

Les fréquences alléliques dans les populations humaines des microsatellites utilisés en sciences forensiques

Jean R. Lobry

22 août 2008

Cette fiche explore les données compilées par Brian Burritt du département de police à San Diego (CA, USA). On dispose par exemple des fréquences alléliques de 9 loci hautement polymorphes dans 202 populations humaines.

Table des matières

1	Introduction	2
2	Importation et pré-traitement des données	2
2.1	Le locus D3S1358	2
2.2	Le locus VWA	5
2.3	Le locus FGA	6
2.4	Le locus D8S1179	7
2.5	Le locus D21S11	7
2.6	Le locus D18S51	8
2.7	Le locus D5S818	9
2.8	Le locus D13S317	10
2.9	Le locus D7S820	10
2.10	Le locus D16S539	11
2.11	Le locus TH01	12
2.12	Le locus TPOX	12
2.13	Le locus CSF1PO	13
2.14	Le locus PentaD	14
2.15	Le locus PentaE	14
2.16	Le locus D2S1338	15
2.17	Le locus D19S433	16
2.18	Sauvegarde	16
3	Vues d'ensemble	16
3.1	Restauration des données pré-traitées	16
3.2	Nombre de populations par locus	16
3.3	Nombre d'allèles par locus	17
3.4	Fréquences alléliques médianes	17

1 Introduction

Les microsatellites représentent environ 3 % du génome humain. Ceux qui sont utilisés en sciences forensiques ont été passés en revue par John M. Butler [1]. Plus de 1000 articles publiés sont des études donnant une estimation des fréquences alléliques dans une population humaine, plus ou moins bien définie. Un effort considérable de compilation a été fait par Brian Burritt du département de police à San Diego (CA, USA). Cette compilation est disponible sous la forme d'un document tableur : `OmniPop200.1.xls`. Les données ci-après sont extraites de ce document.

2 Importation et pré-traitement des données

2.1 Le locus D3S1358

Motif tétranucléotide de type $(TCTR)_{8-21}$ sur le chromosome 3.

```
D3 <- read.table("D3S1358.csv", h = TRUE, sep = "\t", dec = ",",
  row.names = 1)
rownames(D3) <- abbreviate(rownames(D3))
dim(D3)
[1] 12 202
```

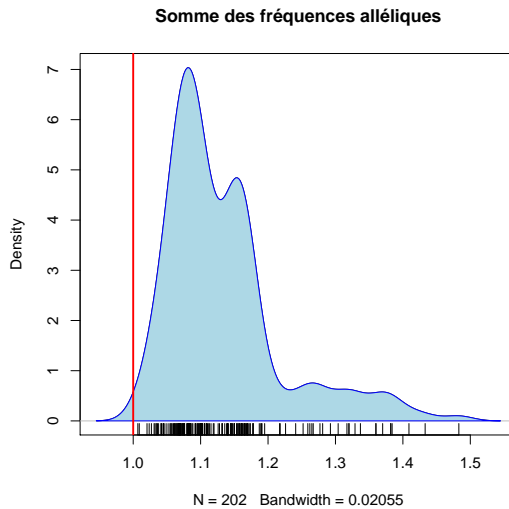
Nous avons donc 12 allèles différents dans 202 populations. Le nom des allèles est une variable qualitative ordonnée donnant le nombre de répétitions du microsatellite :

```
rownames(D3)
[1] "<12" "12,0" "13,0" "14,0" "15,0" "15,2" "16,0" "17,0" "17,1" "18,0" "19,0"
[12] ">19"
```

Les modalités extrêmes correspondent aux valeurs hors-échelle, le nombre de répétition n'est pas forcément entier en cas de délétion d'une ou plusieurs base dans un motif.

La première surprise avec ces données est que la somme des fréquences alléliques ne fasse pas 1 :

```
x <- colSums(D3)
dstx <- density(x)
plot(dstx, main = "Somme des fréquences alléliques")
polycurve <- function(x, y, base.y = min(y), ...) {
  polygon(x = c(min(x), x, max(x)), y = c(base.y, y, base.y),
  ...)
}
polycurve(dstx$x, dstx$y, col = "lightblue", border = "blue")
rug(x)
abline(v = 1, col = "red", lwd = 2)
```



La raison est que les fréquences nulles ont été ici remplacées par une valeur minimum égale à $\frac{5}{2n}$, où n est le nombre d'individus de la population. Par exemple pour la première population :

```
D3[, 1, drop = F]
  X.....FBI.African.American..1.
<12                0.01190
12,0                0.01190
13,0                0.01190
14,0                0.12143
15,0                0.29048
15,2                0.01190
16,0                0.30714
17,0                0.20000
17,1                0.01190
18,0                0.05476
19,0                0.01190
>19                0.01190

table(D3[, 1])
 0.0119 0.05476 0.12143 0.2 0.29048 0.30714
    7      1      1      1      1      1
```

On voit que la fréquence 0.0119 est présente 7 fois, c'est la fréquence minimum pour les allèles non encore observés. On doit donc pouvoir déduire le nombre d'individus :

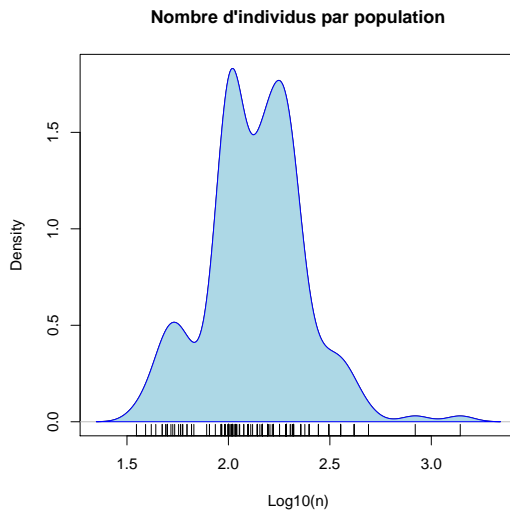
```
t2n <- function(x) {
  tx <- table(x)
  minf <- as.numeric(names(tx)[which.max(tx)])
  res <- 5/(2 * minf)
  if (!is.finite(res))
    res <- NA
  return(res)
}
t2n(D3[, 1])
[1] 210.0840

ntot <- apply(D3, 2, t2n)
sum(ntot)
[1] 32591.45

summary(ntot)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
35.21  100.00  140.90  161.30  192.30 1389.00
```

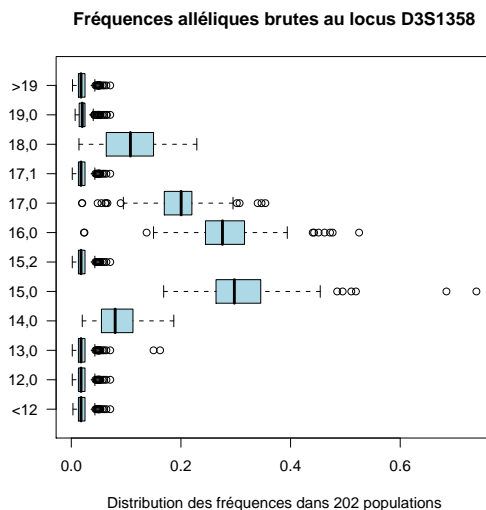
Nous avons donc des données sur plus de 30000 individus. Représentons la distribution du nombre d'individus par population :

```
dstn <- density(log10(ntot))
plot(dstn, main = "Nombre d'individus par population", xlab = "Log10(n)")
polycurve(dstn$x, dstn$y, col = "lightblue", border = "blue")
rug(log10(ntot))
```



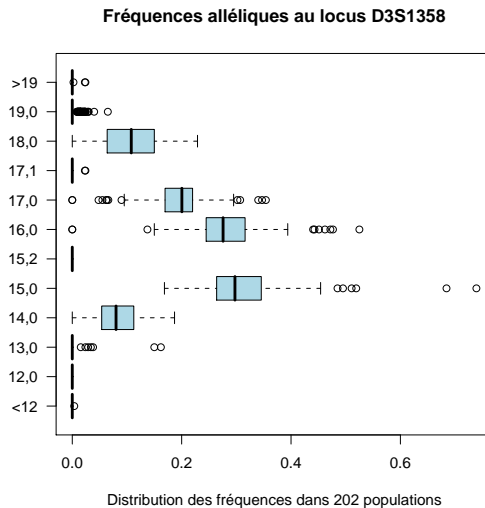
Il y a donc typiquement une centaine d'individus par population. Représentons maintenant la distribution des fréquences alléliques :

```
tmp <- as.data.frame(t(D3))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques brutes au locus D3S1358",
        xlab = "Distribution des fréquences dans 202 populations")
```



Ce graphique est trompeur à cause des fréquences plancher. Remplaçons les fréquences plancher par des fréquences nulles :

```
D3 <- apply(D3, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
0, x))
tmp <- as.data.frame(t(D3))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus D3S1358",
xlab = "Distribution des fréquences dans 202 populations")
burritt <- list()
burritt$D3S1358 <- tmp
```



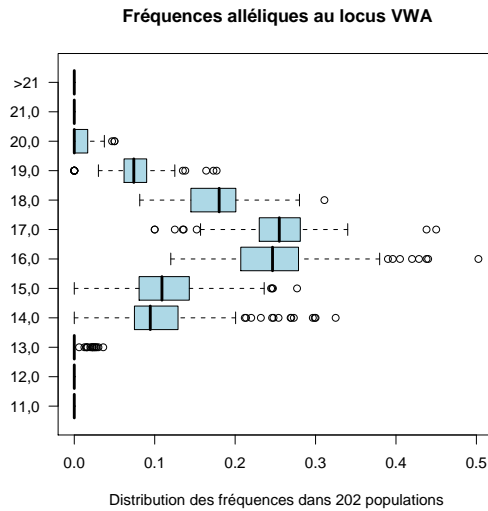
On a donc 5 allèles fréquents à ce locus : 14,0 15,0 16,0 17,0 18,0.

2.2 Le locus VWA

Motif tétranucléotide de type (TCTR)₁₀₋₂₅ sur le chromosome 12.

```
x <- read.table("VWA.csv", h = TRUE, sep = "\t", dec = ",", row.names = 1)
rownames(x) <- abbreviate(rownames(x))
dim(x)
[1] 12 202
rownames(x)
[1] "11,0" "12,0" "13,0" "14,0" "15,0" "16,0" "17,0" "18,0" "19,0" "20,0" "21,0"
[12] ">21"
ntot <- apply(x, 2, t2n)
sum(ntot)
[1] 32862.78
summary(ntot)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 35.21 100.00  140.90  162.70 192.30 1389.00

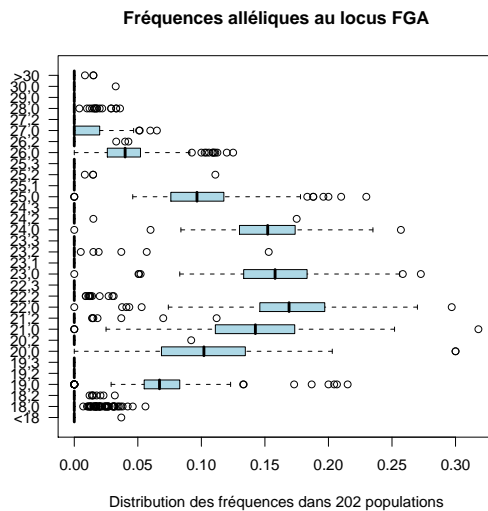
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus VWA",
xlab = paste("Distribution des fréquences dans", nrow(tmp),
"populations"))
burritt$VWA <- tmp
```



2.3 Le locus FGA

```
x <- read.table("FGA.csv", h = TRUE, sep = "\t", dec = ",", row.names = 1)
rownames(x) <- abbreviate(rownames(x))
dim(x)
[1] 32 202
```

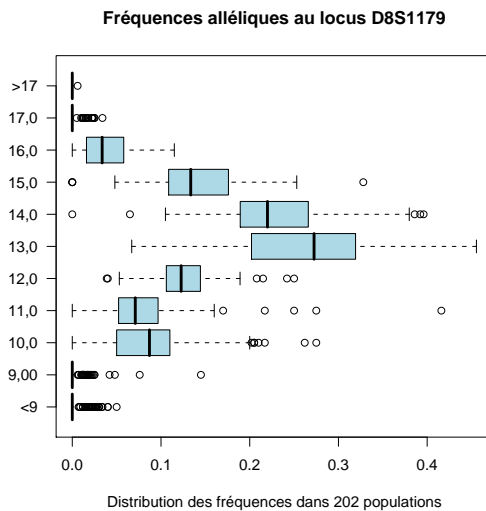
```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus FGA",
xlab = paste("Distribution des fréquences dans", nrow(tmp),
"populations"))
burritt$FGA <- tmp
```



2.4 Le locus D8S1179

```
x <- read.table("D8S1179.csv", h = TRUE, sep = "\t", dec = ",",
  row.names = 1)
rownames(x) <- abbreviate(rownames(x))
dim(x)
[1] 11 202
```

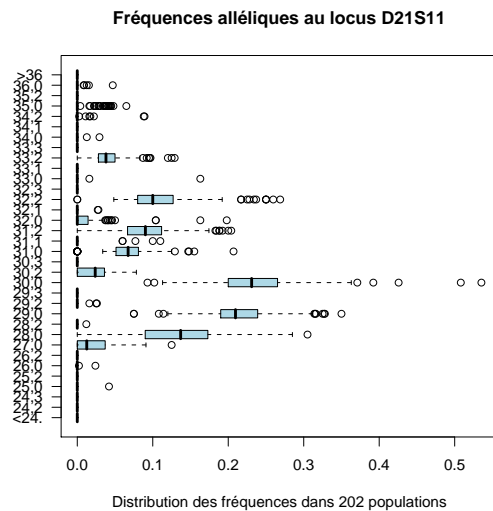
```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
  0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus D8S1179",
  xlab = paste("Distribution des fréquences dans", nrow(tmp),
  "populations"))
burritt$D8S1179 <- tmp
```



2.5 Le locus D21S11

```
x <- read.table("D21S11.csv", h = TRUE, sep = "\t", dec = ",", row.names = 1)
rownames(x) <- abbreviate(rownames(x))
dim(x)
[1] 34 202
```

```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
  0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus D21S11",
  xlab = paste("Distribution des fréquences dans", nrow(tmp),
  "populations"))
burritt$D21S11 <- tmp
```

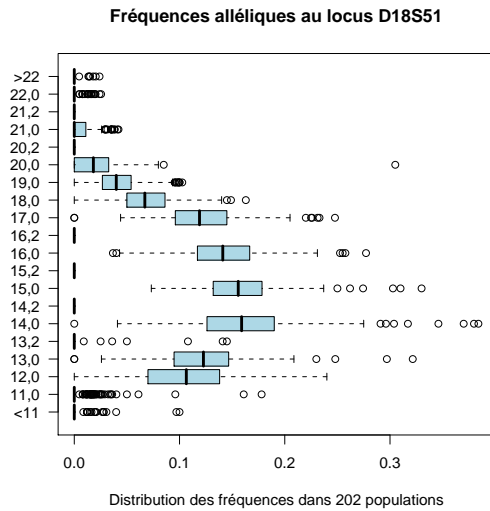


2.6 Le locus D18S51

Il y a une typo ici dans les données originelles.

```
x <- read.table("D18S51.csv", h = TRUE, sep = "\t", dec = ",", row.names = 1)
rownames(x) <- abbreviate(rownames(x))
dim(x)
[1] 20 202
max(x)
[1] 6
which(x == max(x), arr = TRUE)
      row col
14,2   7  12
x[7, 12] <- NA

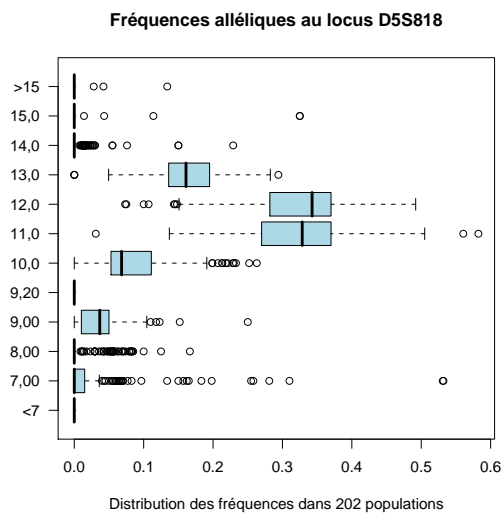
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus D18S51",
xlab = paste("Distribution des fréquences dans", nrow(tmp),
"populations"))
burritt$D18S51 <- tmp
```



2.7 Le locus D5S818

```
x <- read.table("D5S818.csv", h = TRUE, sep = "\t", dec = ",", row.names = 1)
rownames(x) <- abbreviate(rownames(x))
dim(x)
[1] 12 202
```

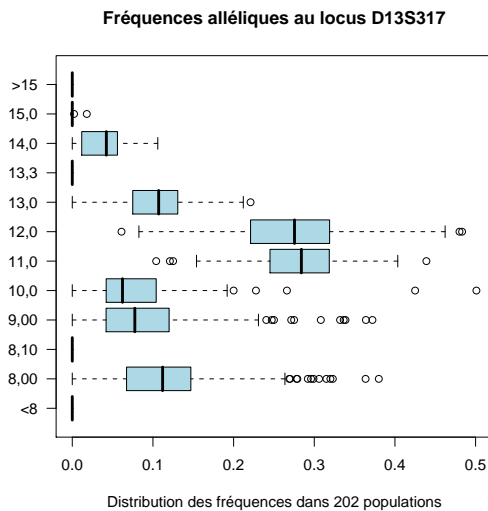
```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
  0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus D5S818",
  xlab = paste("Distribution des fréquences dans", nrow(tmp),
    "populations"))
burritt$D5S818 <- tmp
```



2.8 Le locus D13S317

```
x <- read.table("D13S317.csv", h = TRUE, sep = "\t", dec = ",",
  row.names = 1)
rownames(x) <- abbreviate(rownames(x))
dim(x)
[1] 12 202
```

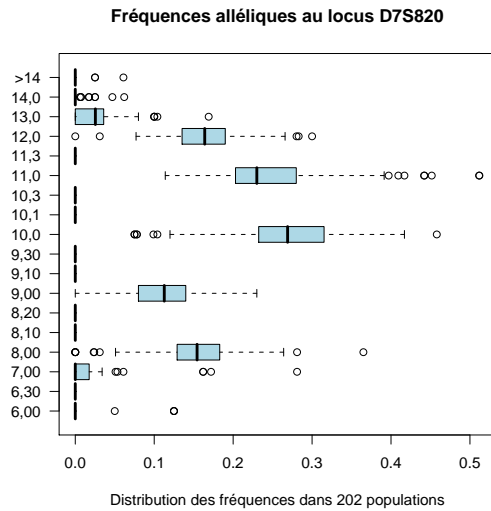
```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
  0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus D13S317",
  xlab = paste("Distribution des fréquences dans", nrow(tmp),
  "populations"))
burritt$D13S317 <- tmp
```



2.9 Le locus D7S820

```
x <- read.table("D7S820.csv", h = TRUE, sep = "\t", dec = ",", row.names = 1)
rownames(x) <- abbreviate(rownames(x))
dim(x)
[1] 18 202
```

```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
  0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus D7S820",
  xlab = paste("Distribution des fréquences dans", nrow(tmp),
  "populations"))
burritt$D7S820 <- tmp
```

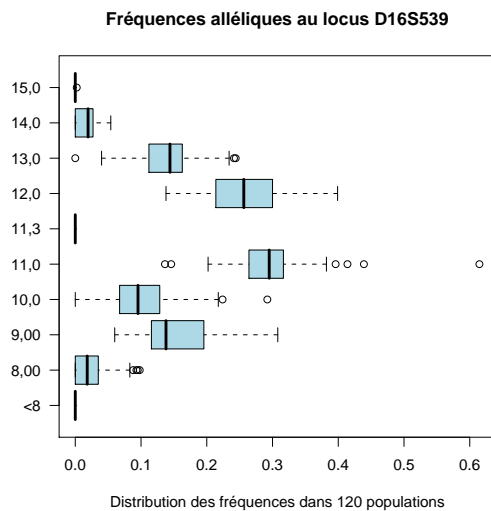


2.10 Le locus D16S539

```
x <- read.table("D16S539.csv", h = TRUE, sep = "\t", dec = ",",
  row.names = 1)
rownames(x) <- abbreviate(rownames(x))
x <- x[, !apply(x, 2, function(x) all(is.na(x)))]
dim(x)
```

```
[1] 10 120
```

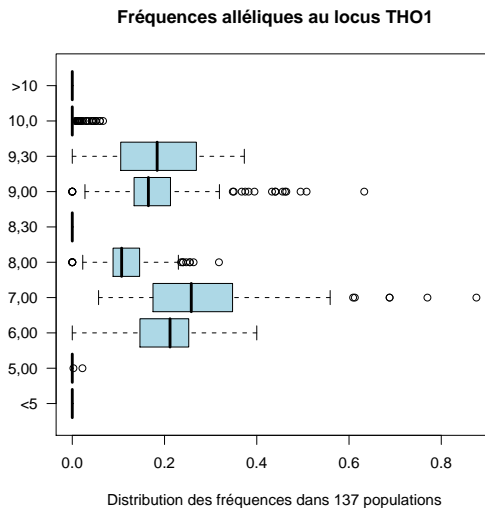
```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
  0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus D16S539",
  xlab = paste("Distribution des fréquences dans", nrow(tmp),
    "populations"))
burritt$D16S539 <- tmp
```



2.11 Le locus THO1

```
x <- read.table("THO1.csv", h = TRUE, sep = "\t", dec = ",", row.names = 1)
rownames(x) <- abbreviate(rownames(x))
x <- x[, !apply(x, 2, function(x) all(is.na(x)))]
dim(x)
[1] 10 137
```

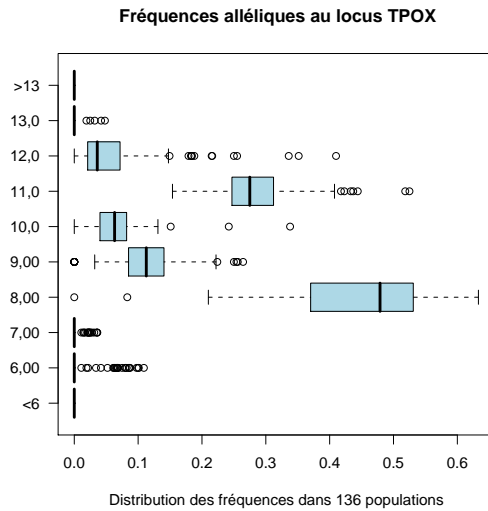
```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus THO1",
xlab = paste("Distribution des fréquences dans", nrow(tmp),
"populations"))
burritt$THO1 <- tmp
```



2.12 Le locus TPOX

```
x <- read.table("TPOX.csv", h = TRUE, sep = "\t", dec = ",", row.names = 1)
rownames(x) <- abbreviate(rownames(x))
x <- x[, !apply(x, 2, function(x) all(is.na(x)))]
dim(x)
[1] 10 136
```

```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus TPOX",
xlab = paste("Distribution des fréquences dans", nrow(tmp),
"populations"))
burritt$TPOX <- tmp
```

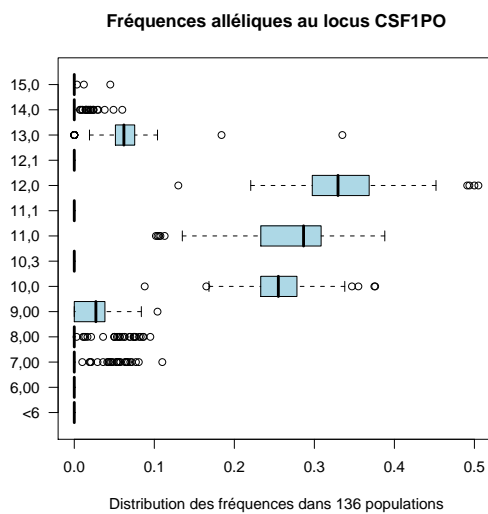


2.13 Le locus CSF1PO

```
x <- read.table("CSF1PO.csv", h = TRUE, sep = "\t", dec = ",", row.names = 1)
rownames(x) <- abbreviate(rownames(x))
x <- x[, !apply(x, 2, function(x) all(is.na(x)))]
dim(x)
```

```
[1] 14 136
```

```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus CSF1PO",
xlab = paste("Distribution des fréquences dans", nrow(tmp),
"populations"))
burritt$CSF1PO <- tmp
```

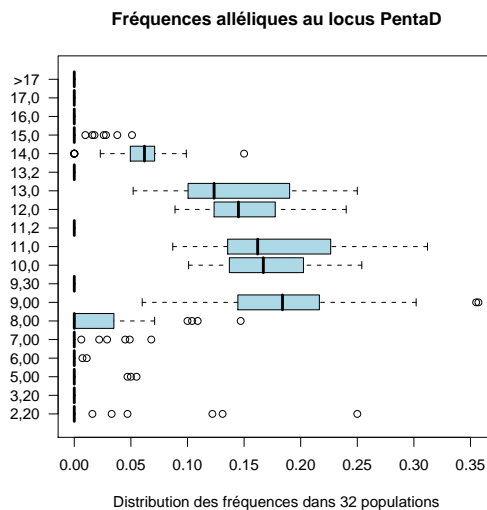


2.14 Le locus PentaD

```
x <- read.table("PentaD.csv", h = TRUE, sep = "\t", dec = ",", row.names = 1)
rownames(x) <- abbreviate(rownames(x))
x <- x[, !apply(x, 2, function(x) all(is.na(x)))]
dim(x)
```

[1] 19 32

```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus PentaD",
xlab = paste("Distribution des fréquences dans", nrow(tmp),
"populations"))
burrittt$PentaD <- tmp
```

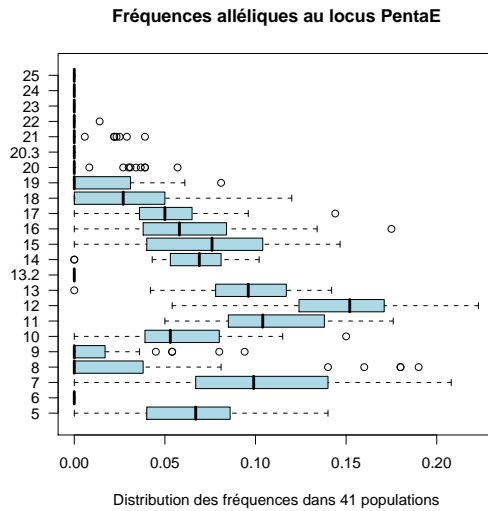


2.15 Le locus PentaE

```
x <- read.table("PentaE.csv", h = TRUE, sep = "\t", dec = ",", row.names = 1)
rownames(x) <- abbreviate(rownames(x))
x <- x[, !apply(x, 2, function(x) all(is.na(x)))]
dim(x)
```

[1] 23 41

```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus PentaE",
xlab = paste("Distribution des fréquences dans", nrow(tmp),
"populations"))
burrittt$PentaE <- tmp
```

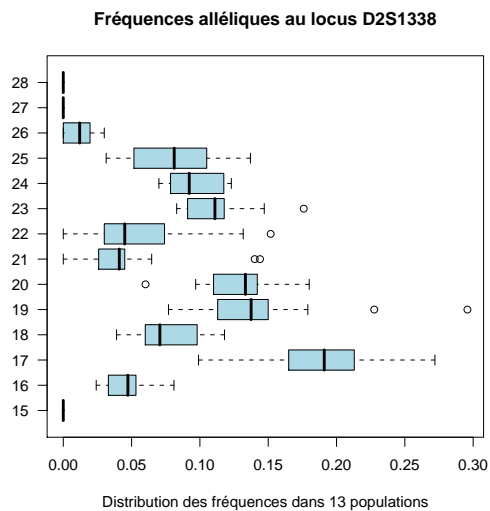


2.16 Le locus D2S1338

```
x <- read.table("D2S1338.csv", h = TRUE, sep = "\t", dec = ",",
  row.names = 1)
rownames(x) <- abbreviate(rownames(x))
x <- x[, !apply(x, 2, function(x) all(is.na(x)))]
dim(x)
```

```
[1] 14 13
```

```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
  0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus D2S1338",
  xlab = paste("Distribution des fréquences dans", nrow(tmp),
    "populations"))
burritt$D2S1338 <- tmp
```

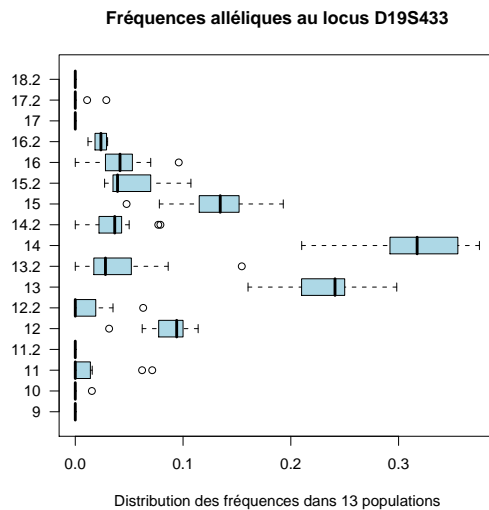


2.17 Le locus D19S433

```
x <- read.table("D19S433.csv", h = TRUE, sep = "\t", dec = ",",
  row.names = 1)
rownames(x) <- abbreviate(rownames(x))
x <- x[, !apply(x, 2, function(x) all(is.na(x)))]
dim(x)
```

[1] 17 13

```
x <- apply(x, 2, function(x) ifelse(x == as.numeric(names(which.max(table(x)))),
  0, x))
tmp <- as.data.frame(t(x))
boxplot(tmp, horizontal = TRUE, las = 1, col = "lightblue", main = "Fréquences alléliques au locus D19S433",
  xlab = paste("Distribution des fréquences dans", nrow(tmp),
    "populations"))
burritt$D19S433 <- tmp
```



2.18 Sauvegarde

```
save(burritt, file = "burritt.RData")
```

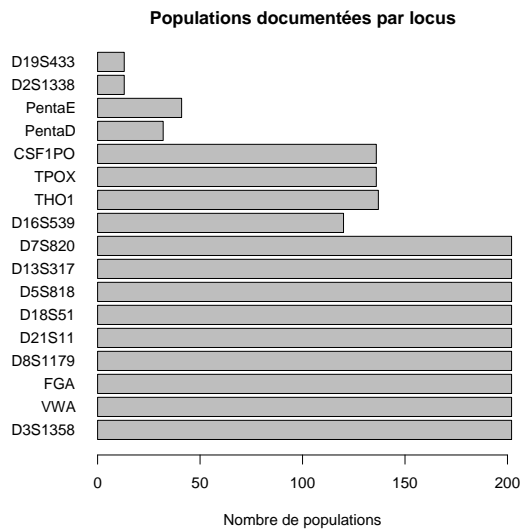
3 Vues d'ensemble

3.1 Restauration des données pré-traitées

```
load("burritt.RData")
```

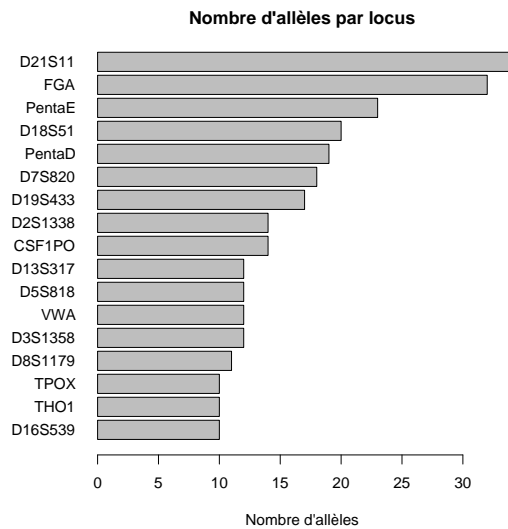
3.2 Nombre de populations par locus

```
par(mar = c(5, 5, 2, 2) + 0.1)
barplot(unlist(lapply(burritt, nrow))), horiz = T, las = 1, main = "Populations documentées par locus",
  xlab = "Nombre de populations")
```



3.3 Nombre d'allèles par locus

```
par(mar = c(5, 5, 2, 2) + 0.1)
apg <- unlist(lapply(burritt, ncol))
barplot(apg[order(apg)], horiz = T, las = 1, main = "Nombre d'allèles par locus",
        xlab = "Nombre d'allèles")
```

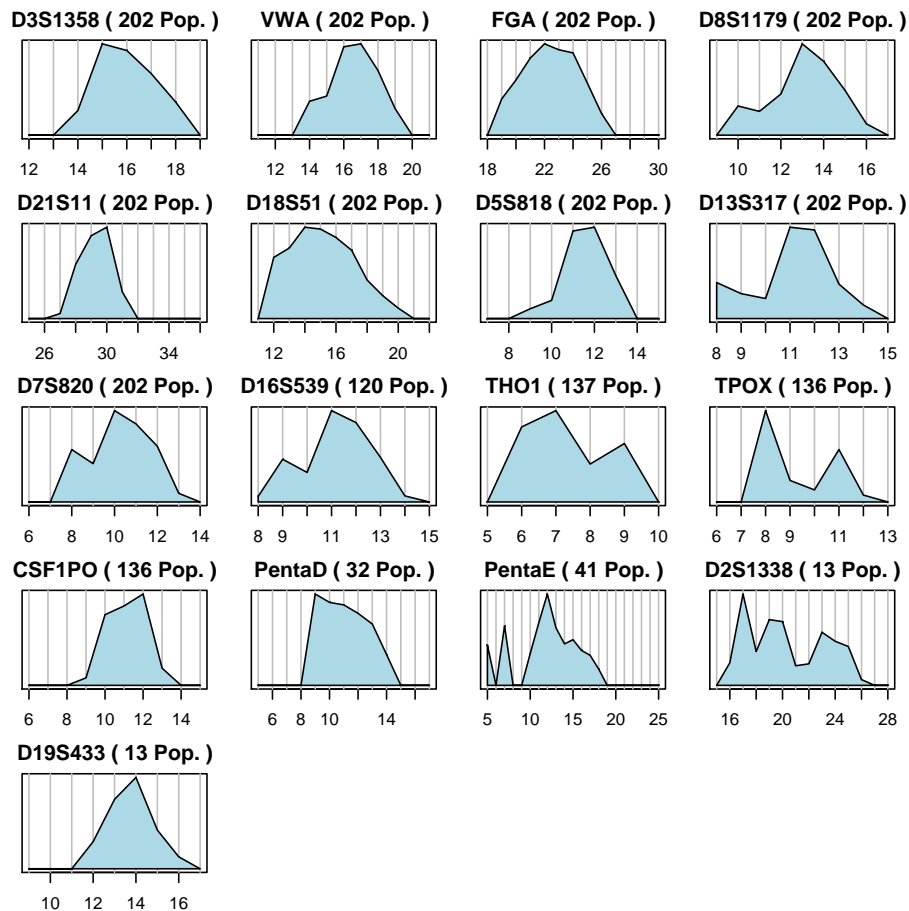


3.4 Fréquences alléliques médianes

Le graphique donne pour les allèles, correspondant à un multiple entier du motif de base et qui ne sont pas hors-échelle, la médiane des fréquences observées au sein les populations en fonction de la taille du microsatellite :

```

par(mfrow = c(5, 4), mar = c(2, 1, 2, 1) + 0.1)
for (i in 1:length(burritt)) {
  x <- burritt[[i]]
  suppressWarnings(xx <- as.numeric(sub(",", ".", colnames(x))))
  entiers <- floor(xx) == xx & !is.na(xx)
  xval <- xx[entiers]
  yval <- apply(x, 2, median)[entiers]
  plot(xval, yval, main = paste(names(burritt)[i], "(", nrow(x),
    "Pop. )"), yaxt = "n", pch = ".")
  for (i in min(xval):max(xval)) abline(v = i, col = "grey")
  polycurve(xx[entiers], yval, col = "lightblue")
}
    
```



Références

- [1] J.M. Butler. Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of Forensic Sciences*, 51 :253–265, 2006.