Ecole supérieure d'agro-développement international
istom

Examen de statistiques et analyse de données

Mercredi 10 décembre 2025

Promotion 114

Antoine Géré

Document(s) autorisé(s) : ☐ Oui ☒ Non Calculatrice autorisée : ☒ Oui ☐ Non

Remarques :

- Les exercices sont indépendants.
- Il sera tenu compte de la propreté de votre copie, ainsi que de la clarté et de la qualité de la rédaction et du raisonnement.
- **Ne pas écrire avec un crayon papier**, sauf pour dessiner et/ou annoter des croquis, le cas échéant.
- Utiliser les **notations** indiquées dans le texte et **justifier toutes vos réponses**.
- Le sujet est à conserver par l'étudiant-e.

Exercice 1 ANOVA à un facteur sous R : la taille des œufs du coucou

Le coucou gris ne couve pas ses propres œufs : il les dépose dans les nids d'oiseaux d'autres espèces, qui les couvent à sa place. Des études antérieures ont montré que les coucous ont parfois évolué pour pondre des œufs dont la *couleur* est proche de celle des œufs de l'espèce hôte. Est-ce également le cas pour leur *taille* ? Les coucous pondent-ils des œufs de taille différente selon l'espèce qui les héberge ?



Le fichier `data_stat-0115.csv` contient une ligne par œuf de coucou et deux colonnes : `host_species` (l'espèce hôte dans le nid de laquelle l'œuf a été trouvé) et `egg_length` (la longueur de l'œuf, en mm).

```
library(dplyr); library(ggplot2); library(car)
cuckooeggs <- read.csv("data_raw/data_stat-0115.csv")
head(cuckooeggs)
summary(cuckooeggs)
```

R affiche :

	host_species	egg_length
1	Hedge Sparrow	20,85
2	Hedge Sparrow	21,65
3	Hedge Sparrow	22,05
4	Hedge Sparrow	22,85
5	Hedge Sparrow	23,05
6	Hedge Sparrow	23,05

head(cuckooeggs)

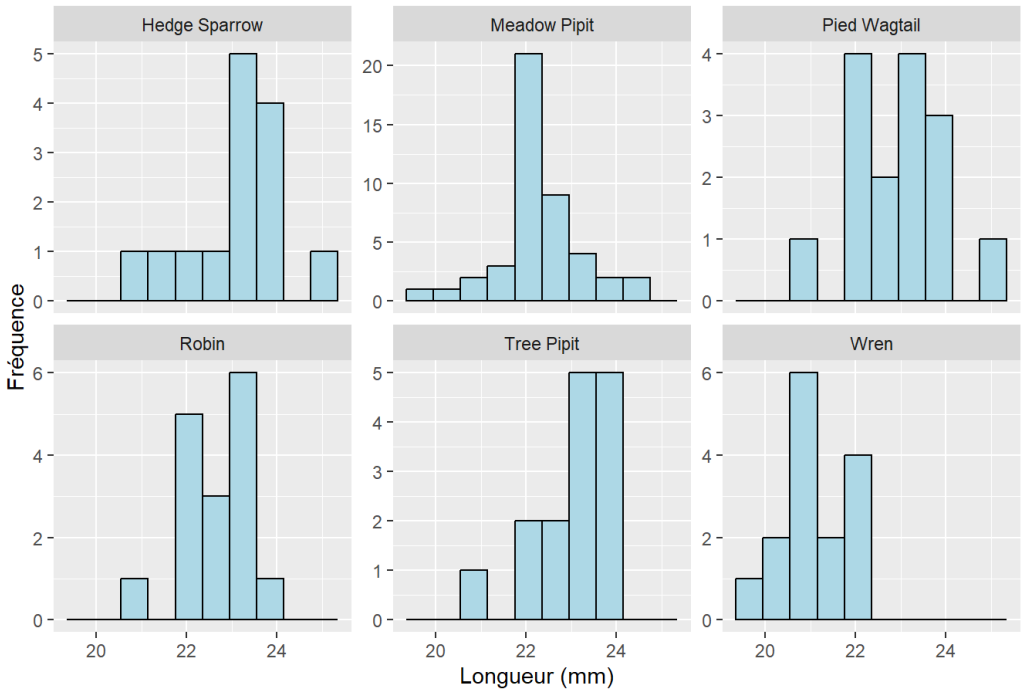
host_species	egg_length
Length : 120	Min. : 19,65
Class : character	1st Qu. : 21,85
Mode : character	Median : 22,35
	Mean : 22,46
	3rd Qu. : 23,25
	Max. : 25,05

summary(cuckooeggs)

1. **Le fichier de données.** Présenter le fichier de données en indiquant le nombre d'observations, le nombre de variables, leur type ainsi que leurs principales caractéristiques.

2. **Interprétation de graphiques.** On trace les graphiques suivants :

```
ggplot(cuckooeggs, aes(x = egg_length)) +
  geom_histogram(bins = 10, fill = "lightblue", color = "black") +
  facet_wrap(~ host_species, scales = "free_y") +
  labs(x = "Longueur (mm)", y = "Frequence")
```



a) Comment se nomme ce type de graphique ?
b) Décrire les différences visibles entre les distributions des longueurs d'œufs selon les espèces hôtes.
c) Les variances semblent-elles comparables ?
d) Les distributions paraissent-elles approximativement normales ?
e) Les moyennes semblent-elles clairement différentes ?

3. **Interprétation d'un tableau descriptif.** On demande à R la moyenne, l'écart-type et l'effectif de la longueur des œufs pour chaque espèce hôte :

```
stats <- cuckooeggs %>%
  group_by(host_species) %>%
  summarise(Moyenne = mean(egg_length), Ecart_type = sd(egg_length), n = n())
```

```
stats
```

host_species	Moyenne	Ecart_type	n
Hedge Sparrow	23,1	1,07	14
Meadow Pipit	22,3	0,921	45
Pied Wagtail	22,9	1,07	15
Robin	22,6	0,685	16
Tree Pipit	23,1	0,901	15
Wren	21,1	0,744	15

a) Quelle espèce hôte présente les œufs les plus grands en moyenne ?

b) Quelle espèce présente les œufs les plus petits en moyenne ?

c) Les différences de moyennes vous semblent-elles importantes ?

d) Les écarts-types semblent-ils homogènes ?

4. **Interprétation d'une erreur.** Afin de savoir si les coucous pondent des œufs de tailles différentes selon l'espèce hôte, on se propose de réaliser un test de Student :

```
t.test(egg_length ~ host_species, data = cuckooeggs)
```

R renvoie le message d'erreur `grouping factor must have exactly 2 levels`. Expliquer pourquoi ce `t.test` ne fonctionne pas dans ce cas précis.

5. **Conditions d'application de l'ANOVA.** On envisage alors une analyse de la variance (ANOVA).

a) Rappeler les conditions nécessaires à l'ANOVA à un facteur.

b) En vous appuyant sur les histogrammes et sur le tableau des moyennes et écarts-types, indiquer si elles semblent raisonnablement satisfaites.

c) Rappeler l'hypothèse nulle de l'ANOVA dans cette étude.

6. **Test de Levene.** Sur le conseil d'un collaborateur, on réalise un test de Levene :

```
leveneTest(egg_length ~ host_species, data = cuckooeggs)
```

Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	5	0,6397	0,6698
	114		

a) Donner l'hypothèse nulle de ce test.

b) Interpréter ce test. Que peut-on en conclure ?

7. **Tests de Shapiro–Wilk.** Le même collaborateur recommande vivement d'effectuer des tests de Shapiro–Wilk. On en réalise donc plusieurs : un par espèce hôte, puis un sur les résidus de l'ANOVA.

```
by(cuckooeggs$egg_length, cuckooeggs$host_species, shapiro.test)
anova_res <- aov(egg_length ~ host_species, data = cuckooeggs)
shapiro.test(residuals(anova_res))
```

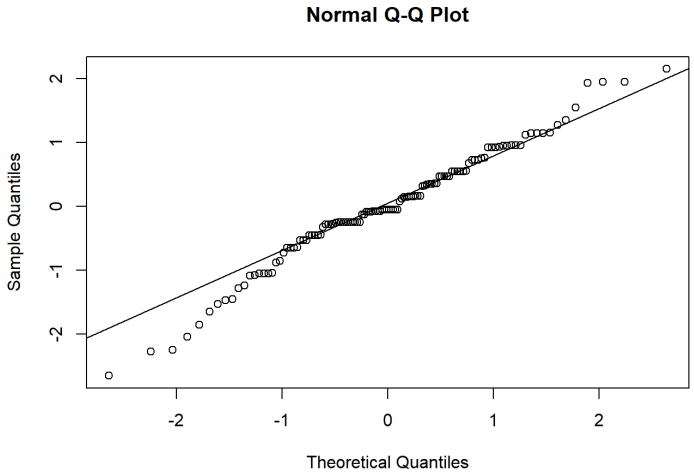

Tournez la page S.V.P.

Données testées	W	p-value
Hedge Sparrow	0,94843	0,5366
Meadow Pipit	0,93006	0,009424
Pied Wagtail	0,96471	0,7736
Robin	0,95212	0,5239
Tree Pipit	0,89772	0,08786
Wren	0,93295	0,3019
`residuals(anova_res)`	0,9804	0,07761

a) Donner l'hypothèse nulle du test de Shapiro–Wilk.
b) Interpréter chacun de ces tests.
c) Étaient-ils tous nécessaires ?
d) Que peut-on en conclure pour l'ANOVA ?

8. Q-Q plot. Un autre collaborateur examine le travail et indique qu'il aurait été pertinent de produire un Q-Q plot. On trace ce graphique :

```
qqnorm(residuals(anova_res))
qqline(residuals(anova_res))
```



a) Rappeler la définition d'un Q-Q plot et expliquer en quoi la remarque de ce collaborateur était pertinente.
b) Analyser et interpréter ce graphique.
c) Ces résultats concordent-ils avec le test de Shapiro–Wilk effectué précédemment ?

9. Résultat de l'ANOVA.

```
summary(anova_res)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
`host_species`	5	42,94	8,588	10,39	$3{,}15 \times 10^{-8}$ ***
`Residuals`	114	94,25	0,827		

a) D'après ces résultats, peut-on conclure que l'effet de l'espèce hôte sur la taille des œufs est statistiquement significatif ?
b) L'ANOVA permet-elle de déterminer exactement quelles espèces diffèrent les unes des autres ?

10. Test post-hoc de Tukey. Toujours sur le conseil du même collaborateur, on réalise un test post-hoc :

```
tukey <- TukeyHSD(anova_res)
tukey
```

Tukey multiple comparisons of means – 95% family-wise confidence level

$host_species	diff	lwr	upr	p adj
Meadow Pipit – Hedge Sparrow	−0,823	−1,629	−0,016	0,0429
Pied Wagtail – Hedge Sparrow	−0,218	−1,198	0,761	0,9872
Robin – Hedge Sparrow	−0,546	−1,511	0,418	0,5726
Tree Pipit – Hedge Sparrow	−0,031	−1,011	0,948	1,0000
Wren – Hedge Sparrow	−1,991	−2,971	−1,012	0,0000006
Pied Wagtail – Meadow Pipit	0,604	−0,181	1,390	0,2325
Robin – Meadow Pipit	0,276	−0,491	1,043	0,9022
Tree Pipit – Meadow Pipit	0,791	0,005	1,577	0,0475
Wren – Meadow Pipit	−1,169	−1,955	−0,383	0,0005
Robin – Pied Wagtail	−0,328	−1,276	0,619	0,9155
Tree Pipit – Pied Wagtail	0,187	−0,776	1,149	0,9932
Wren – Pied Wagtail	−1,773	−2,736	−0,811	0,0000070
Tree Pipit – Robin	0,515	−0,432	1,462	0,6160
Wren – Robin	−1,445	−2,392	−0,498	0,0003
Wren – Tree Pipit	−1,960	−2,922	−0,998	0,0000006

a) Rappeler l'objectif d'un test post-hoc après une ANOVA.

b) D'après les résultats du test de Tukey, quelles paires d'espèces présentent des différences significatives ?

c) Quelles conclusions peut-on tirer de ces résultats ?

Indication ▼ Correction ▼ [stat-0115]

Exercice 2 ACP sous R : l'expression des gènes dans l'embryon humain précoce

Le jeu de données de cet exercice est inspiré d'une étude publiée dans la revue *Cell* (Petropoulos et al., 2016), dans laquelle des chercheurs ont mesuré l'activité de nombreux gènes dans des cellules d'embryons humains très précoces. Leur objectif était d'observer comment les cellules évoluent au cours du développement et comment elles se différencient en plusieurs types cellulaires.

On utilise ici une version simplifiée de ces données. Le fichier `data_stat-0116.xlsx` contient :

- 300 **gènes** : chaque ligne correspond à un gène différent ;
- une colonne `Gene`, qui donne le nom du gène ;
- 12 colonnes quantitatives, une par **condition biologique** (un jour de développement et, à partir du jour 5, un type cellulaire : épiblaste `EPI`, endoderme primitif `PE` ou trophectoderme `TE`). Chaque valeur mesure l'expression du gène dans la condition correspondante : `Day_3`, `Day_4`, `Pre_Day_5`, `Day_5_EPI`, `Day_5_PE`, `Day_5_TE`, `Day_6_EPI`, `Day_6_PE`, `Day_6_TE`, `Day_7_EPI`, `Day_7_PE`, `Day_7_TE`.

L'objectif de l'analyse en composantes principales (ACP) est d'identifier les grands axes qui résument les principales différences d'expression entre ces 12 conditions, et de visualiser si certaines conditions se regroupent selon des profils similaires.

```
library(readxl); library(corrplot); library(FactoMineR); library(factoextra)
data <- read_excel("data_raw/data_stat-0116.xlsx")
var.quanti  <- sapply(data, is.numeric)    # colonnes numeriques
data.quanti <- data[, var.quanti]
head(data.quanti)
```

R affiche (les trois dernières colonnes, `Day_7_EPI`, `Day_7_PE` et `Day_7_TE`, ne sont pas reproduites) :

Gene	Day_3	Day_4	Pre_Day_5	Day_5_EPI	Day_5_PE	Day_5_TE	Day_6_EPI	Day_6_PE	Day_6_TE	...
SEPT6	1,5	0	0,2	2,2	0,6	0,3	9,8	0,7	0,4	...
AARS2	2,3	13,8	17,3	44,1	19,5	8,5	21,8	7,5	4,8	...
ABCG2	0,2	0,3	7,6	3,3	10,4	35,1	23	54	71,3	...
ABHD12B	5,8	11	16,7	33,3	17	12,3	57,4	18,3	9,8	...
ABHD6	0,8	4,2	9,7	7,5	29,2	8,1	11,8	14,8	8,8	...
ACE	0,3	17,5	10,3	15,4	6,7	3,9	11,1	1,3	0,8	...

1. **Questions préliminaires.**

   a) Combien de gènes et de conditions le jeu de données contient-il ? Quels sont les individus statistiques ?

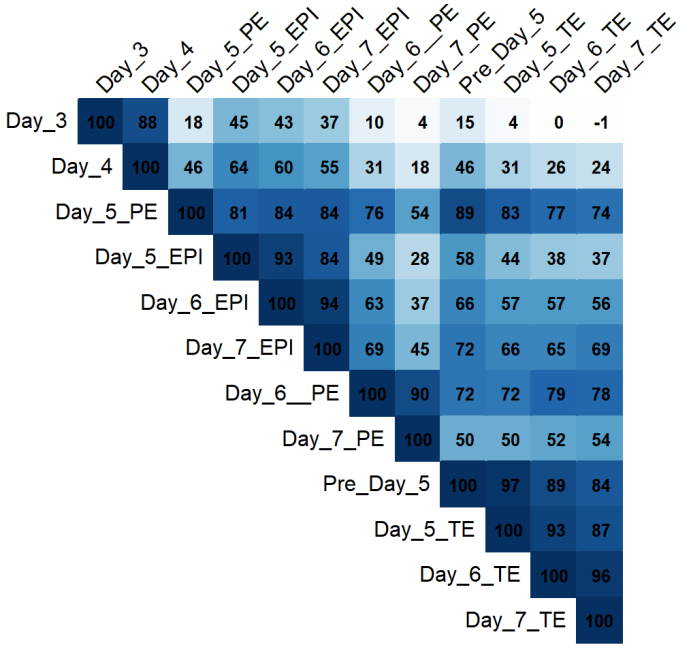
   b) Quelles sont les variables quantitatives ?

   c) Pourquoi est-il important de distinguer les variables quantitatives et qualitatives avant l'analyse ?

   d) Expliquer en quelques phrases l'objectif de l'ACP. Pourquoi cette méthode est-elle utilisée ici ?

2. **Matrice des corrélations.** On représente la matrice des corrélations entre les 12 conditions (les coefficients sont affichés en %) :

```
Matrice.Correlation <- cor(data.quanti, use = "complete.obs")
corrplot(Matrice.Correlation, method = "color", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, addCoef.col = "black",
         addCoefasPercent = TRUE, number.cex = 0.8)
```



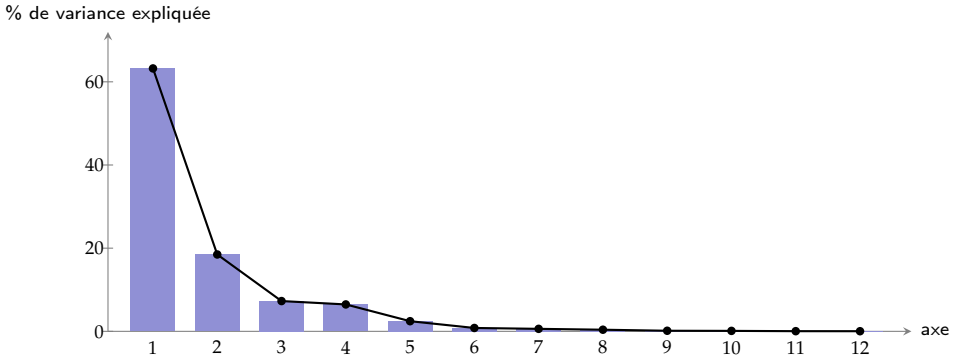
   a) Quelles conditions semblent fortement corrélées ? Lesquelles semblent peu corrélées ?

   b) Que peut-on en déduire pour l'analyse en composantes principales ?

3. **Variance expliquée.** On réalise l'ACP sur les données centrées réduites, puis on affiche les valeurs propres et l'éboulis :

```
data.CR  <- scale(data.quanti, center = TRUE, scale = TRUE)
pca.data <- PCA(data.CR, graph = FALSE)
vp <- pca.data$eig
vp
barplot(vp[, 2], names.arg = 1:nrow(vp), xlab = "Composantes principales",
        ylab = "Pourcentage de variance expliquee", col = "steelblue")
lines(x = 1:nrow(vp), vp[, 2], type = "b", pch = 19, col = "black")
```

Axe	1	2	3	4	5	6	7	8	9	10	11	12
Valeur propre	7,586	2,217	0,873	0,776	0,293	0,097	0,072	0,046	0,016	0,014	0,005	0,003
% de variance	63,22	18,48	7,28	6,47	2,44	0,81	0,60	0,39	0,14	0,11	0,04	0,03
% cumulé	63,22	81,69	88,97	95,44	97,88	98,69	99,29	99,68	99,82	99,93	99,97	100



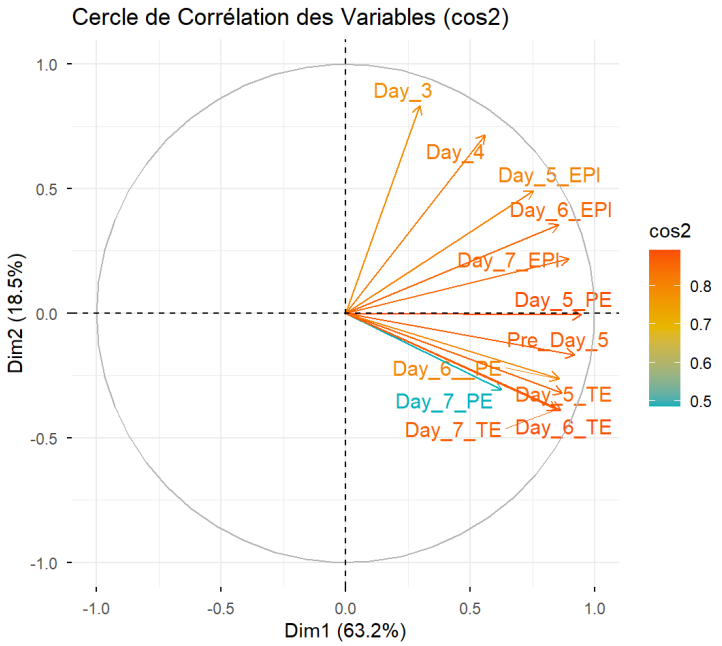
a) Que représente la variance expliquée par chaque composante principale ?

b) Combien de composantes principales retiendriez-vous pour représenter l'essentiel de l'information ? Justifier.

c) Quelle proportion de la variance est expliquée par les deux premières composantes ?

4. Cercle des corrélations.

```
fviz_pca_var(pca.data, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE,
             title = "Cercle de correlation des variables (cos2)")
```



a) Identifier deux variables fortement corrélées et deux variables faiblement corrélées.

b) Que signifient la longueur et l'orientation des flèches ?

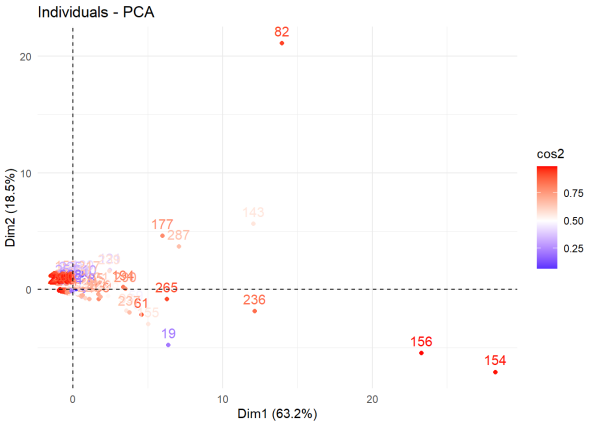
c) Quelles variables sont fortement liées à l'axe 2 ?

5. Représentation des individus. On affiche les coordonnées des premiers gènes, la carte des gènes coloriée selon leur qualité de représentation, puis un biplot ne retenant que les gènes et les variables dont le $\cos^2$ dépasse 0,5 :

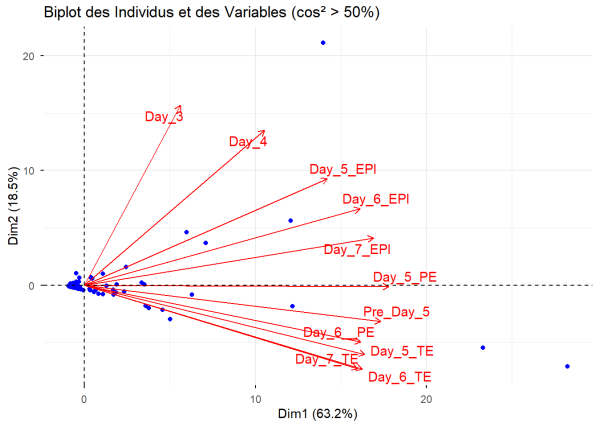
Tournez la page S.V.P.

```
head(pca.data$ind$coord)
fviz_pca_ind(pca.data, col.ind = "cos2") +
  scale_color_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0.50) +
  theme_minimal()
fviz_pca_biplot(pca.data, select.ind = list(cos2 = 0.5), select.var = list(cos2 = 0.5),
                repel = TRUE, col.ind = "blue", col.var = "red", geom.ind = "point",
                title = "Biplot des individus et des variables (cos2 > 0.5)")
```

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	−0,886	−0,057	−0,060	0,033	0,038
2	−0,606	0,109	−0,122	−0,069	−0,105
3	−0,370	−0,320	0,032	0,104	0,316
4	−0,506	0,111	−0,114	−0,140	0,027
5	−0,689	−0,077	−0,029	0,001	−0,076
6	−0,779	0,061	−0,069	0,077	−0,059



Carte des gènes (couleur : $\cos^2$)



Biplot ($\cos^2 > 0{,}5$)

Que révèle la position des individus sur le plan des deux premiers axes (proximité, similarité, groupes, individus atypiques) ?

Référence. Petropoulos S., Edsgärd D., Reinius B., Deng Q., Panula S. P., Codeluppi S., Reyes A. P., Linnarsson S., Sandberg R., Lanner F. (2016). Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell*, 165, 1012–1026.

Indication ▼ Correction ▼ [stat-0116]

Indication pour l'exercice 1 ▲

1. `Length : 120` donne le nombre de lignes. Une colonne `character` est une variable qualitative : combien a-t-elle de modalités (regarder le tableau de la question 3) ?
2. Les barres sont accolées et l'axe horizontal est continu. Comparer la *position* des distributions (où se trouve le pic ?), leur *étalement* et leur *forme* (symétrie, un ou plusieurs pics).
3. Exprimer l'écart entre la plus grande et la plus petite moyenne en mm, en % de la longueur moyenne, puis en nombre d'écarts-types. Pour les écarts-types, calculer le rapport du plus grand au plus petit (règle pratique : il doit rester inférieur à 2).
4. Combien de groupes un test de Student compare-t-il ? Combien de comparaisons deux à deux faudrait-il ici ? Que devient alors le risque de première espèce global ?
5. Trois conditions : indépendance, normalité (de quoi exactement ?), homogénéité des variances. L'hypothèse nulle porte sur les moyennes des six groupes.
6. Le test de Levene compare des variances. Comparer la p-valeur à $0{,}05$.
7. Pour chaque ligne, comparer la p-valeur à $0{,}05$. Le modèle de l'ANOVA fait une hypothèse sur la loi des *résidus*, pas sur celle de chaque groupe.
8. Si les résidus suivent une loi normale, les points sont alignés sur la droite. Les tests de normalité sont peu puissants pour de petits échantillons.
9. Comparer la p-valeur à $0{,}05$. On peut aussi calculer la part de la variabilité totale expliquée par l'espèce hôte : $R^2 = \dfrac{SCE_{\text{inter}}}{SCE_{\text{totale}}}$.
10. Une différence est significative si la p-valeur ajustée est inférieure à $0{,}05$, ce qui revient à dire que l'intervalle `[ lwr ; upr ]` ne contient pas 0. On peut résumer les résultats par des lettres : deux espèces qui partagent une lettre ne diffèrent pas significativement.

Indication pour l'exercice 2 ▲

1. Les lignes sont les individus, les colonnes numériques les variables actives. Une ACP calcule des moyennes, des variances et des corrélations : cela a-t-il un sens pour un nom de gène ?
2. Une case foncée avec un nombre proche de 100 signale une forte corrélation. Comparer les conditions d'un même type cellulaire à des jours consécutifs, puis les jours les plus éloignés. L'ACP « fusionne » des variables très corrélées sur un même axe.
3. En ACP normée, la somme des valeurs propres vaut le nombre de variables, ici 12. Critères usuels : valeur propre supérieure à 1 (Kaiser), coude de l'éboulis, pourcentage cumulé d'au moins $80\,\%$.
4. Angle entre deux flèches : proche de $0°$, corrélation positive forte ; proche de $90°$, corrélation faible ; proche de $180°$, corrélation négative. Une flèche n'est interprétable que si elle est longue (proche du cercle).
5. Deux gènes proches sur la carte ont des profils d'expression semblables sur les 12 conditions, *à condition* d'être bien représentés ($\cos^2$ élevé). Que signifie l'axe 1 si toutes les flèches pointent du même côté ?

Tournez la page S.V.P.

Correction de l'exercice 1 ▲

1. **Le fichier de données.** Le tableau `cuckooeggs` contient **120 observations** (une par œuf) et **2 variables** :

Variable	Type	Description
`host_species`	qualitative nominale (facteur)	espèce hôte dans le nid de laquelle l'œuf a été pondu
`egg_length`	quantitative continue	longueur de l'œuf (mm)

   - `host_species` a 6 modalités : Hedge Sparrow (accenteur mouchet), Meadow Pipit (pipit farlouse), Pied Wagtail (bergeronnette grise), Robin (rougegorge), Tree Pipit (pipit des arbres) et Wren (troglodyte). Les effectifs sont déséquilibrés : 45 œufs pour Meadow Pipit, 14 à 16 pour les autres.
   - `egg_length` varie de 19,65 à 25,05 mm, avec une moyenne de 22,46 mm et une médiane de 22,35 mm, très proches : la distribution d'ensemble est à peu près symétrique.

   Une variable quantitative et une variable qualitative à plusieurs modalités : les données se prêtent à une *comparaison de moyennes entre groupes*.

2. **Interprétation des graphiques.**

   a) Ce sont des **histogrammes**, un par espèce hôte. L'axe horizontal porte des classes de longueur (mm), l'axe vertical l'effectif de chaque classe. Les classes sont contiguës (barres accolées), ce qui distingue un histogramme d'un diagramme en bâtons. L'option `scales = "free_y"` donne à chaque panneau sa propre échelle verticale : on ne compare donc pas les hauteurs d'un panneau à l'autre.

   b)
   - *Position* : les histogrammes ne sont pas centrés au même endroit. Hedge Sparrow et Tree Pipit sont centrés vers 23 mm ; Wren est nettement décalé vers la gauche (surtout entre 20 et 22 mm) ; Meadow Pipit, Pied Wagtail et Robin occupent une position intermédiaire.
   - *Dispersion* : certaines distributions sont un peu plus resserrées que d'autres, mais aucune n'est beaucoup plus étalée.
   - *Forme* : toutes les distributions sont unimodales, à peu près symétriques, sans valeur aberrante nette.

   c) Oui, à l'œil : les étendues et le nombre de classes occupées sont du même ordre d'une espèce à l'autre. Cette impression devra être confirmée par un test (question 6).

   d) Globalement oui : formes en cloche, observations concentrées autour d'une valeur centrale, effectifs qui décroissent de part et d'autre, pas de longue queue. Avec 14 à 16 valeurs par groupe, un histogramme reste toutefois un outil grossier.

   e) Pas toutes. Wren semble nettement en dessous des autres ; Hedge Sparrow, Tree Pipit et Pied Wagtail semblent proches. Les autres différences sont trop faibles pour être jugées à l'œil : il faut un test.

3. **Interprétation du tableau descriptif.**

   a) Hedge Sparrow et Tree Pipit : 23,1 mm en moyenne.

   b) Wren : 21,1 mm en moyenne, nettement en dessous de toutes les autres espèces.

   c) L'écart maximal est $23{,}1 - 21{,}1 = \mathbf{2{,}0}$ mm, soit $\dfrac{2{,}0}{22{,}46} \approx 9\,\%$ de la longueur moyenne d'un œuf : ce n'est pas négligeable. Rapporté à la dispersion à l'intérieur des groupes (écarts-types de 0,7 à 1,1 mm), il représente environ 2 écarts-types : une telle différence a de bonnes chances d'être significative. Les autres écarts (de l'ordre de 0,2 à 0,8 mm) sont plus difficiles à juger.

   d) Les écarts-types vont de 0,685 (Robin) à 1,07 (Hedge Sparrow et Pied Wagtail). Leur rapport vaut

$$\frac{1{,}07}{0{,}685} \approx 1{,}56.$$

   En pratique, un rapport des écarts-types inférieur à 2 est jugé acceptable (cela correspond à un rapport des variances inférieur à 4 ; ici $1{,}56^2 \approx 2{,}4$). Les dispersions sont donc comparables sans être identiques : l'homogénéité des variances n'exige pas une égalité parfaite.

4. **Interprétation de l'erreur.** Le test de Student (pour échantillons indépendants) compare les moyennes de **deux** groupes : sa statistique repose sur *une* différence de moyennes. Or, dans la formule `egg_length ~ host_species`, la variable explicative `host_species` a **six** modalités : il n'existe pas de différence unique à tester, d'où le message d'erreur.

   On pourrait faire tous les tests de Student deux à deux, soit $\binom{6}{2} = 15$ comparaisons, mais chaque test comporte un risque $\alpha = 5\,\%$ de conclure à tort à une différence. Si les 15 tests étaient indépendants, la probabilité de commettre au moins une erreur de première espèce serait

   $$1 - 0{,}95^{15} \approx 0{,}54,$$

   soit plus d'une chance sur deux ! Il faut un test qui compare les six moyennes *en une seule fois* : l'**ANOVA à un facteur**.

5. **Conditions d'application de l'ANOVA.**

   a)
   - **Indépendance des observations** : chaque œuf est mesuré une seule fois et la mesure d'un œuf n'influence pas celle d'un autre (en pratique : des œufs provenant de nids, et si possible de coucous, différents).
   - **Normalité des résidus** : les écarts entre chaque observation et la moyenne de son groupe suivent une loi normale.
   - **Homogénéité des variances** (homoscédasticité) : la variance est la même dans tous les groupes.

   b)
   - *Indépendance* : elle relève du protocole et ne se vérifie pas sur les graphiques ; les œufs provenant de nids d'espèces différentes, on peut raisonnablement la supposer.
   - *Normalité* : les histogrammes sont unimodaux, à peu près symétriques, sans valeur aberrante ; la condition paraît raisonnablement satisfaite (on la testera à la question 7).
   - *Homogénéité des variances* : rapport des écarts-types $\approx 1{,}56 < 2$ et étalements visuellement comparables ; la condition paraît satisfaite (on la testera à la question 6).

   c) En notant $\mu_1, \ldots, \mu_6$ les longueurs moyennes des œufs de coucou dans les nids des six espèces hôtes :

   $$H_0 : \mu_1 = \mu_2 = \cdots = \mu_6 \qquad \text{contre} \qquad H_1 : \text{au moins deux moyennes diffèrent.}$$

   Autrement dit, sous H_0, les coucous pondent des œufs de même longueur moyenne quelle que soit l'espèce hôte.

6. **Test de Levene.**

   a) Le test de Levene compare les variances de plusieurs groupes :

   $$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_6^2 \qquad \text{contre} \qquad H_1 : \text{au moins deux variances diffèrent.}$$

   b) $p = 0{,}6698 > 0{,}05$: on **ne rejette pas** H_0. Les données sont compatibles avec des variances égales ; la condition d'homogénéité des variances est considérée comme satisfaite, ce qui confirme l'impression laissée par les histogrammes et par le rapport des écarts-types. (Avec `center = median`, option par défaut de `leveneTest`, il s'agit de la version de Brown–Forsythe, peu sensible à un écart à la normalité.)

7. **Tests de Shapiro–Wilk.**

   a) H_0 : « les données sont issues d'une loi normale ». Si $p < 0{,}05$, on rejette la normalité ; sinon, on ne la rejette pas.

   b)

Données	p-valeur	Conclusion
Hedge Sparrow	0,537	$p > 0{,}05$: normalité non rejetée
Meadow Pipit	0,009	$p < 0{,}05$: **normalité rejetée**
Pied Wagtail	0,774	$p > 0{,}05$: normalité non rejetée
Robin	0,524	$p > 0{,}05$: normalité non rejetée
Tree Pipit	0,088	$p > 0{,}05$: normalité non rejetée
Wren	0,302	$p > 0{,}05$: normalité non rejetée
Résidus de l'ANOVA	0,078	$p > 0{,}05$: normalité non rejetée

   Seule la distribution de Meadow Pipit (le groupe le plus nombreux, 45 œufs) s'écarte significativement d'une loi normale.

c) Non. Le modèle de l'ANOVA suppose la normalité des **résidus** : c'est le test sur `residuals(anova_res)` qui compte. Les six tests par groupe sont une inspection exploratoire utile, mais multiplier les tests augmente le risque d'en voir un « significatif » par hasard (même problème qu'à la question 4), et chacun porte sur peu de données (14 à 16 valeurs), donc avec une faible puissance.

d) Sur les résidus, $p = 0{,}078 > 0{,}05$: on ne rejette pas la normalité. Le léger écart observé pour Meadow Pipit ne remet pas en cause l'analyse, d'autant que l'ANOVA est robuste à de faibles écarts à la normalité lorsque les groupes comptent au moins une quinzaine d'observations. Avec l'homogénéité des variances (question 6) et l'indépendance, les conditions de l'ANOVA sont réunies : on peut l'appliquer.

8. Q-Q plot.

a) Un Q-Q plot (*quantile–quantile plot*, ou droite de Henry) place en abscisse les quantiles théoriques d'une loi normale et en ordonnée les quantiles observés (ici ceux des résidus). Si les données suivent une loi normale, les points sont à peu près alignés sur la droite tracée par `qqline` ; une courbure signale une asymétrie, des points qui s'écartent aux extrémités signalent des queues trop lourdes ou des valeurs extrêmes.

La remarque est pertinente : un test de normalité ne donne qu'une p-valeur, très dépendante de la taille de l'échantillon (peu puissant sur de petits échantillons, trop sensible sur de très grands). Le Q-Q plot montre *où* et *comment* les données s'écartent de la normalité : il complète le test.

b) Les points suivent bien la droite sur presque toute leur étendue. On observe de légers écarts aux deux extrémités (quelques résidus un peu plus grands, en valeur absolue, que prévu), ce qui est courant avec des données biologiques. Aucune courbure systématique ni valeur isolée : pas de déviation majeure.

c) Oui : le test de Shapiro–Wilk sur les résidus ne rejetait pas la normalité ($p = 0{,}078$), et le Q-Q plot le confirme visuellement. La p-valeur, proche de $0{,}05$ sans l'atteindre, s'explique par les petits écarts aux extrémités.

9. Résultat de l'ANOVA.

a) La statistique vaut $F = \dfrac{8{,}588}{0{,}827} \approx 10{,}39$, avec 5 et 114 degrés de liberté, et $p = 3{,}15 \times 10^{-8}$, bien inférieure à $0{,}05$ (et même à $0{,}001$, d'où les trois étoiles). On **rejette** H_0 : l'effet de l'espèce hôte sur la longueur des œufs de coucou est **hautement significatif**. Il est extrêmement improbable que des différences de moyennes aussi marquées soient dues au seul hasard de l'échantillonnage.

On peut quantifier cet effet : la part de la variabilité totale de la longueur des œufs expliquée par l'espèce hôte vaut

$$R^2 = \frac{SCE_{\text{inter}}}{SCE_{\text{totale}}} = \frac{42{,}94}{42{,}94 + 94{,}25} \approx 0{,}31,$$

soit environ $31\,\%$: l'effet est net, mais une grande partie de la variabilité reste propre à chaque œuf.

b) Non. L'ANOVA conclut qu'*au moins deux* moyennes diffèrent, sans dire lesquelles. Pour le savoir, il faut comparer les moyennes deux à deux avec un test **post-hoc**.

10. Test post-hoc de Tukey.

a) Un test post-hoc compare toutes les paires de groupes *après* une ANOVA significative, en **contrôlant le risque global** d'erreur de première espèce (ici $5\,\%$ pour l'ensemble des 15 comparaisons, *family-wise confidence level*), ce que ne font pas 15 tests de Student successifs. Le test de Tukey convient lorsque toutes les comparaisons deux à deux intéressent et que les variances sont homogènes (vérifié à la question 6).

b) Une différence est significative quand la p-valeur ajustée est inférieure à $0{,}05$, c'est-à-dire quand l'intervalle $[\,\texttt{lwr}\ ;\ \texttt{upr}\,]$ ne contient pas 0. Sept paires sur quinze sont significatives :

Comparaison	Différence (mm)	p ajustée	Conclusion
Wren – Hedge Sparrow	$-1{,}99$	$< 10^{-6}$	Wren < Hedge Sparrow
Wren – Tree Pipit	$-1{,}96$	$< 10^{-6}$	Wren < Tree Pipit
Wren – Pied Wagtail	$-1{,}77$	7×10^{-6}	Wren < Pied Wagtail
Wren – Robin	$-1{,}45$	3×10^{-4}	Wren < Robin
Wren – Meadow Pipit	$-1{,}17$	5×10^{-4}	Wren < Meadow Pipit
Meadow Pipit – Hedge Sparrow	$-0{,}82$	$0{,}043$	Meadow Pipit < Hedge Sparrow
Tree Pipit – Meadow Pipit	$+0{,}79$	$0{,}047$	Meadow Pipit < Tree Pipit

Les huit autres comparaisons ne sont pas significatives (p ajustées entre 0,23 et 1). Les deux différences impliquant Meadow Pipit sont tout juste significatives ($p \approx 0{,}04$ et la borne de l'intervalle est très proche de 0) : elles sont à interpréter avec prudence.

c) On résume classiquement ces résultats par des **lettres** : deux espèces qui partagent une lettre ne diffèrent pas significativement.

Espèce hôte	Moyenne (mm)	Groupe
Hedge Sparrow	23,1	a
Tree Pipit	23,1	a
Pied Wagtail	22,9	ab
Robin	22,6	ab
Meadow Pipit	22,3	b
Wren	21,1	c

- Les œufs de coucou pondus dans les nids de **Wren** (troglodyte) sont significativement plus petits que tous les autres, d'environ 1,2 à 2 mm. Le troglodyte est un très petit oiseau, dont les propres œufs sont petits : c'est cohérent avec l'hypothèse d'une adaptation de la taille de l'œuf à l'hôte.
- Les œufs pondus chez Meadow Pipit sont un peu plus petits que ceux pondus chez Hedge Sparrow et Tree Pipit (différences d'environ 0,8 mm, à la limite de la significativité).
- Pied Wagtail et Robin ne se distinguent significativement d'aucune espèce sauf Wren.

Conclusion : la longueur des œufs de coucou dépend de l'espèce hôte, ce qui va dans le sens d'une adaptation évolutive à l'hôte, comme pour la couleur. Une étude observationnelle comme celle-ci montre une association, pas le mécanisme : on ne sait pas, par exemple, si ce sont des lignées de coucous différentes qui parasitent chaque espèce hôte.

Correction de l'exercice 2 ▲

1. Questions préliminaires.

a) Le tableau compte 300 lignes et 13 colonnes : une colonne qualitative (`Gene`, le nom du gène) et 12 colonnes quantitatives (les conditions). Les **individus** statistiques sont les **300 gènes** ; ils sont décrits par **12 variables quantitatives**, les conditions biologiques, qui croisent :

- le *jour de développement* de l'embryon (jours 3 à 7) ;
- à partir du jour 5, le *type cellulaire* : épiblaste (EPI), à l'origine de l'embryon proprement dit ; endoderme primitif (PE) ; trophectoderme (TE), à l'origine du placenta.

b) Les 12 variables quantitatives sont les colonnes numériques `Day_3`, `Day_4`, `Pre_Day_5`, `Day_5_EPI`, `Day_5_PE`, `Day_5_TE`, `Day_6_EPI`, `Day_6_PE`, `Day_6_TE`, `Day_7_EPI`, `Day_7_PE` et `Day_7_TE`. Chaque valeur est un niveau d'expression d'un gène dans une condition : une mesure numérique continue, sur laquelle moyennes, variances et corrélations ont un sens.

c) L'ACP est une méthode *algébrique* : elle calcule des moyennes, des variances, des corrélations, puis diagonalise la matrice des corrélations. Ces calculs n'ont de sens que pour des variables quantitatives ; une colonne de texte comme `Gene` les rendrait impossibles (`PCA` renverrait une erreur). C'est le rôle de `sapply(data, is.numeric)` : ne garder que les colonnes numériques. Une variable qualitative peut en revanche servir à *interpréter* l'ACP, comme variable supplémentaire, ou servir de nom aux individus.

d) L'ACP résume l'information contenue dans un grand nombre de variables corrélées par un petit nombre de nouvelles variables non corrélées, les **composantes principales**, qui conservent le maximum de variance. Elle permet de réduire la dimension, de visualiser les données sur un plan, et de repérer les structures dominantes : groupes de variables liées, oppositions, individus semblables ou atypiques.

Elle est bien adaptée ici : 12 conditions, très probablement corrélées entre elles puisqu'elles relèvent du même processus de développement, et une démarche *exploratoire* (on ne teste pas une hypothèse précise). On attend de l'ACP qu'elle montre si les conditions se regroupent par *jour* de développement ou par *type cellulaire*, et quelles conditions s'opposent.

2. Matrice des corrélations.

a) Les coefficients étant affichés en pourcentage, une valeur proche de 100 indique une corrélation forte, une valeur proche de 0 une corrélation faible.

Conditions fortement corrélées	**Interprétation**
`Day_6_EPI` et `Day_7_EPI`	même type cellulaire (EPI), jours consécutifs
`Day_5_TE` et `Day_6_TE`	même type cellulaire (TE), jours consécutifs
`Day_5_PE` et `Day_6_PE`	même type cellulaire (PE), jours consécutifs
Conditions faiblement corrélées	
`Day_3` et `Day_7_TE`	jours éloignés et types cellulaires différents
`Day_3` et `Day_5_PE`	profils biologiques distincts
EPI et TE ou PE, à un même jour	types cellulaires différenciés

Biologiquement : une fois engagée dans un lignage, une cellule poursuit son programme de développement, et les gènes actifs dans un type cellulaire le restent à des niveaux comparables les jours suivants. À l'inverse, entre le jour 3 (cellules encore indifférenciées) et le trophectoderme du jour 7, le programme d'expression a profondément changé.

b) Des variables fortement corrélées apportent une information en grande partie *redondante* : c'est précisément la situation où l'ACP est efficace, car elle les résume par un même axe.

Observation	**Conséquence pour l'ACP**
corrélations fortes entre conditions proches	résumées par un même axe principal
corrélations faibles entre types ou jours éloignés	différences portées par des axes secondaires
blocs de corrélations (`order = "hclust"`)	groupes de conditions à retrouver sur le cercle

On s'attend donc à ce qu'un petit nombre d'axes suffise à résumer les 12 conditions.

3. Variance expliquée.

a) Chaque composante principale est une combinaison linéaire des variables de départ. La première est construite pour que la variance des individus projetés soit maximale ; la deuxième maximise la variance restante en étant non corrélée (orthogonale) à la première ; et ainsi de suite. La variance de la composante s est la valeur propre λ_s, et

$$\%\text{ de variance expliquée par l'axe } s = 100 \times \frac{\lambda_s}{\sum_t \lambda_t} = 100 \times \frac{\lambda_s}{12},$$

car en ACP normée la somme des valeurs propres vaut le nombre de variables. C'est la part de l'information totale (les différences d'expression entre gènes, sur les 12 conditions) que conserve l'axe. Par exemple, l'axe 1 restitue à lui seul $100 \times 7{,}586/12 \approx 63{,}2\,\%$ de cette information.

b) On retient **deux axes**, car trois critères concordent :

- *Kaiser* : seuls $\lambda_1 = 7{,}59$ et $\lambda_2 = 2{,}22$ dépassent 1, c'est-à-dire résument plus d'information qu'une variable de départ ($\lambda_3 = 0{,}87 < 1$) ;
- *coude* : l'éboulis chute brutalement jusqu'à l'axe 2 puis devient presque plat ; les axes 3 et 4 (7,3 et 6,5 %) apportent peu et sont de hauteur voisine ;
- *variance cumulée* : les deux premiers axes dépassent le seuil usuel de 80 %.

c) $63{,}22 + 18{,}48 \approx \mathbf{81{,}7}\,\%$. Plus de 80 % de l'information est conservée sur le plan $(1, 2)$: il donne une image fidèle de la structure des données, ce qui justifie de représenter variables et individus sur ce plan.

4. Cercle des corrélations. Les coordonnées d'une variable sont ses corrélations avec les deux premiers axes ; la couleur indique sa qualité de représentation sur le plan, $\cos^2 = \cos_1^2 + \cos_2^2$.

a) Deux variables bien représentées sont *fortement corrélées* si leurs flèches sont proches (angle voisin de $0°$), *faiblement corrélées* si elles sont à peu près perpendiculaires (angle voisin de $90°$).

- Fortement corrélées : `Day_6_EPI` et `Day_7_EPI`, ou `Day_5_TE` et `Day_6_TE` (même type cellulaire, jours consécutifs).
- Faiblement corrélées : `Day_3` et `Day_7_TE`, ou `Day_4` et `Day_6_PE` (flèches presque orthogonales).

On retrouve les liaisons lues dans la matrice des corrélations.

b)
- La **longueur** d'une flèche mesure la qualité de représentation de la variable sur le plan : une flèche qui atteint presque le cercle est bien représentée et s'interprète sans risque ; une flèche courte (comme celle de `Day_7_PE`) a l'essentiel de sa variance sur d'autres axes (3, 4...) et ne doit pas être interprétée sur ce plan.
- L'**orientation** indique à quels axes la variable est liée (une flèche horizontale est liée à l'axe 1, une flèche verticale à l'axe 2) et, par l'angle entre deux flèches, la corrélation entre deux variables : angle proche de $0°$, corrélation positive forte ; proche de $90°$, corrélation faible ; proche de $180°$, corrélation négative forte.

c) `Day_3` et `Day_4` ont des flèches longues et orientées verticalement : ce sont elles qui sont le plus fortement corrélées à l'axe 2, et qui contribuent le plus à sa construction. L'axe 2 oppose donc les gènes exprimés aux stades précoces, avant la différenciation des lignages, aux autres.

5. Représentation des individus.

- **Proximité** : deux gènes proches sur le plan ont des profils d'expression semblables sur les 12 conditions ; deux gènes éloignés ont des profils différents. Cette lecture n'est fiable que pour des gènes bien représentés ($\cos^2$ élevé) : un gène mal représenté peut paraître proche d'un autre alors que leurs profils diffèrent sur les axes non représentés. C'est pourquoi le biplot ne garde que les gènes et les conditions dont le $\cos^2$ dépasse $0{,}5$.
- **Signification de l'axe 1** : les conditions étant presque toutes corrélées positivement entre elles, l'axe 1 traduit un *effet taille* : il classe les gènes selon leur niveau global d'expression. Les premiers gènes du tableau, peu exprimés, ont d'ailleurs des coordonnées négatives sur l'axe 1, la plus négative étant celle de SEPT6 ($-0{,}886$), presque pas exprimé dans aucune condition.
- **Groupes** : la majorité des gènes forment un nuage dense près de l'origine. Comme beaucoup d'entre eux ont un $\cos^2$ supérieur à $0{,}5$, cette position centrale n'est pas un artefact de projection : ces gènes ont réellement un profil proche du profil moyen, avec une expression modérée et sans spécificité marquée pour un jour ou un type cellulaire.
- **Gènes atypiques** : les gènes éloignés du centre ont une expression forte ou spécifique dans certaines conditions. Du côté positif de l'axe 1, ce sont des gènes fortement exprimés dans de nombreuses conditions ; dans la direction d'une flèche (par exemple vers `Day_3` et `Day_4` sur l'axe 2), ce sont des gènes caractéristiques de ces conditions. Ce sont des candidats intéressants comme *marqueurs* d'un stade ou d'un lignage, mais ils peuvent aussi refléter des variations expérimentales : il faut revenir aux données brutes pour les examiner.

En conclusion, la carte des gènes, lue avec le cercle des corrélations et contrôlée par le $\cos^2$, résume la structure des profils d'expression : un grand nombre de gènes au profil moyen, et quelques gènes spécifiques d'un stade ou d'un type cellulaire, qui portent l'essentiel des différences entre conditions.
