Exercices avec le logiciel R

Épreuve M1 A9

J.R. Lobry & A.B. Dufour

Contrôle - 21 novembre 2008

M1 - A9 - 21 novembre 2008 *Tous documents autorisés - échanges strictement interdits*

1 Répondre directement sur la feuille

Nom :
Prénom :
Numéro carte étudiant :

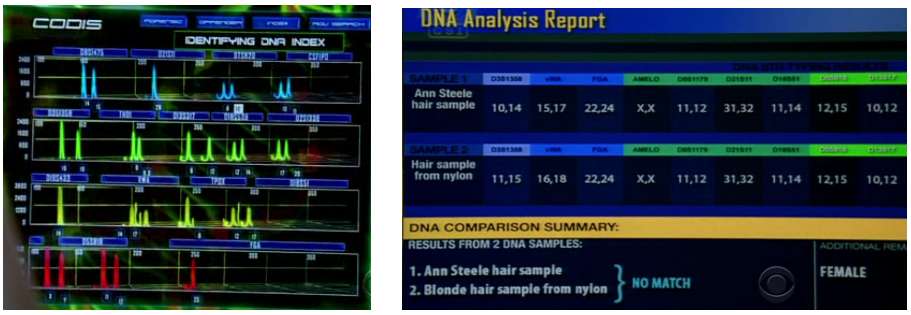


FIGURE 1 – Copies d'écran de séries de télévision présentant des profils génétiques utilisés par la police scientifique. À gauche, la version variable quantitative (extrait de la série CSI Miami). Les allèles sont caractérisés par leur taille déterminée par leur temps de migration dans une électrophorère capillaire. Les quatre couleurs correspondent aux quatre fluorochromes utilisés. Les 16 rectangles bleus donnent les amplitudes de taille attendues pour les 16 loci utilisés (`D8S1179`, `D21S11`, `D7S820`, `CSF1PO`, `D3S1358`, `TH01`, `D13S317`, `D16S539`, `D2S1338`, `D19S433`, `vWA`, `TPOX`, `D18S51`, `AMEL`, `D5S818`, `FGA` du kit AmpFℓSTR© Identifiler© [1]). À droite la version variable qualitative (extrait de la série CSI New-York). Les allèles sont codés par une variable qualitative. Il s'agit ici d'échantillon de cellules diploïdes, il y a donc deux allèles pour chaque locus. Les conventions de nomemclature pour les modalités de ces variables seront données dans le texte.

2 Introduction

L'utilisation de profils génétiques par la police scientifique a été popularisée par de nombreuses séries de télévision parfois remarquablement exactes dans les

détails (*cf* figure 1). Les données utilisées ici [3] sont les profils génétiques de 700 individus anonymes (uniquement des Etats-Unis d'Amérique) ayant auto-déclaré leur *ethnicity* ou *race* (*cf* figure 2).

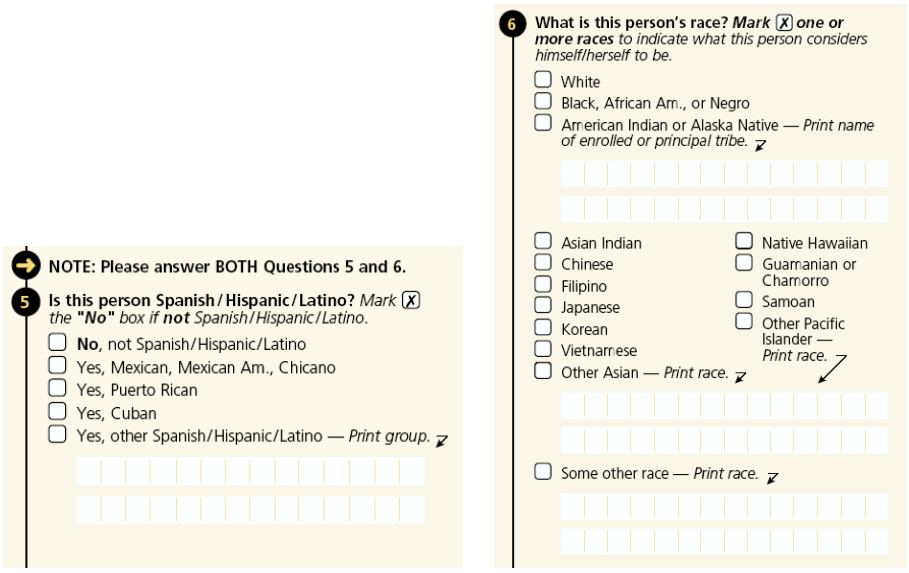


FIGURE 2 – Les Etats-Unis d'Amérique utilisent un concept assez flou d'*ethnicity* (question 5) et de *race* (question 6). Cette figure est une copie du formulaire utilisé pour le recencemment de l'année 2000. Notez la possibilité pour un individu de cocher plusieurs cases pour *race*. La collecte de ce type d'informations est interdite en France par la loi informatique et libertés du 6 janvier 1978 qui dispose en son article 8 qu'*[i]l est interdit de collecter ou de traiter des données à caractère personnel qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou qui sont relatives à la santé ou à la vie sexuelle de celles-ci.*

3 Importation des données

```
codis <- read.table("http://pbil.univ-lyon1.fr/R/donnees/codis.txt",
  header = TRUE, sep = "\t", row.names = 1, na.string = 0, colClasses = "factor")
dim(codis)
```

```
[1] 700  33
```

```
names(codis)
```

```
 [1] "D8S1179.1" "D8S1179.2" "D21S11.1"  "D21S11.2"  "D7S820.1"  "D7S820.2"
 [7] "CSF1PO.1"  "CSF1PO.2"  "D3S1358.1" "D3S1358.2" "TH01.1"    "TH01.2"
[13] "D13S317.1" "D13S317.2" "D16S539.1" "D16S539.2" "D2S1338.1" "D2S1338.2"
[19] "D19S433.1" "D19S433.2" "vWA.1"     "vWA.2"     "TPOX.1"    "TPOX.2"
[25] "D18S51.1"  "D18S51.2"  "AMEL.1"    "AMEL.2"    "D5S818.1"  "D5S818.2"
[31] "FGA.1"     "FGA.2"     "ethnicity"
```

À quoi sert la fonction `read.table()` et les arguments `sep` et `header` ?

Réponse :

Nous n'utilisons pas dans la suite le locus `AMEL` qui ne sert qu'à déterminer le sexe des individus. De plus, nous ne conservons que les individus entièrement documentés.

```
codis <- codis[,-grep("AMEL", colnames(codis))]
codis <- codis[complete.cases(codis),]
dim(codis)
```

```
[1] 699  31
```

```
codis[1:5, 1:5]
```

```
        D8S1179.1 D8S1179.2 D21S11.1 D21S11.2 D7S820.1
GA05071        13        14     30.2     32.2        7
GT37306        13        14       31     31.2       10
GT37312        13        14       29       30       10
GT37349        13        14       30       31       10
GT37351        10        12       28       29       10
```

Combien d'individus n'étaient pas entièrement documentés ?

Réponse :

```
table(codis$ethnicity)
```

```
Afric  Cauc  Hisp
  257   302   140
```

La colonne `ethnicity` est une variable qualitative non ordonnée dont la signification [3] est la suivante :

- `Afric` *African American*
- `Cauc` *U.S. Caucasian*
- `Hisp` *Hispanic*

Cet échantillon est-il représentatif de la variabilité existant (*cf* figure 2) aux Etats-Unis d'Amérique ?

Réponse :

3.1 Nomenclature des allèles

3.1.1 Cas général

Les loci utilisés en sciences forensiques sont des microsatellites, ou encore STR pour *Short Tandem Repeats*, c'est-à-dire des répétitions de petite taille (de 2 à 5 paires de bases) en tandem. Les loci utilisés correspondent principalement à

des microsatellites ayant un motif élémentaire de 4 paires de bases. Ainsi, dans ce jeu de données, tous les microsatellites sont formés de la répétition directe de tétranucléotides. Par exemple, le locus `D8S1179` correspond à la séquence $(TATC)_n$, où n est le nombre de répétitions du motif. C'est ce nombre n qui est très variable d'un individu à l'autre. Dans notre jeu de données pour le locus `D8S1179` :

```
unique(c(levels(codis$D8S1179.1),levels(codis$D8S1179.2)))
```

```
 [1] "10" "11" "12" "13" "14" "15" "16" "8"  "9"  "17" "18"
```

le nombre de répétitions varie donc de 8 à 18. La convention [2] est de nommer les allèles par le nombre de répétitions du motif élémentaire. On utilise une échelle allélique pour déterminer les allèles (*cf* figure 3).

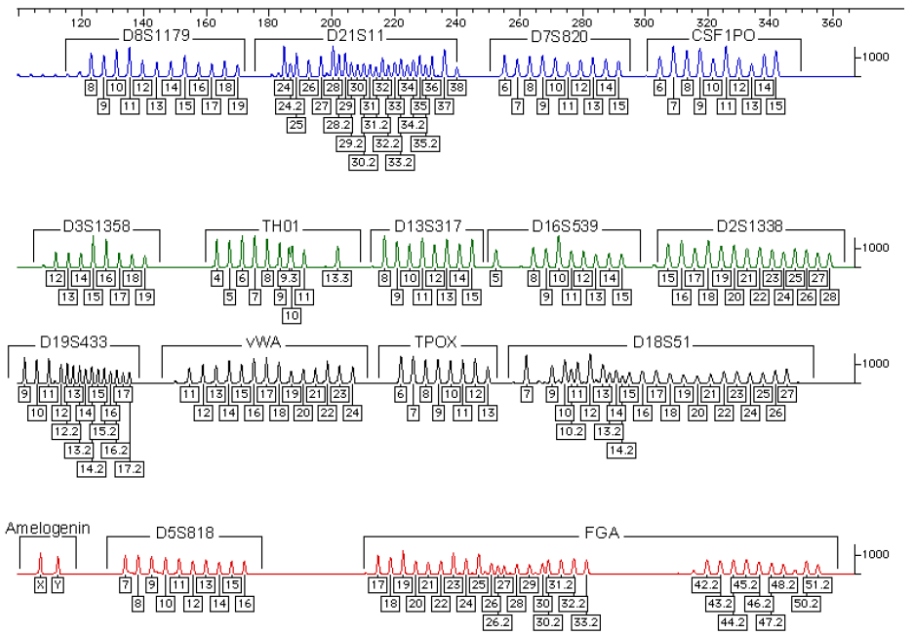


FIGURE 3 – Un exemple d'échelle allélique (copie d'écran de la figure 5-1 de [1]). C'est un mélange des allèles les plus fréquents dans les populations humaines qui va permettre de transformer la variable quantitative *position d'un pic* observé en une variable qualitative ordonnée *nom de l'allèle* correspondant.

3.1.2 Cas particulier des microvariants

Il arrive qu'un des motifs du microsatellite ne soit pas complet. Dans ce cas il est désigné par le nombre de motifs complets suivi du nombre de paires de bases dans le motif incomplet. Ces deux valeurs sont séparées par un point [2]. Par exemple, il existe pour le locus `TH01` un microvariant plus court d'une paire de base que la répétition de 10 motifs élémentaires, on utilise donc 9.3 pour le désigner :

```
unique(c(levels(codis$TH01.1),levels(codis$TH01.2)))
```

```
[1] "5"   "6"   "7"   "8"   "9"   "9.3" "10"  "11"
```

Sa taille est donc de $9 \times 4 + 3 = 39$ paires de bases. La figure 4 vous donne la distribution des tailles observées dans ce jeu de données en mettant en évidence les microvariants. Elle a été produite avec le code suivant :

D'après la figure 4, quels sont les trois principaux loci pour lesquels on trouve des microvariants ?

Réponse :

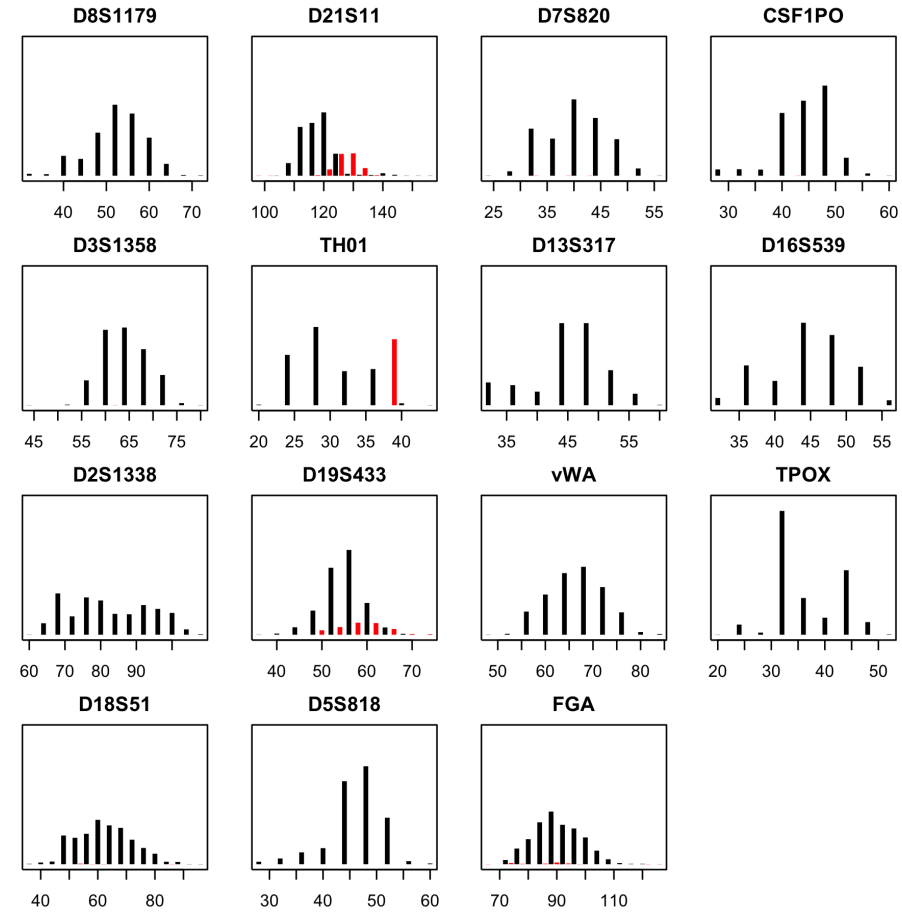


FIGURE 4 – Il y a 15 graphiques pour les 15 loci étudiés (leur nom figure en titre). Ces figures donnent les fréquences absolues ($\times 100$) observées dans le jeu de données en fonction de la taille (en paires de bases) des microsatellites. Les microvariants (dont la taille n'est pas un multiples de 4) sont indiqués en rouge.

4 Point de vue de quantitatif

On décide d'analyser les données du point de vue de la longueur des allèles, c'est-à-dire les données originelles avant leur conversion en nom d'allèle. La fonction `al2bp()` suivante permet de calculer la longueur d'un allèle en paires de bases à partir de son nom :

```
al2bp <- function(al){
        dec <- unlist(strsplit(as.character(al), split = "\\."))
        res <- 4*as.numeric(dec[1])
        if(length(dec) > 1) res <- res + as.numeric(dec[2])
  return(res)
}
al2bp(9)
```

```
[1] 36
```

```
al2bp(9.3)
```

```
[1] 39
```

On construit l'objet `codis.q` qui contient la longueur des allèles :

```
codis.q <- apply(codis[,1:30], c(1,2), al2bp)
dim(codis.q)
```

```
[1] 699  30
```

```
codis.q[1:5,1:5]
```

```
        D8S1179.1 D8S1179.2 D21S11.1 D21S11.2 D7S820.1
GA05071        52        56      122      130       28
GT37306        52        56      124      126       40
GT37312        52        56      116      120       40
GT37349        52        56      120      124       40
GT37351        40        48      112      116       40
```

```
par(mfrow=c(4,4), mar = c(2,2,2,0)+0.1, lend = "butt")
for(i in seq(1,30,by=2)){
  cnt <- table(c(codis.q[,i], codis.q[,i+1]))
        x <- as.numeric(names(cnt))
        y <- cnt/100
        plot(x, y, type = "h", col = ifelse(x %% 4 == 0, "black","red"),
        lwd = 3, las = 1, ylim = c(0,7),
        main = unlist(strsplit(colnames(codis)[i], split = "\\."))[1])
}
```

Le code suivant a été utilisé pour produire la figure 5 :

```
par(mfrow=c(4,4), mar = c(2,2,2,0)+0.1, lend = "butt")
for(i in seq(1,30,by=2)){
        x <- c(codis.q[,i], codis.q[,i+1])
        grp <- c(codis$eth, codis$eth)
        boxplot(x~grp, varwidth = TRUE, horizontal = TRUE,
        col = c("red3","green3",grey(0.3)), las = 1,
        main = unlist(strsplit(colnames(codis.q)[i], split = "\\."))[1])
}
plot.new()
legend("center", legend = c("1. Afric", "2. Cauc", "3. Hisp"), lty = 1,
lwd = 5,
col = c("red3","green3",grey(0.3)), bty = "n")
```

Au vu de la figure 5, que peut-on dire sur la différence de taille des allèles entre les trois groupes ?

Réponse :

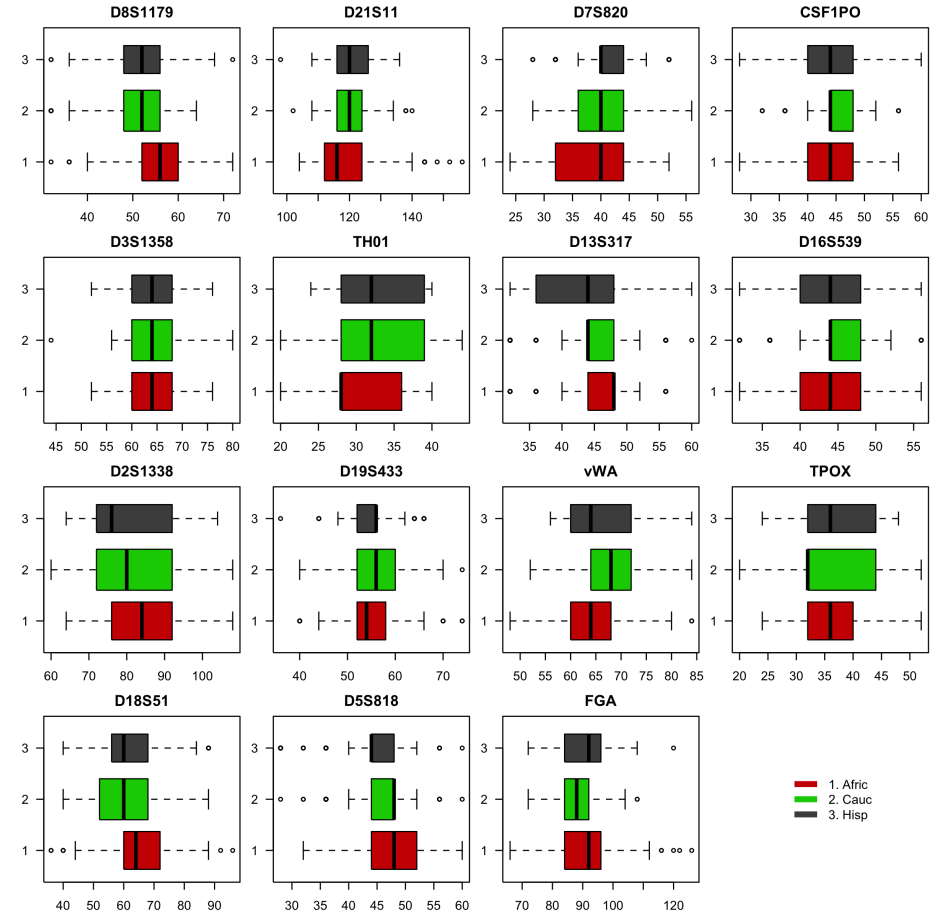


FIGURE 5 – Il y a 15 graphiques pour les 15 loci étudiés (leur nom figure en titre). Chaque graphique donne la distribution des tailles des allèles observés pour les trois groupes d'individus.

Les individus étudiés sont diploïdes, ils ont donc tous deux allèles (éventuellement le même en cas d'homozygotie) pour chaque locus. Par exemple le premier individu a un allèle de taille 52 et un allèle de taille 56 au locus `D8S1179`.

```
codis.q[1,1:2]
```

```
D8S1179.1 D8S1179.2
       52        56
```

On s'intéresse à la relation entre la taille des deux allèles. Le code suivant donne la figure 6.

```
par(mfrow = c(4,4), mar = c(1,1,1,1)+0.1)
for(i in seq(1,30,by=2)){
        x <- codis.q[,i]
        y <- codis.q[,i+1]
         sunflowerplot(x,y, seg.lwd = 0.5,
         size = 1/20, asp = 1)
         abline(c(0,1), col = "blue")
         legend("bottomright", unlist(strsplit(colnames(codis.q)[i], split = "\\."))[1], bty = "n")
}
```

Au vu de la figure 6, que peut-on dire sur la relation entre la longueur des deux allèles pour chaque locus ?

Réponse :

Cette structure dans les données est complètement artificielle : chaque individu possède un allèle issu de son géniteur et un allèle issu de sa génitrice, mais on ne peut pas dire lequel est lequel. L'ordre utilisé ici est purement conventionnel. On se demande si une telle structure dans les données est susceptible de conduire à des artefacts. Pour ce faire, on fait l'expérience suivante :

```
x <- rnorm(699)
head(x)
```

```
[1]  0.1399803  0.2697330  0.9514394  0.3271120  0.6562713 -0.2095637
```

```
y <- rnorm(699)
head(y)
```

```
[1]  0.06964759  0.82877209 -0.39780477  0.40314377 -0.59062451  1.97335442
```

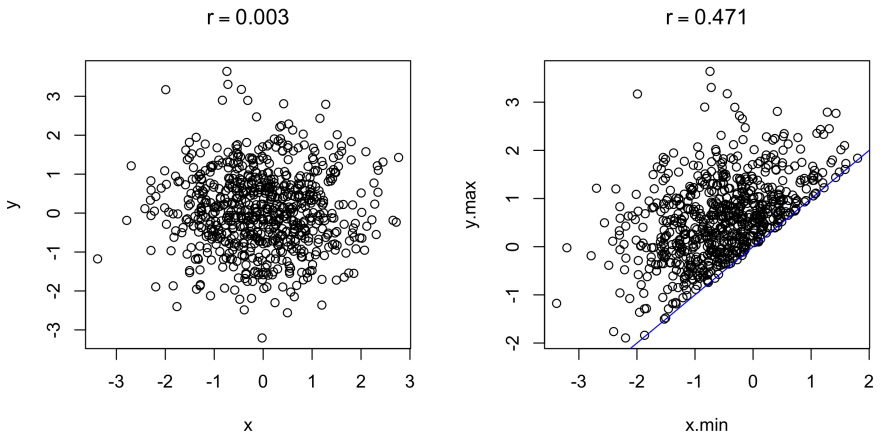
```
x.min <- pmin(x,y)
head(x.min)
```

```
[1]  0.06964759  0.26973302 -0.39780477  0.32711201 -0.59062451 -0.20956373
```

```
y.max <- pmax(x,y)
head(y.max)
```

```
[1] 0.1399803 0.8287721 0.9514394 0.4031438 0.6562713 1.9733544
```

```
par(mfrow = c(1,2))
plot(x,y, main = bquote(r == .(round(cor(x,y),3))))
plot(x.min, y.max, main = bquote(r == .(round(cor(x.min,y.max),3))))
abline(c(0,1), col = "blue")
```

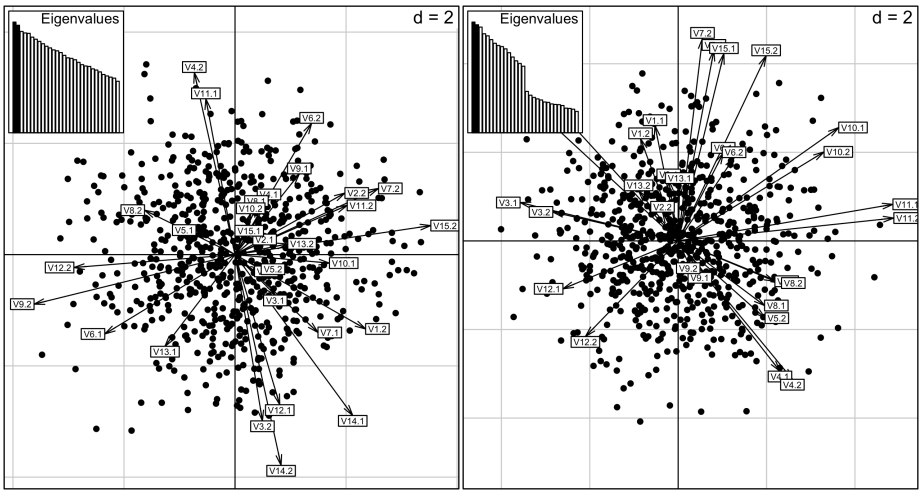


Au vu de ces résultats, pensez-vous que la structure artificielle dans les données soit susceptible de conduire à des artefacts ?

Réponse :

Toujours pour étudier l'effet de la structure artificielle des données, on simule un jeu de données de même dimension que le jeu étudié que l'on résume par une ACP :

```
set.seed(1)
rndtab <- as.data.frame(matrix(rnorm(30*699), ncol = 30))
colnames(rndtab) <- paste(rep(paste("V",1:15, sep = ""),each=2),1:2, sep = ".")
ordtab <- rndtab
for(i in seq(1,30,by=2)){
  ordtab[,i] <- pmin(rndtab[,i], rndtab[,i+1])
    ordtab[,i+1] <- pmax(rndtab[,i], rndtab[,i+1])
}
library(ade4)
rndtab.acp <- dudi.pca(rndtab, scann=FALSE)
ordtab.acp <- dudi.pca(ordtab, scann=FALSE)
par(mfrow = c(1,2))
scatter(rndtab.acp, clab.row = 0, clab.col = 0.5)
scatter(ordtab.acp, clab.row = 0, clab.col = 0.5)
```

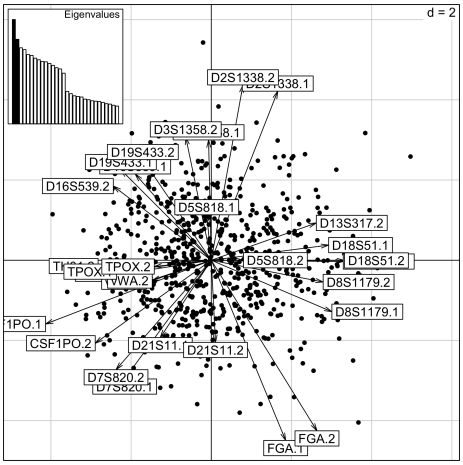


Au vu de ces résultats, quel est l'effet attendu de la structure artificielle des données sur le résultat de l'ACP ?

Réponse :

On effectue maintenant l'ACP du jeu de données pour voir si l'artefact est présent.

```
acp <- dudi.pca(codis.q, scann = FALSE)
scatter(acp, clab.row = 0)
```

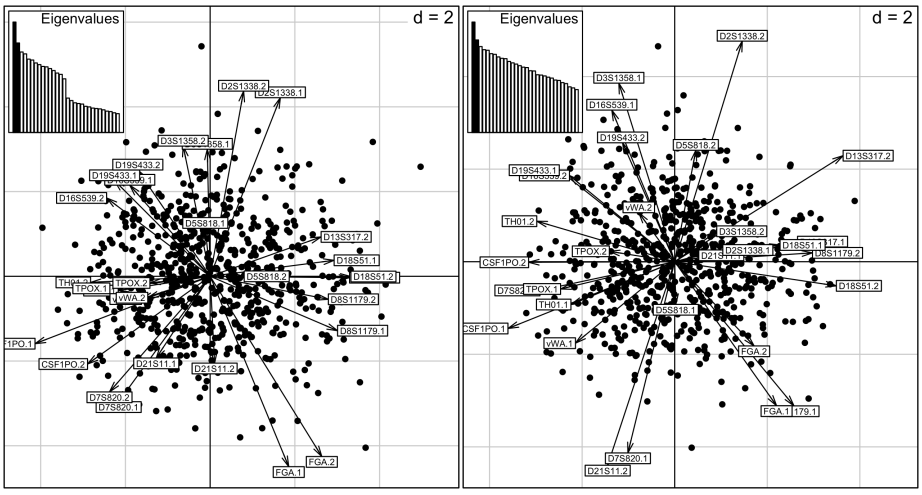


Au vu de ces résultats, l'artefact attendu est-il présent ?

Réponse :

Pour pallier cet inconvénient, on décide alors de faire la chose suivante :

```
codis.q2 <- codis.q
for(j in seq(1,30,by=2)){
  for(i in 1:699)
    codis.q2[i, c(j,j+1)] <- sample(codis.q[i, c(j,j+1)])
}
acpq2 <- dudi.pca(codis.q2, scann=FALSE)
par(mfrow = c(1,2))
scatter(acp, clab.row = 0, clab.col = 0.5)
scatter(acpq2, clab.row = 0, clab.col = 0.5)
```



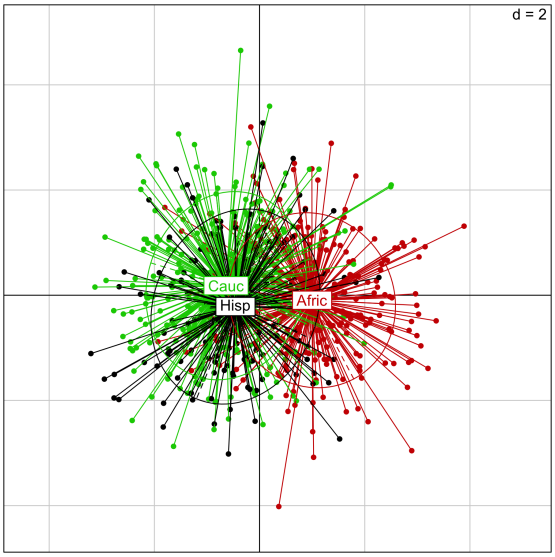
Expliquer ce que l'on a fait ici et pourquoi.

Réponse :

4.1 Interprétation du plan factoriel

Pour aider l'interprétation du premier plan factoriel, on utilise les groupes comme variables illustratives :

```
s.class(acpq2$li, codis$ethnicity, col = c("red3","green3","black"))
```

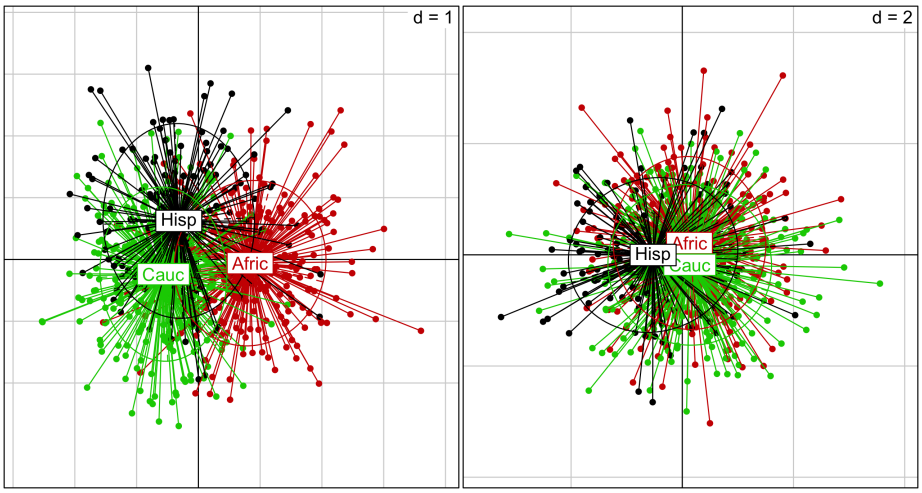


Au vu de ce résultat, quelle est votre interprétation des deux premiers facteurs de l'ACP ?

Réponse :

Les loci étudiés ici sont ceux qui sont utilisés dans les bases de données ADN telles que le CODIS (*Combined DNA Index System*) géré par le FBI (*Federal Bureau of Investigation*) aux Etats-Unis d'Amérique ou le FNAEG (Fichier National Automatisé des Empreintes Génétiques) géré par la police nationale et la gendarmerie nationale en France. La question de la prédictibilité de l'appartenance d'un individu à un groupe ethnique à partir de la connaissance de ses allèles est une question sociétale sensible. On fait l'expérience suivante pour étudier le pouvoir prédictif de ce type de données :

```
ad2 <- discrimin(acpq2, codis$ethnicity, scann = FALSE)
codis.q3 <- codis.q2
for(j in 1:ncol(codis.q2))
  codis.q3[,j] <- sample(codis.q2[,j])
acpq3 <- dudi.pca(codis.q3, scann = FALSE)
ad3 <- discrimin(acpq3, codis$ethnicity, scann = FALSE)
par(mfrow = c(1,2))
s.class(ad2$li, codis$ethnicity, col = c("red3","green3","black"))
s.class(ad3$li, codis$ethnicity, col = c("red3","green3","black"))
```



Au vu de ces résultats, que pouvez-vous dire du pouvoir prédictif des loci utilisés en sciences forensiques ? Est-il beaucoup plus important que celui obtenu avec des données aléatoires ?

Réponse :

Références

[1] Anonymous. *AmpFℓSTR© Identifiler© PCR Amplification Kit. User's Manual.* Applied Biosystems, Foster City, CA, USA, 2006. PN 4323291D.

[2] W. Bar, B. Brinkmann, P. Lincoln, W.R. Mayr, and U. Rossi. DNA recommendations. 1994 report concerning further recommendations of the DNA commission of the ISFH regarding PCR-based polymorphisms in STR (short tandem repeat) systems. *Int. J. Leg. Med.*, 107 :159–160, 1994.

[3] J.M. Butler, R. Schoske, P.M. Vallone, J.W. Redman, and M.C. Kline. Allele frequencies for 15 autosomal STR loci on U.S. caucasian, african american, and hispanic populations. *Journal of Forensic Sciences*, 48 :908–911, 2003.

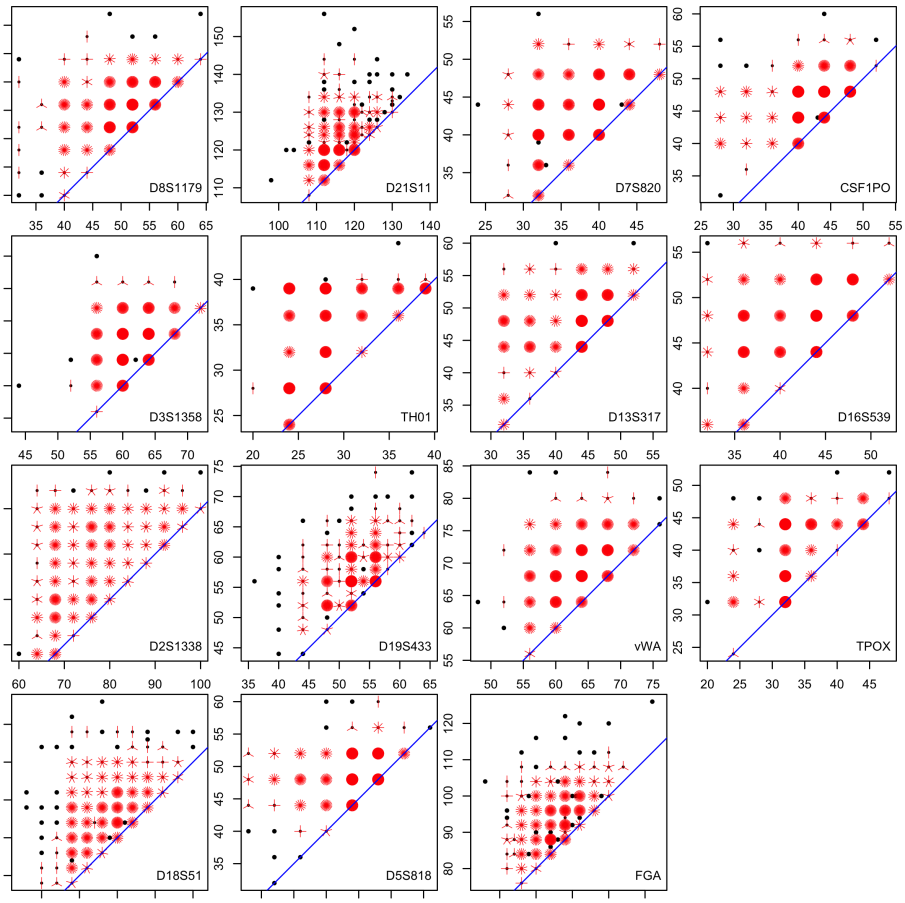


FIGURE 6 – Il y a 15 graphiques pour les 15 loci étudiés (leur nom figure en bas à droite). Les 699 individus sont positionnés à la taille de leur premier allèle en abscisse et à celle de leur deuxième en ordonnée. La superposition des points est gérée ici par une représentation graphique de type tournesol. La ligne en bleu correspond à la première bissectrice ($y = x$).