

Exercices avec le logiciel 

Épreuve M1 A9

J.R. Lobry

Contrôle - 4 décembre 2009

M1 - A9 - 4 décembre 2009 *Tous documents autorisés - échanges strictement interdits*

1 Répondre directement sur la feuille

Nom : Prénom : Numéro carte étudiant :
--

2 Introduction

Les données utilisées ici sont extraites du travail de deuxième année de master d'Alexandra Debernardi [4, 5].

L'identification humaine à partir de traces d'ADN est basée sur l'analyse de profils génétiques pour des loci de type microsatellite à base de tétranucléotides. Après une phase d'amplification PCR, les profils sont établis en séparant les fragments amplifiés en fonction de leur taille par électrophorèse capillaire. Pour vérifier qu'il n'y a pas eu de problème lors de la phase d'amplification on utilise un contrôle positif qui dans notre cas est une solution ADN de qualité industrielle dont le profil génétique attendu est donné dans la table 1. Au vu de cette table, donner le nombre de loci de type homozygote.

Réponse :

Donner le nombre de loci lus dans le vert.

Réponse :

Un exemple de résultat obtenu pour un contrôle positif est donné dans la figure 1.

Abbréviation	Locus	Canal	Chromophore	λ [nm]	Genotype
D8	D8S1179	bleu	6-FAM	522	13
D21	D21S11	.	.	.	30
D7	D7S820	.	.	.	10, 11
CSF	CSF1PO	.	.	.	10, 12
D3	D3S1358	vert	VIC	554	14, 15
TH	TH01	.	.	.	8, 9.3
D13	D13S317	.	.	.	11
D16	D16S539	.	.	.	11, 12
D2	D2S1338	.	.	.	19, 23
D19	D19S433	jaune	NED	575	14, 15
vWA	vWA	.	.	.	17, 18
TP	TPOX	.	.	.	8
D18	D18S51	.	.	.	15, 19
AM	AMEL	rouge	PET	596	x
D5	D5S818	.	.	.	11
FGA	FGA	.	.	.	23, 24

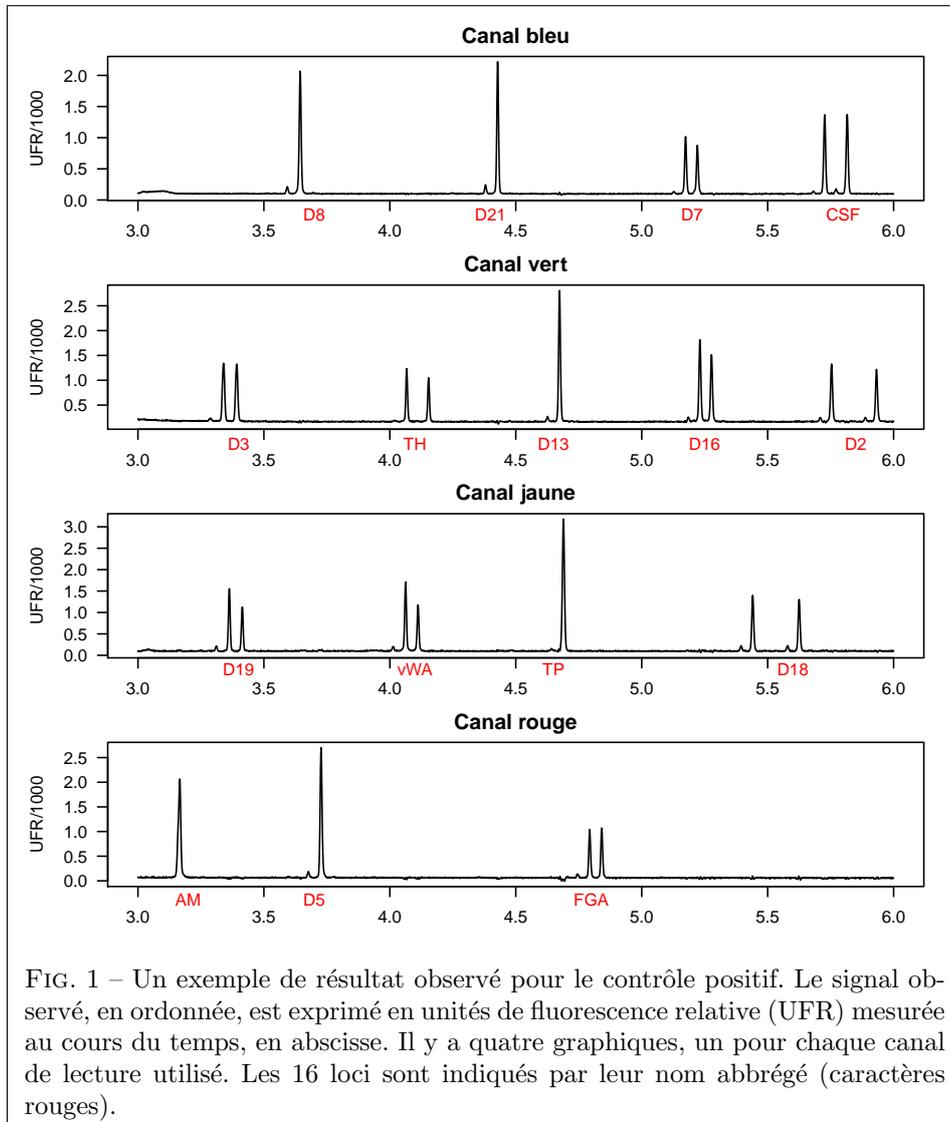
TAB. 1 – Cette table donne le génotype attendu pour le contrôle positif dit DNA 9947A aux 16 loci étudiés. Les deux premières colonnes donnent le nom abrégé des loci et leur nom complet en suivant la nomenclature de l'appendice I de [2] et de la table 2 de [3]. La dernière colonne donne le génotype attendu, par exemple au locus D8 on s'attend à avoir un homozygote portant l'allèle 13, au locus D7 un hétérozygote portant les allèles 10 et 11 (le nom des allèles suit la nomenclature de [1]). Les colonnes du milieu (Canal, Chromophore et λ) donnent la longueur d'onde (λ) à laquelle se fait la lecture du signal, cette longueur d'onde est contrôlée par l'utilisation de différents chromophores (par exemple 6-FAM) qui ont des spectres d'émission caractéristiques. Par commodité on résume cette information en disant que la lecture a été faite dans la couleur d'émission principale (colonne Canal). Par exemple la lecture de D8 se fait dans le bleu, la lecture de D3 dans le vert.

3 La hauteur des pics

La hauteur des pics est théoriquement proportionnelle à la quantité initiale d'ADN avant amplification, la hauteur des pics doit être suffisante pour pouvoir exploiter les résultats. Pour pouvoir comparer directement un locus homozygote à un locus hétérozygote on utilise en fait la somme des hauteurs des pics à chaque locus. Chaque expérience de contrôle positif est donc caractérisée par 16 variables quantitatives.

3.1 Importation des données

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/hauteurs2.RData"))
class(hauteurs)
[1] "data.frame"
colnames(hauteurs)
[1] "D8" "D21" "D7" "CSF" "D3" "TH" "D13" "D16" "D2" "D19" "vWA" "TP" "D18"
[14] "AM" "D5" "FGA"
dim(hauteurs)
```



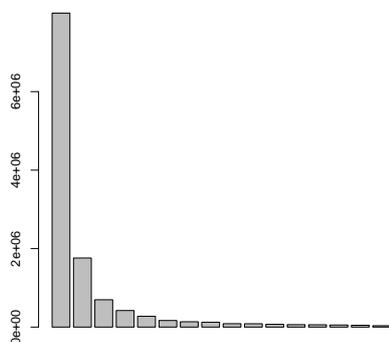
[1] 205 16

De combien de d'expériences de contrôle positif disposons nous en tout ici ?

Réponse :

On décide de synthétiser l'information disponible avec l'analyse en composantes principales suivante :

```
library(ade4)
h.acp <- dudi.pca(hauteurs, scale = FALSE, scann = FALSE, nf = 2)
barplot(h.acp$eig)
```



Combien de facteurs ont-ils été conservés pour l'analyse ?

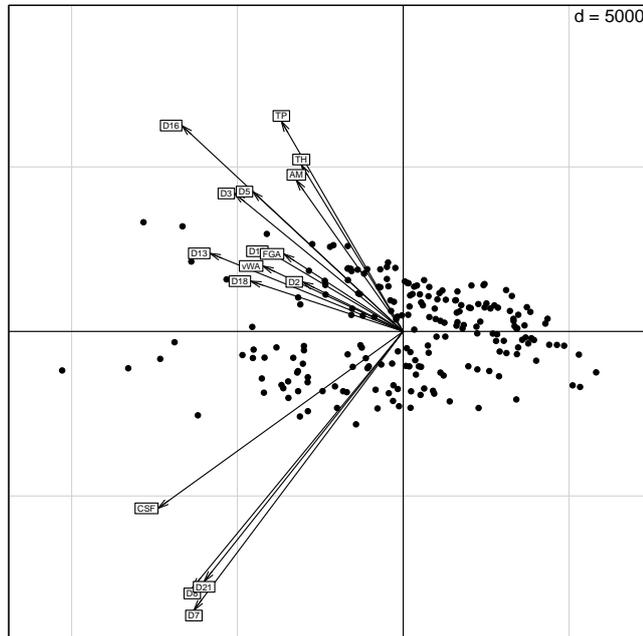
Réponse :

Pourquoi a-t-on utilisé l'option `scale = FALSE` ?

Réponse :

Le premier plan factoriel est le suivant :

```
scatter(h.acp, clab.row = 0, posieig = "none", clab.col = 0.5)
```



Au vu du premier plan factoriel, quelle interprétation pouvez-vous donner du premier et du deuxième facteur ?

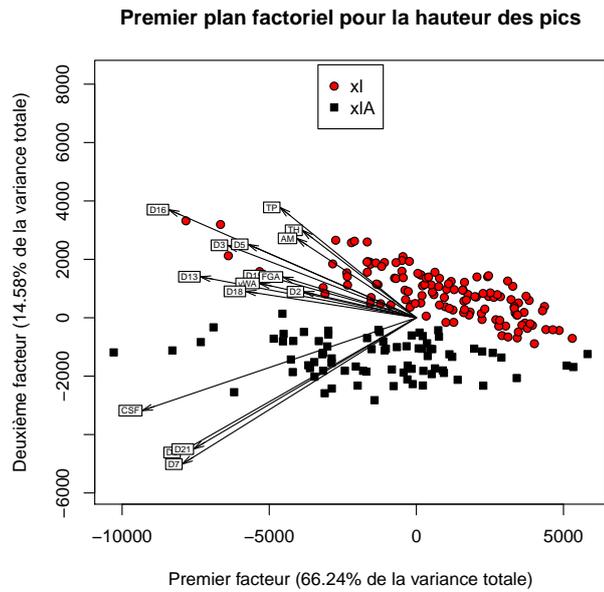
Réponse :

3.2 Utilisation d'une information supplémentaire

Les expériences de contrôle positif ont été réalisées sur deux appareils différents notés `x1` et `x1A`. Dans la table analysée les individus de 1 à 126 correspondent à `x1` et les individus 127 à 205 à `x1A`. On décide de porter cette information supplémentaire sur le premier plan factoriel ainsi :

```
x1 <- 1:126
x1A <- 127:nrow(hauteurs)
x <- h.acp$li[, 1]
y <- h.acp$li[, 2]
xlab <- paste("Premier facteur (", round(100 * h.acp$eig[1]/sum(h.acp$eig),
2), "% de la variance totale)", sep = "")
ylab <- paste("Deuxième facteur (", round(100 * h.acp$eig[2]/sum(h.acp$eig),
2), "% de la variance totale)", sep = "")
plot(x[x1], y[x1], asp = 1, xlab = xlab, ylab = ylab, pch = 21,
bg = "red", main = "Premier plan factoriel pour la hauteur des pics",
```

```
xlim = range(x)
points(x[x1A], y[x1A], bg = "black", pch = 22)
legend("top", inset = 0.01, legend = c("x1", "x1A"), pch = c(21,
22), pt.bg = c("red", "black"))
s.corcircle(9 * h.acp$co, add = TRUE, clabel = 0.5)
```



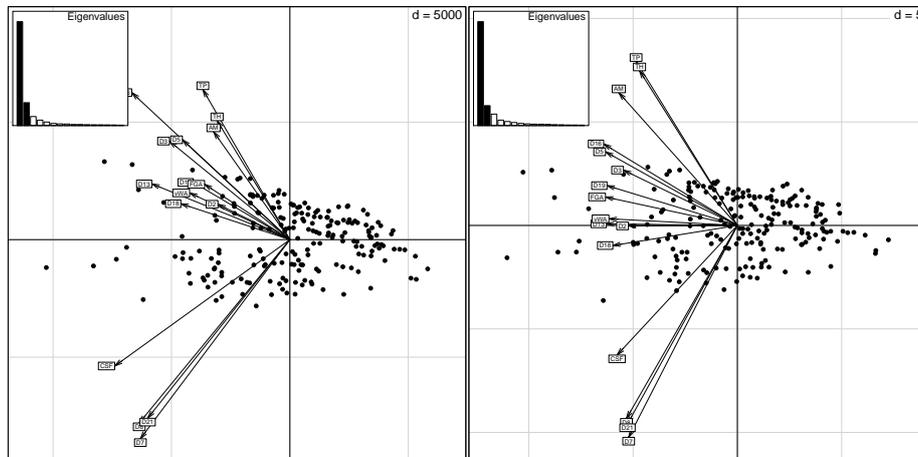
Comment interprétez vous maintenant le deuxième facteur de l'analyse ?
Quelle est la différence entre les deux appareils ?

Réponse :

3.3 Comparaison avec une autre analyse

On compare les résultats avec une autre analyse en composantes principales :

```
h2.acp <- dudi.pca(hauteurs, scale = TRUE, scann = FALSE, nf = 2)
par(mfrow = c(1, 2))
scatter(h.acp, clab.row = 0, clab.col = 0.5)
scatter(h2.acp, clab.row = 0, clab.col = 0.5)
```



Est-ce qu'il y a une grande différence entre les résultats de ces deux analyses ?
Qu'en déduisez vous ?

Réponse :

4 L'équilibre d'hétérozygotie

Dans le cas d'un locus hétérozygote il y a deux pics. Ces pics ont théoriquement la même hauteur puisque dans un mélange cellulaire de cellules diploïdes il y a *a priori* autant de d'allèles d'origine paternelle que maternelle.

On appelle équilibre d'hétérozygotie ϕ_i (en utilisant les notations de [6]) au locus i la valeur :

$$\phi_i = 100 \times \frac{y_i^{\min}}{y_i^{\max}} \quad (1)$$

où y_i^{\min} et y_i^{\max} sont les hauteurs du plus petit et du plus grand pic au locus i , respectivement.

4.1 Importation des données

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/eqh2.RData"))
class(eqh)
[1] "data.frame"
colnames(eqh)
[1] "D7" "CSF" "D3" "TH" "D16" "D2" "D19" "vWA" "D18" "FGA"
```

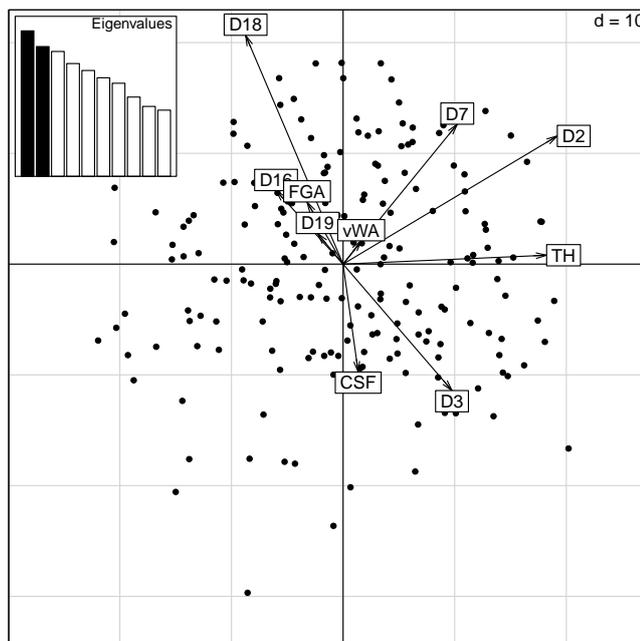
```
dim(eqh)
[1] 205 10
```

Combien de colonnes sont elles présentes dans ce jeu de données ? Pourquoi ce nombre est il moindre qu'avec le jeu de données précédent avec les hauteurs de pics ?

Réponse :

4.2 Analyse multivariée

```
e.acp <- dudi.pca(eqh, scale = FALSE, scann = FALSE, nf = 2)
scatter(e.acp, clab.row = 0)
```



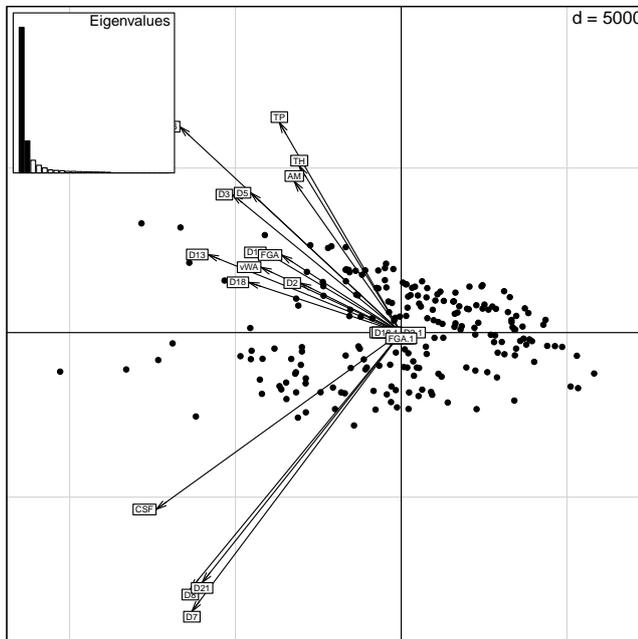
Comment interprétez-vous ces résultats ?

Réponse :

5 Analyses simultanées

On décide de regrouper les données sur la hauteur des pics et celles sur l'équilibre d'hétérozygotie dans une même table pour l'analyse :

```
tout <- cbind(hauteurs, eqh)
colnames(tout) <- make.names(colnames(tout), unique = TRUE)
class(tout)
[1] "data.frame"
colnames(tout)
[1] "D8" "D21" "D7" "CSF" "D3" "TH" "D13" "D16" "D2" "D19"
[11] "vWA" "TP" "D18" "AM" "D5" "FGA" "D7.1" "CSF.1" "D3.1" "TH.1"
[21] "D16.1" "D2.1" "D19.1" "vWA.1" "D18.1" "FGA.1"
dim(tout)
[1] 205 26
tout.acp <- dudi.pca(tout, scale = FALSE, scann = FALSE, nf = 2)
scatter(tout.acp, clab.row = 0, clab.col = 0.5)
```



Expliquez le résultat obtenu.

Réponse :

Références

- [1] 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.
- [2] J.M. Butler. *Forensic DNA typing : biology, technology, and genetics of STR markers*. Elsevier, New York, USA, 2005.
- [3] 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.
- [4] A. Debernardi. Développement d'un outil informatique pour l'analyse automatisée de données de validation de méthodes d'analyses de profils génétiques. Technical report, Rapport technique du Master Santé et Populations spécialité Biostatistiques Bioinformatique Génome, Université Claude Bernard - Lyon 1, 2009.
- [5] A. Debernardi, E. Suzanne, A. Formant, L. Pène, A.B. Dufour, and J.R. Lobry. One year variability of peak heights, heterozygous balance and inter-locus balance for the DNA positive control of AmpF ℓ STR $^{\circledR}$ Identifiler $^{\circledR}$ STR kit. *Forensic Science International : Genetics*, in revision :0–0, 2009.
- [6] P. Gill, R. Sparkes, and C. Kimpton. Development of guidelines to designate alleles using an STR multiplex system. *Journal of Forensic Sciences*, 89 :185–197, 1997.