidicola	420
524	Nanoarchaeum		Nanoarchaeum equitans	490
1563	Mycoplasma		Mycoplasma genitalium	560
684	Mycoplasma		Mycoplasma genitalium	580
568	Buchnera		Buchnera aphidicola	615
651	Buchnera		Buchnera aphidicola	640
595	Buchnera		Buchnera aphidicola	641
926	Blattabacterium		Blattabacterium sp.	650
2052	Candidatus	Phytoplasma	Western X phytoplasma	670
1568	Mycoplasma		Mycoplasma orale	675
2100	unclassified	candidate division TM7	genomosp. GTL1	680
311	unclassified	Candidatus Baumannia cicadellinicola		686
579	Wigglesworthia		Wigglesworthia glossinidia	697

Les plus grands génomes bactériens étaient :

		genus	species	gs
9		Sorangium	Sorangium cellulosum	13033
1347	Microchaete		Fremyella diplosiphon	12000
1187	Plesiocystis		Plesiocystis pacifica	10585
1932	Stigmatella		Stigmatella aurantiaca	10265
931	Bradyrhizobium		Bradyrhizobium japonicum	10231
240	Solibacter		Solibacter usitatus	9965
1493	Microscilla		Microscilla marina	9771
1569	Myxococcus		Myxococcus xanthus	9450
321	Burkholderia		Burkholderia xenovorans	9279
1479	Magnetospirillum	Magnetospirillum magnetotacticum		9211
295	Myxococcus		Myxococcus xanthus	9139

570	Bradyrhizobium	Bradyrhizobium japonicum	9105
27	Frankia	Frankia sp. EAN1pec	9035
558	Streptomyces	Streptomyces avermitilis	9025
1591	Nostoc	Nostoc punctiforme	9020

## Références

- [1] S.G. Andersson, A. Zomorodipour, J.O. Andersson, T. Sicheritz-Ponten, U.C. Alsmark, R.M. Podowski, A.K. Naslund, A.S. Eriksson, H.H. Winkler, and C.G. Kurland. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, 396 :133–140, 1998.
- [2] A. Bernal, U. Ear, and N. Kyrpides. Genomes online database (GOLD) : a monitor of genome projects world-wide. *Nucleic Acids Res.*, 29 :126–127, 2001.
- [3] S.T. Cole, K. Eiglmeier, J. Parkhill, K.D. James, N.R. Thomson, P.R. Wheeler, N. Honore, T. Garnier, C. Churcher, D. Harris, K. Mungall, D. Basham, D. Brown, T. Chillingworth, R. Connor, R.M. Davies, K. Devlin, S. Duthoy, T. Feltwell, A. Fraser, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, C. Lacroix, J. Maclean, S. Moule, L. Murphy, K. Oliver, M.A. Quail, M.A. Rastandream, K.M. Rutherford, S. Rutter, K. Seeger, S. Simon, M. Simmonds, J. Skelton, R. Squares, S. Squares, K. Stevens, K. Taylor, S. Whithehead, J.R. Woodward, and B.G. Barrell. Massive gene decay in the leprosy bacillus. *Nature*, 409 :1007–1011, 2001.
- [4] N.C. Kyrpides. Genomes online database (GOLD 1.0) : a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, 15 :773–774, 1999.