Analyses statistiques des données acquises lors d'analyses fonctionnelles chez la souris

Laura Neuhart

Laura Neuhart

Master 2 Mathématiques et Applications

Spécialité Statistique

Analyses statistiques des données acquises lors d'analyses fonctionnelles chez la souris.

Du 01 février 2012 au 31 juillet 2012

Laboratoire d'accueil : ICS Institut Clinique de la Souris – Illkirch

Maitre de stage : M. Yann Herault & Mme Sophie LEBLANC

Responsable Universitaire : Mme Armelle GUILLOU

Remerciements

Avant de débuter ce mémoire, je tiens à remercier Dr. Yann HERAULT, Directeur de l'Institut Clinique de la Souris, pour m'avoir permis d'effectuer mon stage dans son institut et surtout de m'avoir fait confiance pour ce projet. Merci pour cette belle opportunité et surtout pour la disponibilité dont il a fait preuve.

Un grand merci à Sophie LEBLANC, pour m'avoir accompagnée, guidée et conseillée tout au long de ces 6 mois. Je lui suis reconnaissante de l'autonomie qu'elle m'a accordée pour mener à bien cette étude.

Je remercie également tout le service « Bioinformatique » de m'avoir accueillit au sein de leur équipe.

J'adresse encore mes remerciements à toutes les personnes que j'ai pu rencontrer lors de ce stage.

Pour finir, merci à tous les enseignants qui m'ont permis d'acquérir un grand nombre de connaissances durant toutes ces années.

Glossaire

ACP : **A**nalyse en **C**omposantes **P**rincipales

ANOVA: (**AN**alysis **O**f **VA**riance) Analyse de la variance

CAH : **C**lassification **A**scendante **H**iérarchique

Cohorte : Ensemble d'individus.

EUMODIC: The **EU**ropean **MO**use **DI**sease **C**linic

FDR (**F**alse **D**iscovery **R**ate) : Taux de fausses découvertes.

Gène : Unité de transmission héréditaire de l'information génétique. Un gène est un segment d'ADN (ou d'ARN chez virus), situé à un locus précis sur un chromosome, qui comprend la séquence d'acide désoxyribonucléique (ADN) codant pour une protéine, et les séquences qui en permettent et régulent l'expression.

ICS : **I**nstitut **C**linique de la **S**ouris

Phénotype : Ensemble des caractères observables d'un individu. Le phénotype correspond à la réalisation du génotype (expression des gènes) mais aussi des effets du milieu, de l'environnement.

Phénotypage : Analyse des phénotypes.

Pipeline : Désigne l'ensemble des tests fonctionnels réalisés sur une cohorte de souris.

Souris Mutante : souris pour laquelle un gène a subi une mutation.

Souris WT (**W**ild **T**ype = Type sauvage) : souris contrôle, souris pour laquelle le gène étudié n'a subi aucune mutation.

Remerciements _____ *- 3 -*

Glossaire _____ *- 4 -*

Introduction _____ *- 1 -*

Présentation de l'ICS _____ **- 1 -**

EUMODIC (The EUropean MOuse DIsease Clinic) _____ **- 1 -**

Description du jeu de données _____ **- 2 -**

Objectifs _____ **- 2 -**

Matériels et Méthodes _____ *- 3 -*

Le Logiciel R _____ **- 3 -**

Les Données _____ **- 3 -**

Les Méthodes Statistiques _____ **- 4 -**

I. Mesures de liaison entre deux caractères quantitatifs _____ - 4 -
II. Comparaison de groupes de données _____ - 6 -
III. Données manquantes : Imputation multiple _____ - 9 -
IV. Analyse en composantes principales _____ - 10 -
V. Classification non supervisée _____ - 13 -

Résultats _____ *- 16 -*

Etude des corrélations _____ **- 16 -**

Etude de l'effet du génotype : Détection de phénotype _____ **- 18 -**

I. Réduction du nombre de variables _____ - 18 -
II. Etude des phénotypes _____ - 25 -

Classification non supervisée des gènes _____ **- 30 -**

I. Etude du tableau complet _____ - 30 -
II. Etude de tableaux partiels _____ - 37 -

Conclusion _____ *- 40 -*

Discussions et Perspectives _____ *- 40 -*

Bibliographie _____ *- 41 -*

Annexes _____ *i*

Liste des gènes et leurs identifiants _____ **ii**

Description du déroulement des différents Pipelines _____ **iii**

ACP : Cercles de corrélations _____ **iv**

Introduction

Présentation de l'ICS

L'Institut Clinique de la Souris (ICS) est une infrastructure de recherche d'excellence pour la génomique fonctionnelle. Elle combine la capacité de générer des souris mutantes à grande échelle et d'effectuer une analyse phénotypique complète des animaux.
L'institut est composé de trois services interactifs, l'Ingénierie Génétique et la validation de modèle, le Phénotypage et la Zootechnie qui génèrent environ 200 lignées de souris génétiquement modifiées par an.
Ces différents services de l'ICS sont là afin d'aider les scientifiques à comprendre l'ensemble du génome humain grâce au modèle murin et ainsi développer une meilleure compréhension des maladies humaines.

EUMODIC (The EUropean MOuse DIsease Clinic)

Le programme EUMODIC réunit 18 instituts de recherche (dont l'ICS) de 8 pays européens qui sont experts dans le domaine de la génomique fonctionnelle et du phénotypage de souris.
EUMODIC a entreprit une évaluation phénotypique primaire de 500 lignées de souris mutantes issues de plusieurs souches.
Le phénotypage (analyse des caractères observables) concerne, dans ce programme, les principales fonctions biologiques de la souris telles que la fonction respiratoire, la fonction neurologique et comportementale, l'os et le cartilage, la chimie clinique, l'immunologie, les systèmes hormonaux et métaboliques, cardiovasculaires.
Cette évaluation se déroule dans 4 centres de phénotypage :

- HMGU en Allemagne
- ICS en France
- MRC Harwell au Royaume-Uni
- Sanger Institut au Royaume-Uni

Les données recueillies sont rendues publiques sur le site EuroPhenome[1] :

http://www.europhenome.org/

[1] Mallon, A-M., Blake, A., Hancock, J. M. (2008). EuroPhenome and EMPReSS: online mouse phenotyping resource, *Nucleid Acids Research*. **36,** D715-D718.

Description du jeu de données

Les 20 tests sont séparés en 2 groupes distincts ou pipelines. Pour chaque lignée mutante, 7 souris males et 7 souris femelles suivent les tests du pipeline 1 et, 7 souris males et 7 souris femelles suivent les tests du pipeline 2. Lors des tests sont également mesurées des souris contrôles (non mutées) mâles et femelles. La stratégie de passage des souris contrôles diffère d'un centre à l'autre. L'ICS a choisi de passer en même temps un groupe contrôle pour 2 lignées mutantes. Pour des raisons techniques, les souris d'une même lignée mutante sont parfois passées en plusieurs fois.

Objectifs

L'objectif global du stage est de réaliser des analyses statistiques sur l'ensemble du jeu de données obtenu à l'ICS dans le cadre de ce projet depuis 2009, afin d'identifier les gènes impliqués dans les différents processus ou voies biologiques.

Dans un premier temps, les analyses consistent à étudier les liens entre les variables sur les populations de souris mutantes et contrôles.
Une étude de l'effet des différents phénotypes est ensuite effectuée.
Pour finir, le but est de classer les gènes étudiés et de déterminer des familles ou profiles de gènes.

Matériels et Méthodes

Le Logiciel R

En référence à ses deux auteurs, Ross Ihaka et Robert Gentleman, le logiciel nommé de la lettre ***R*** renvoie également à son équivalent payant le logiciel S.
Ce logiciel possède son propre langage de programmation et de part son caractère libre, il est devenu une référence dans le monde statistique. Le logiciel R fournit à l'utilisateur les procédures usuelles afin d'explorer les données grâce à des facilités graphiques performantes et mener à bien une étude statistique. Sa simplicité d'utilisation permet la programmation rapide d'algorithmes complexes.

Les Données

Dans un premier temps, le projet se concentre sur les données récoltées par un seul centre de phénotypage : l'Institut Clinique de la Souris. Il sera possible d'étendre l'étude aux autres centres par la suite.

Les données sont ensuite séparées selon le paramètre « *Sexe* », du faite de la non homogénéité des résultats des expériences sur ces deux groupes. L'étude se fait en priorité sur les données des souris de sexe Male, de souche C57Bl/6N Tac.

Pour des raisons biologiques, toutes les mesures ne peuvent pas être effectuées sur une seule souris. Les différents paramètres étudiés sont donc séparés en deux Pipelines distincts. Une souris ne réalisera que l'un de ces deux Pipelines. Cela entraine donc que les résultats du premier pipeline ne pourront pas être comparés à ceux du deuxième lors de l'étude des corrélations.

Les Méthodes Statistiques

I. Mesures de liaison entre deux caractères quantitatifs

Soit *(X,Y)* un couple de caractères quantitatifs.

1. Covariance

La covariance de ce couple est défini par :

$$Cov(X,Y) = \mu(XY) - \mu(X)\mu(Y)$$

Avec $\mu(X), \mu(Y)$ les moyennes de X et Y, et $\mu(XY)$ la moyenne de XY :

$\mu(XY) = \sum_{i,j} x_i y_j f_{ij}$ avec f_{ij} la fréquence du couple $(x_i\ , y_j)$

2. Coefficient de corrélation linéaire

Le coefficient de corrélation linéaire théorique s'écrit :

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Avec $\sigma(X), \sigma(Y)$les écarts-types de X et Y.

Le coefficient de corrélation linéaire observé est :

$$r(x,y) = \frac{Cov_n(x,y)}{s_n(x)s_n(y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x}_n)^2)(\sum_{i=1}^{n}(y_i - \bar{y}_n)^2)}}$$

Avec $\bar{x}_n$ et $\bar{y}_n$ les moyennes calculées dans l'échantillon et

$$Cov_n(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)$$

$s_n^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ Et $s_n^2(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$

3. Coefficient de corrélation partielle

Lors de précédentes études, il a été observé que le poids des souris avait une forte influence sur un grand nombre de paramètres. Il serait donc intéressant de prendre en compte l'effet de cette variable lors de l'analyse de corrélations entre paramètres.

Bien souvent, deux variables *X* et *Y* sont fortement corrélées avec une troisième que l'on nommera *Z*. Il est possible alors de conclure à une forte corrélation entre *X* et *Y* alors qu'à priori il n'y a aucune relation entre ces deux grandeurs mise à part leur liaison avec la troisième variable *Z*.

Dans ce cas, il est nécessaire d'introduire une mesure de liaison traduisant la relation entre *X* et *Y* en éliminant l'influence de *Z*. Il s'agit de la notion de corrélation partielle[2] :

$$r(x,y|z) = \frac{r(x,y) - r(x,z)r(y,z)}{\sqrt{1-r(x,z)^2}\sqrt{1-r(y,z)^2}}$$

Avec $r(x,y), r(x,z)$et $r(y,z)$ les coefficients de corrélations linéaires entre *X* et *Y, X et* Z, *Y* et *Z.*

4. Test de l'hypothèse $r(x,y|z) = 0$

Il est intéressant de tester la nullité du coefficient de corrélation partielle entre les variables *X* et *Y* connaissant l'influence de *Z.*

$$H_0 : r(x,y|z) = 0$$
Contre
$$H_1 : r(x,y|z) < ou > 0$$

Sous l'hypothèse nulle H_0 :

$$t(x,y|z) = \sqrt{n-2}\frac{r(x,y|z)}{\sqrt{1-r(x,y|z)^2}}$$

est une réalisation d'une variable T qui suit une loi de Student à *n-2* degrés de liberté.

Pour un seuil α donné, soit $t_{n-2,\alpha}$ le quantile de la loi de Student à n-2 degrés de liberté d'ordre α.

Nous décidons alors : $\begin{cases} |t(x,y|z)| > t_{n-2,\alpha}, on\ rejette\ H_0. \\ |t(x,y|z)| \leq t_{n-2,\alpha}, on\ accepte\ H_0. \end{cases}$

5. Test de l'hypothèse $r_1(x,y|z) = r_2(x,y|z)$

Le but de l'étude est de discerner des modifications de liens entre les paramètres après une mutation génétique. Afin de déterminer si deux coefficients de corrélations sont significativement différents ou non, le test suivant est utilisé :

$$H_0 : r_1(x,y|z) = r_2(x,y|z)$$
Contre
$$H_1 : r_1(x,y|z) \neq r_2(x,y|z)$$

La transformation de Fisher est utilisée lors de ce test :

$$Z_i = tanh^{-1}(r_i), i = 1,2$$

Sous l'hypothèse nulle H_0 :

$$Z = \frac{|Z_1 - Z_2|}{\sqrt{\frac{1}{(n_1-3)} + \frac{1}{(n_2-3)}}}$$

[2] Fisher, R.A. (1924). The distribution of the partial correlation coefficient. *Metron* 3 (3–4), 329–332.

Pour un seuil α donné, soit $z_{\alpha/2}$ le quantile de la loi normale centrée-réduite d'ordre $\frac{\alpha}{2}$.

Nous décidons alors : $\begin{cases} Z \geq z_{\alpha/2}, on\ rejette\ H_0. \\ Z < z_{\alpha/2}, on\ accepte\ H_0. \end{cases}$

II. Comparaison de groupes de données

1. Analyse de la Variance à un facteur

Proposé par Sir Ronald Fisher dans un article[3] datant de 1918, l'analyse de la variance est une procédure permettant la comparaison des espérances de variables aléatoires indépendantes gaussiennes, de même variance entre plusieurs populations.
Pour l'analyse de la variance à un facteur, deux variables sont observées :

- Une variable quantitative continue, appelée la réponse et notée Y
- Une variable qualitative, totalement contrôlée par l'expérimentateur, appelée le facteur.

Dans le cas déséquilibré, le modèle statistique est donc, le plus souvent, écrit de la façon suivante :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Y_{ij} suit une loi Normale (μ_i, σ) pour i= 1,…, I et j= 1,…, n_i

les ε_{ij} suivent une loi Normale (0, σ) pour i= 1,…, I et j= 1,…, n_i

les α_i vérifient dans notre cas la contrainte : $\alpha_1 = 0$

Les conditions fondamentales d'application de l'analyse de la variance sont donc que les variables erreurs ε_{ij_i} :

- soient indépendantes
- aient même variance inconnue σ²
- suivent une loi normale centrée

Le but est donc de déterminer si toutes les espérances μ_i sont égales. Avec la correspondance pour $i = 1, \dots, I$:

$$\mu_i = \mu + \alpha_i.$$

Test de comparaison de plusieurs espérances.

H_0: $\mu_1 = \mu_2 = \cdots = \mu_I$

Contre

H_1: *Les espérances* μ_i*ne sont pas toutes égales.*

Si les trois conditions fondamentales sont satisfaites et si l'hypothèse nulle est vraie, alors $F(obs) = \frac{CM_{fac}}{CM_{res}}$ est une réalisation de la variable aléatoire *F* qui suit la loi de Fisher $F(I-1, n-I)$.

[3] Fisher, R.A. (1918). The correlations between Relatives on the Supposition of Mendelian Inheritance, *Philosophical Transactions of the Royal Society of Edinburgh*, **52**, 399-433

Avec *n* l'effectif total, $CM_{fac} = \frac{SC_{fac}}{I-1} = \frac{\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{y}_n)^2}{I-1}$ et $CM_{res} = \frac{SC_{res}}{n-I} = \frac{\sum_{i=1}^{I}(\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2)}{n-I}$.
Pour un seuil α donné, les tables de la loi de Fisher nous fournissent une valeur critique c_α telle que $P_{H_0}(F \leq c_\alpha) = 1 - \alpha$.

Alors nous décidons : $\begin{cases} si\ F(obs) \geq c_\alpha\ on\ rejette\ H_0. \\ si\ F(obs) < c_\alpha\ on\ accepte\ H_0. \end{cases}$

Les estimateurs $\hat{\mu}, \widehat{\alpha_1}, \dots, \widehat{\alpha_I}, \widehat{\sigma^2}$ des paramètres $\mu, \alpha_1, \dots, \alpha_I, \sigma^2$ du modèle sont donnés par les formules suivantes :

$$\hat{\mu} = Y_{\cdot\cdot} = \bar{Y}$$

$$\widehat{\alpha_i} = Y_{i\cdot} - \hat{\mu}, 1 \leq i \leq I$$

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{I}(\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2)}{n - I}$$

Comparaisons multiples : Procédure de Benjamini-Hochberg

Quand lors de la comparaison de plusieurs espérances l'hypothèse nulle a été rejetée, il est intéressant d'analyser ces différences, et donc de comparer ces différences entres elles.
Pour cela, le $i^{ème}$ test est le suivant :

$H_0: \mu_1 = \mu_i$
Contre
$H_1: \mu_1 \neq \mu_i$

Récemment, Benjamini et Hochberg[4] ont proposés une nouvelle approche basée sur le taux de fausses découvertes (False Discovery Rate).

	Déclaré non significatif	Déclaré significatif	Total
Hypothèse nulle est vraie	**U**	**V**	m_0
Hypothèse nulle est fausse	**T**	**S**	$m - m_0$
Total	$m - R$	R	m

Tableau 1: Nombre d'erreurs commises lors de *m* tests

$$FDR = E\left(\frac{V}{R}\right)$$

Nous considérons *m* probabilités critiques ordonnées correspondants aux *m* tests :

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$$

Nous définissons ensuite :

$$k := \max\left\{i \mid p_{(i)} \leq \frac{i\alpha}{m}\right\}$$

[4] Benjamini, I., Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society*, **57**, 289-300

Nous décidons alors : $\begin{cases} si\ i \leq k\ , on\ rejette\ H_0\ pour\ le\ i^{ème}\ test. \\ si\ i > k\ , on\ accepte\ H_0\ pour\ le\ i^{ème}\ test. \end{cases}$

Dans le contexte exploratoire, le FDR, qui conduit à des procédures moins restrictives que d'autres méthodes, semble un critère bien adapté du fait du grand nombre de comparaisons à effectuer.

2. Analyse de la variance à un facteur par la méthode des rangs de Kruskal-Wallis

Lorsque les conditions fondamentales ne sont pas vérifiées, il est nécessaire d'utiliser un test non paramétrique.

Présenté en 1952 par Wiliam Kruskal et W. Allen Wallis[5], l'analyse de la variance à un facteur par la méthode des rangs de Kruskal-Wallis est un test non paramétrique utilisée pour vérifier l'appartenance de *k* échantillons indépendants à une même population.

Soit Y une variable aléatoire de loi continue observée sur une population divisée en $k \geq 3$ sous populations. Nous supposons ainsi que nous disposons de *k* échantillons aléatoires indépendants et de $k \geq 3$ séries d'observations. Notons $\mathcal{L}_i(Y)$ la loi de la variable aléatoire Y sur la sous-population d'ordre *i* avec $1 \leq i \leq k$.

Le test de Kruskal-Wallis est utilisé pour tester les hypothèses suivantes :

$$H_0 : \mathcal{L}_1(Y) = \mathcal{L}_2(Y) = \cdots = \mathcal{L}_k(Y)$$
Contre
$$H_1 : Les\ lois\ \mathcal{L}_1(Y), \ldots, \mathcal{L}_k(Y)\ ne\ sont\ pas\ toutes\ identiques.$$

Soient R_{ij} le rang de Y_{ij} parmi les *n* valeurs, $R_{i.} = \sum_{j=1}^{n_i} R_{ij}$ la somme des rangs associée à chaque échantillon, $\overline{R}_{i.} = \frac{R_{i.}}{n_i}$ la moyenne des rangs de chaque échantillon et la statistique de Kruskal-Wallis s'écrit :

$$KW_{n.} = \frac{12}{n.(n.+1)} \sum_{i=1}^{k} n_i \left(\overline{R}_{i.} - \frac{n.+1}{2} \right)^2 = \frac{12}{n.(n.+1)} \sum_{i=1}^{k} \frac{R_{i.}^2}{n_i} - 3(n.+1).$$

Pour un seuil α donné, les tables de la loi du Khi-deux nous fournissent une valeur critique c_α telle que $P_{H_0}(-c_\alpha < KW_{n.} < c_\alpha) = 1 - \alpha$.

Alors nous décidons : $\begin{cases} si\ KW_{n.}(obs) \geq c_\alpha\ on\ rejette\ H_0. \\ si\ KW_{n.}(obs) < c_\alpha\ on\ accepte\ H_0\ . \end{cases}$

[5] Kruskal, W., Wallis, W.A. (1952) Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association*, **47** (260): 583–621

Comparaisons multiples : Test de Steel-Dwass-Critchlow-Fligner

Le test de Steel-Dwass-Critchlow-Fligner[6] est utilisé afin de tester les hypothèses suivantes :

$$H_0 : \mathcal{L}_1(Y) = \mathcal{L}_i(Y)$$
Contre
$$H_1 : \mathcal{L}_1(Y) \neq \mathcal{L}_i(Y)$$

Les observations des deux sous-populations sont ordonnées.

Nous décidons qu'au seuil α l'hypothèse H_0 est rejetée si :

$$\bar{R}_{1.} - \bar{R}_{i.} \geq q(k; +\infty; 1-\alpha) \sqrt{\frac{n.(n.+1)}{12}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_i}}$$

Où $q(k; +\infty; 1-\alpha)$ est le quartile d'ordre (1-α) pour la loi de l'étendue studentisée pour k moyennes et $+\infty$ degrés de liberté.

III. Données manquantes : Imputation multiple

Toute collecte de données à grande échelle est nécessairement confrontée au problème des données manquantes.

Bien qu'un grand nombre de méthodes d'imputation simple (imputation par une constante, par une valeur aléatoire ou non) aient été développées, la méthode de l'imputation multiple, proposée et décrite par DB Rubin[7], conduit à des résultats généralement plus adéquats.

Visant à corriger la sous estimation de la variance caractéristique des imputations simples, l'imputation multiple est une méthode qui fait l'hypothèse que les valeurs manquantes sont aléatoires ou de type MAR (*missing at random*).
La condition MAR signifie que sachant les données observées, le mécanisme de non réponse ne dépend pas de données non observées.

Soit $m > 1$ le nombre de jeux de données complets.

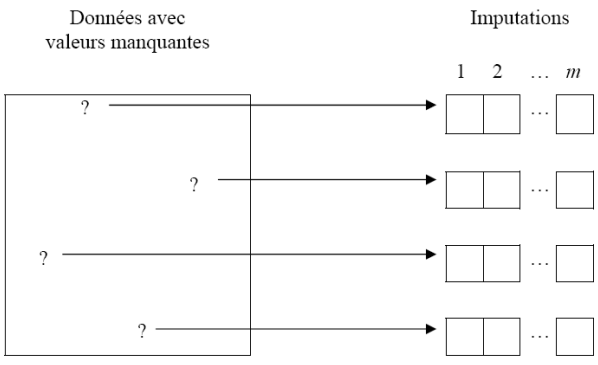


Figure 1: Principe de l'imputation multiple

[6] Steel, R. (1959) A multiple comparison rank sum test : treatment versus control, *Biometrics*, **15** (4), 560-572
[7] Rubin, D. B. (1988). An overview of multiple imputation, *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, 79-84

Après l'obtention de ces *m* ensembles de données, l'analyse se fait au moyen de n'importe quelle méthode appropriée aux données complètes.

Les résultats qui dépendent de l'imputation multiple sont généralement valides car elle prend en compte l'incertitude générée par les valeurs manquantes. Cette incertitude est habituellement sous estimée par les méthodes d'imputation simple.

Grâce à son haut niveau d'efficacité, peu d'imputations suffisent à l'obtention d'excellents résultats. Rubin[8] montre que moins de 5 imputations sont nécessaires.

IV. Analyse en composantes principales

L'**A**nalyse en **C**omposantes **P**rincipales (ACP) est une méthode descriptive multidimensionnelle formalisée par H. Hotteling[9] dans les années 1930. A partir d'un tableau de données à *p* variables quantitatives $(v_1, v_2, \dots, v_p)$ et *n* individus $(u_1, u_2, \dots, u_n)$, l'ACP propose des représentations graphiques de ces variables et de ces individus.

Soit ***X*** le tableau de données brut à partir duquel l'analyse va être effectuée :

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Chaque individu et chaque variable pourront être représentés dans l'espace lui correspondant.

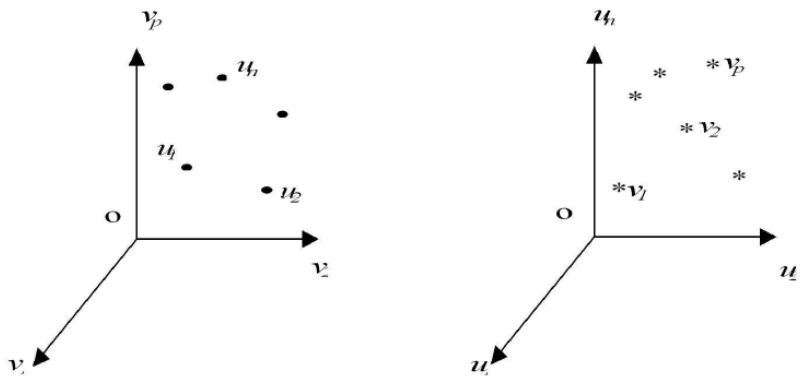


Figure 2: Nuage des individus/Nuage des variables

1. Nuage des individus

Il apparait judicieux de choisir pour origine, le centre de gravité du nuage des individus. En ACP, on choisit de donner le même poids à chaque individu, soit $p_i = \frac{1}{n}$. Le centre de gravité *G* du nuage est alors le point dont les coordonnées sont les valeurs moyennes des variables.

[8] Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, Wiley, New-York.
[9] Hotteling, H. (1933). Analysis of a Complex of Statistical Variables with Principal Components, *Journal of Educational Psychology,* **24**(6), 417-441.

Et afin d'éviter tout problème d'hétérogénéité des variables, on cherchera à normaliser les variables et donc faire notre étude sur des variables sans dimensions.

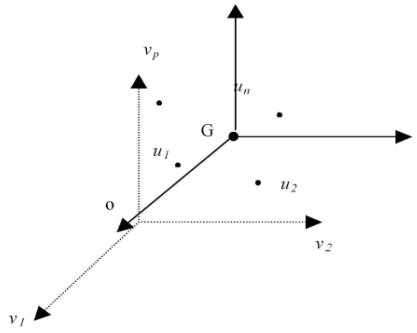


Figure 3 : Après centrage des données

Nous travaillons alors avec le tableau des données centrées –réduites :

$$\boldsymbol{X} = \begin{pmatrix} \frac{x_{11}-x_{.1}}{\sigma_1} & \cdots & \frac{x_{1p}-x_{.p}}{\sigma_p} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1}-x_{.1}}{\sigma_1} & \cdots & \frac{x_{np}-x_{.p}}{\sigma_p} \end{pmatrix}$$

2. Inertie du nuage

Soit I_G le moment d'inertie totale, il correspond à une mesure de la dispersion du nuage des individus :

$$I_G = \frac{1}{n}\sum_{i=1}^{n} d^2(G,u_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij}-x_{.j})^2 = \sum_{j=1}^{p}(\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-x_{.j})^2) = \sum_{j=1}^{p} Var(v_j)$$

I_G est donc égal à la trace de la matrice de variance-covariance des *p* variables v_j.

Soit I_Δ l'inertie du nuage des individus par rapport à un axe Δ passant par *G* et mesurant la proximité de l'axe Δ du nuage des individus :

$$I_\Delta = \frac{1}{n}\sum_{i=1}^{n} d^2(h_{\Delta i}, u_i)$$

Avec $h_{\Delta i}$ la projection orthogonale de u_i sur l'axe Δ.

Soit I_V l'inertie du nuage des individus par rapport à un sous espace vectoriel *V* passant par *G* :

$$I_V = \frac{1}{n}\sum_{i=1}^{n} d^2(h_{Vi}, u_i)$$

Avec h_{Vi} la projection orthogonale de u_i sur le sous espace *V*.

Le théorème de Huygens nous dit :

$$I_V + I_{V^*} = I_G$$

Avec *V** le complémentaire orthogonal de *V* dans $\mathbb{R}^p$.

Dans le cas où le sous espace est de dimension 1, I_{V^*} est dite « l'inertie expliquée par l'axe ».

3. Recherche des axes

En projetant le nuage sur le sous espace *V*, on perd l'inertie mesurée par I_V, on ne conserve alors que celle mesurée par I_{V^*}.

Le but est donc de trouver l'axe Δ_1 passant par G d'inertie minimum afin de disposer de son sous-espace vectoriel complémentaire Δ_{1^*} avec une inertie maximum et donc une dispersion du nuage qui soit maximale.

Soit $\overrightarrow{Ga_1}$ le vecteur directeur de l'axe Δ_1. Il faut donc trouver ce vecteur directeur tel que Δ_{1^*} soit maximum avec la contrainte $\left\|\overrightarrow{Ga_1}\right\| = 1$.
Par la symétrie du produit scalaire, on déduit que :

$$I_{\Delta_1^*} = {}^t a_1 \sum a_1$$

Avec $\sum$ la matrice de variance covariance empirique des *p* variables.

$$\text{Et } \left\|\overrightarrow{Ga_1}\right\| = {}^t a_1 a_1.$$

Grâce à la méthode des multiplicateurs de Lagrange, on trouve a_1 qui soit maximum avec la contrainte ${}^t a_1 a_1 = 1$.
Pour trouver a_1, il suffit de calculer les dérivées partielles de : $g(a_1) = {}^t a_1 \sum a_1 - \lambda_1 ({}^t a_1 a_1 - 1)$
En utilisant la dérivée matricielle, on obtient :

$$\frac{\partial g(a_1)}{\partial a_1} = 2\sum a_1 - 2\lambda_1 a_1 = 0$$

Le système à résoudre est donc :

$$\begin{cases} \sum a_1 - \lambda_1 a_1 = 0 \quad (1) \\ {}^t a_1 a_1 - 1 = 0 \qquad (2) \end{cases}$$

De (1) on n'en déduit que a_1 le vecteur propre de la matrice $\sum$ associé à la valeur propre λ_1.
Grâce à (2) on trouve :

$${}^t a_1 \sum a_1 = \lambda_1 \Leftrightarrow I_{\Delta_1^*} = \lambda_1.$$

Or on veut justement $I_{\Delta_1^*}$ maximum. Cela signifie donc que la valeur propre λ_1 est la plus grande valeur propre de la matrice $\sum$ et cette valeur propre est égale à l'inertie portée par l'axe Δ_1.

L'axe Δ_1 a donc comme vecteur directeur le premier vecteur propre associé à la plus grande valeur propre de la matrice de variance-covariance.

On cherche de la même façon les axes suivants.

Il reste à représenter les individus dans les plans définis par ces nouveaux axes.
Soit y_{ik} la coordonnée de u_i sur l'axe Δ_k:

$$y_{ik} = \langle \overrightarrow{Gu_i}, \overrightarrow{a_k} \rangle = {}^t a_k U_i$$

avec $U_i = \begin{bmatrix} \frac{x_{1i} - x_{.i}}{\sigma_i} \\ \frac{x_{2i} - x_{.i}}{\sigma_i} \\ \\ \frac{x_{ni} - x_{.i}}{\sigma_i} \end{bmatrix}$.

4. Contribution des axes

La contribution absolue de l'axe Δ_k à l'inertie totale du nuage des individus est égale à :

$$c_{abs}(\Delta_k|I_G) = \lambda_k.$$

Et sa contribution relative est donc :

$$c_{rel}(\Delta_k|I_G) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Ces pourcentages d'inertie indiquent la part de variabilité du nuage des individus expliquée par ces axes. On pourra donc réduire notre analyse aux $d < p$ premiers axes qui expliquent un pourcentage d'inertie proche de 1.

V. Classification non supervisée

Afin de découvrir une structure dans les données il est intéressant d'effectuer une classification de celles-ci. Ne disposant d'aucune information a priori sur les données à traiter, une classification non supervisée est la méthode la plus appropriée.

Deux sortes principales de classification de ce type sont connues :

- Classification hiérarchique
- Classification à partitionnement

Le but de ces classifications est de trouver des classes homogènes et relativement bien séparées des autres classes. Par un calcul élémentaire on peut voir que le nombre de classes possibles d'un ensemble de *n* éléments croit plus qu'exponentiellement avec *n*. Il est donc impensable de tester toutes les classes possibles.
Les différentes méthodes se limitent à l'exécution d'un algorithme itératif convergeant vers une partition relativement correcte.

Plusieurs caractéristiques de la classification sont laissées au choix de l'utilisateur :

- Mesure d'éloignement
- Critère d'homogénéité des classes
- Méthode hiérarchique ou à partitionnement
- Nombre de classes

Mesures d'éloignement

Une dissimilarité *d* est une application de $E \times E$ dans $\mathbb{R}^+$ telle que :

- $d(i,i) = 0 \ \forall i \in E$
- $d(i,i') = d(i',i) \ \forall i, i' \in E \times E$

Une distance satisfait les propriétés d'une dissimilarité. La matrice de similarité est donc sous la forme :

$$\begin{pmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & 0 & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{pmatrix}$$

Soit y_{jk} la coordonnée de l'individu u_j sur l'axe Δ_k.

Différentes distances peuvent être définies :

- Distance Euclidienne :

$$d(u_m, u_n) = \sqrt{\sum_{k=1}^{d} (y_{mk} - y_{nk})^2}$$

- Distance de Manhattan (ou « City-block ») :

$$d(u_m, u_n) = \sum_{k=1}^{d} |y_{mk} - y_{nk}|$$

1. Classification ascendante hiérarchique

Algorithme :

Initialisation :

Les classes initiales seront les singletons.

Pour $i = 1$ à n :

Regroupement des deux classes les plus proches au sens de la distance préalablement choisie.

Les critères d'agrégation :

- Lien minimum : la distance entre deux classes est la plus petite distance entre un individu de la première classe et un de la seconde.

$$D(C_1, C_2) = \min\left(d(u_{i1}, u_{j2}), u_{i1} \in C_1 et\ u_{j2} \in C_2\right)$$

- Lien maximum : la distance entre deux classes est la plus grande distance entre un individu de la première classe et un de la seconde.

$$D(C_1, C_2) = \max\left(d(u_{i1}, u_{j2}), u_{i1} \in C_1 et\ u_{j2} \in C_2\right)$$

- Critère de Ward : (uniquement si l'on est muni d'un espace euclidien) le regroupement des classes est choisi de telle sorte que l'augmentation de l'inertie intra classe soit minimale.

$$D(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_{C_1} + n_{C_2}} d^2(g_{c_1}, g_{c_2})$$

Avec g_{c_i} le centre de gravité de la classe *i*.

La classification ascendante hiérarchique se présente comme la succession de partitions emboitées et peut être représentée graphiquement à l'aide d'un dendrogramme.

La méthode descendante n'est que très rarement utilisée en pratique.

2. Classification à partitionnement

Contrairement aux méthodes hiérarchiques qui construisent les classes progressivement, ici les partitions de l'ensemble des individus sont créées directement et sont améliorées par la suite.

Algorithme :

Initialisation :

Sélectionner *k* points dans l'espace des individus appelés « centres ».

Jusqu'à ce que le critère de variance interclasse ne croit plus significativement :

- Assigner chaque individu au centre de la classe le plus proche au sens de la distance choisie précédemment.
- Recalculer le centre de gravité de chaque classe.

La méthode K-means

Soit $(u_1, u_2, \dots, u_n)$ un ensemble d'individus où chaque individu est un vecteur de dimension *d*. La méthode K-means à comme but de partitionner ces individus en $K < n$ ensembles $\boldsymbol{S} = \{S_1, S_2, \dots, S_K\}$ afin de minimiser la somme des carrées à l'intérieur de ces classes :

$$\underset{\boldsymbol{S}}{argmin} \sum_{i=1}^{k} \sum_{u_j \in S_i} \|u_i - \mu_i\|^2$$

Où μ_i la moyenne des points de l'ensemble S_i.

La méthode K-medoids

La méthode K-medoids peut être une alternative à la méthode K-means qui résoudra le problème de sensibilité aux données aberrantes lors de l'initialisation de cette dernière.

Dans cette nouvelle méthode, les classes sont représentées par un de leurs membres, le plus central de la classe, appelé « *medoids* ».

De la même façon que pour la méthode K-means, l'algorithme des K-medoids minimise l'erreur quadratique moyenne.

L'indice des silhouettes

L'indice des silhouettes a été défini par Rousseeuw[10] en 1987, il permet de mesurer pour tout individu X_i, s'il est plus ou moins bien classé :

$$\forall X_i, i = 1, \dots, I \; s(X_i) = \frac{b(X_i) - a(X_i)}{\max\left(a(X_i), b(X_i)\right)}$$

Avec $a(X_i)$ la dissimilarité moyenne entre l'individu X_i et tous les autres individus de la classe à laquelle il appartient. Et $b(X_i)$ le minimum des dissimilarités moyennes entre l'individu X_i et tous les autres individus des autres classes.

Plus $s(X_i)$ est proche de 1, mieux l'individu X_i est classé. Si $s(X_i)$ est proche de 0, il est situé entre deux classes. Enfin si $s(X_i)$ est proche de -1, l'individu est mal classé.

[10] Rousseeuw, P. J., (1987). Silhouettes : A graphical aid to the interpretation and validation of clusters analysis, *Journal of Computational and Applied Mathematics*, **20** (53-65).

Résultats

Etude des corrélations

Lors de précédentes études, il a été observé que le poids des souris avait une forte influence sur un grand nombre de paramètres. De ce fait, nous effectuerons les corrélations linéaires partielles en fonction du poids de l'animal au moment de l'expérience.

Sachant que la souris peut avoir entre 9 et 16 semaines lors de la mesure, les données conditionnelles ne seront donc pas les mêmes suivant les 2 paramètres étudiés. Le tableau de corrélation obtenu ne sera donc pas totalement symétrique du fait de la variable conditionnelle qui sera le poids de l'animal au moment de l'expérience placée en ligne.

Dans un premier temps, regardons le tableau de corrélations partielles pour les souris contrôles :

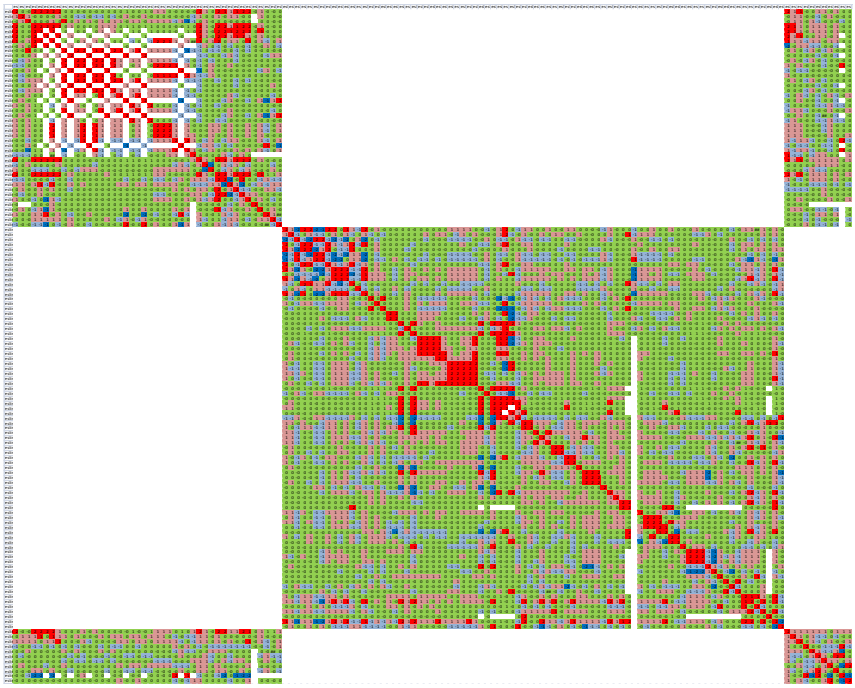


Figure 4 : Matrice de corrélation, souris contrôles.

Les cases blanches représentent les coefficients n'ayant pu être mesurés, entre un paramètre du Pipeline 1 et un paramètre du Pipeline 2 par exemple, ou à cause de données manquantes trop nombreuses.

En rouge et bleu foncé les coefficients respectivement significativement supérieurs et inférieurs à 0 et dont la valeur absolue est au moins supérieure à 0.4. En rose et bleu pâle, s'ils sont significativement différents de 0, mais leur valeur absolue est inférieure à 0.4. En vert, les coefficients n'étant pas significativement supérieurs à 0.

Comparons celle-ci à la matrice de corrélations partielles de l'une des lignées mutantes :

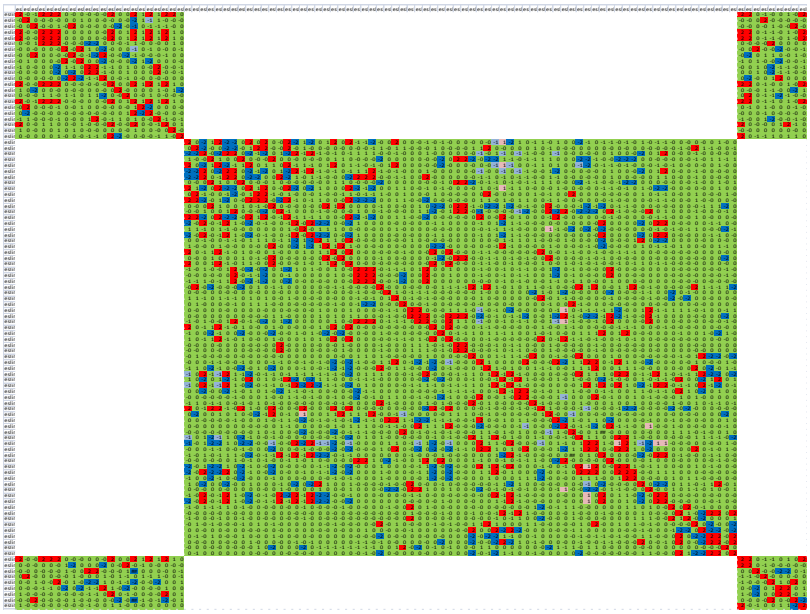


Figure 5 : Matrice de Corrélation, souris de la lignée Aste1

Si l'on compare chaque coefficient de corrélations partielles du groupe contrôle par rapport à celui correspondant chez le groupe mutant on obtient la matrice suivante :

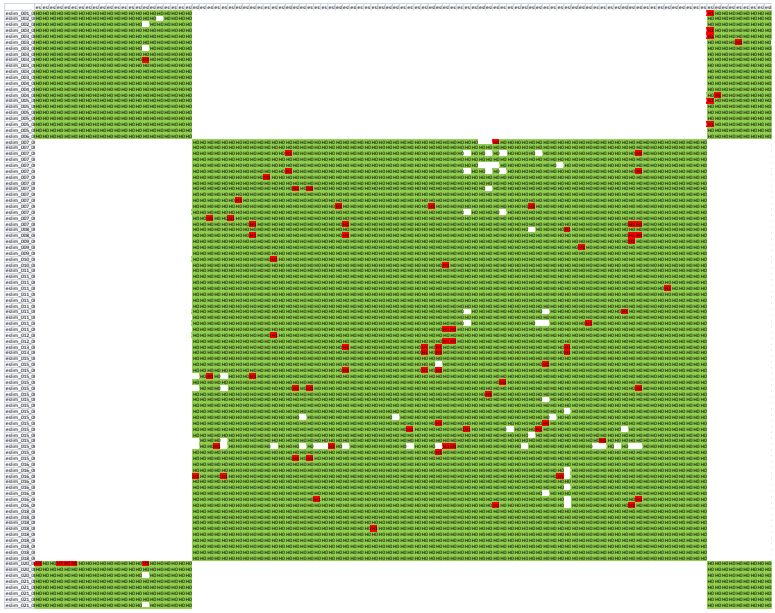


Figure 6 : Comparaisons de deux coefficients de corrélation, contrôle/Aste1

Ces comparaisons étant effectuées pour tous les groupes mutants, un ajustement de la p-valeur au moyen de la méthode de Benjamini-Hochberg est mis en place. Au final peu de différences significatives (en rouge) entre les coefficients du groupe contrôle et celui du groupe mutant sont détectées.

Conclusion :

Il est difficile d'interpréter de tels résultats, il n'y a pas de « groupe » de paramètres qui semblent avoir des résultats bien différents chez la lignée mutante. Il en est de même pour les autres lignées, nous passerons donc directement à la partie « Etude de l'effet du génotype : Détection de phénotypes » en espérant obtenir des résultats plus faciles d'interprétations. Pour cela, il semble nécessaire de réduire le nombre de paramètres.

Etude de l'effet du génotype : Détection de phénotype

I. Réduction du nombre de variables

Avec plus d'une centaine de variables à étudier, il s'avère nécessaire de trouver une méthode permettant la réduction de ce nombre sans toute fois perdre l'information contenue dans l'ensemble des données.

Dans le contexte de notre étude, la méthode qui semble la plus appropriée est l'Analyse en Composantes Principales.

Avant d'appliquer la méthode aux données, les différents paramètres seront regroupés. On devrait retrouver au sein de ces groupes des relations de corrélations plus ou moins fortes. Comme il n'y a pas de regroupement « parfait », deux sortes seront testées. La première méthode consiste à regrouper les paramètres en fonction du test biologique dans lequel ils sont mesurés. Dans la seconde méthode de regroupement, les différents paramètres seront regroupés selon la fonction biologique qu'ils sont censés expliquer.

Le poids étant un paramètre particulier et mesuré régulièrement, il sera considéré comme une fonction biologique à lui tout seul.

1ère façon : Les paramètres regroupés par test biologique

Fonctions Biologiques	Nombre de paramètres Pipeline 1	Nombre de paramètres Pipeline 2
Poids	6	4
Biochimie Sanguine	6	18
Métabolisme énergétique	16	0
Anxiété	0	14
Schizophrénie	0	10
Hématologie	0	8
Immunologie	0	15
Capacités motrices	0	3
Sensibilité à la douleur	0	1
Métabolisme osseux	2	0
Homéostasie du glucose	2	0
Histologie	2	0
Cardiovasculaire	2	0
Santé générale	0	1

Figure 7: Regroupement des variables selon les fonctions biologiques

La méthode retenue pour ses conclusions de meilleure qualité étant la seconde, nous y décrirons plus en détails les résultats obtenus.

2ème façon : Les paramètres regroupés par fonction biologique

Fonctions Biologiques	Nombre de paramètres Pipeline 1	Nombre de paramètres Pipeline 2
Poids	6	4
audition	0	6
Métabolisme énergétique	22	2
Anxiété	0	3
Schizophrénie	0	10
Hématologie	0	7
Immunologie	0	16
Capacités motrices	5	16
Métabolisme osseux	2	3
Homéostasie du glucose	3	1
Métabolisme hépatique	0	5
Cardiovasculaire	3	2
Métabolisme rénal	0	9

Figure 8: Regroupement des variables selon les fonctions biologiques

Les fonctions biologiques contenant moins de 3 variables seront laissées telles quelles. Pour les autres une ACP est effectuée.
Afin de garder le maximum d'information, les variables retenues après l'analyse devront contenir un minimum de 75 % du pourcentage cumulé d'explication de la variance.

1. Poids

Pour la fonction biologique représentant le poids, les deux composantes retenues sont des combinaisons linaires des différentes mesures du poids lors du Pipeline 1 et du Pipeline 2.

2. Audition

6 paramètres mesurent l'audition de l'animal grâce à un test de stimulation acoustique, le PPI (Pre-Pulse Inhibition). Ces paramètres sont fortement corrélés.

audition	PPI_BN.65.	PPI_P70	PPI_P80	PPI_P85	PPI_P90	PPI_ST
PPI_BN.65.	1	0,912	0,712	0,347	0,141	-0,05
PPI_P70	0,912	1	0,736	0,377	0,156	-0,06
PPI_P80	0,712	0,736	1	0,47	0,262	-0,05
PPI_P85	0,347	0,377	0,47	1	0,716	0,182
PPI_P90	0,141	0,156	0,262	0,716	1	0,347
PPI_ST	-0,05	-0,06	-0,05	0,182	0,347	1

Figure 9: Matrice de corrélations des paramètres expliquant l'audition.

Avec un pourcentage cumulé d'explication de la variance de plus de 76.6%, deux composantes principales sont sélectionnées après l'application de l'ACP aux données :

- La première explique les paramètres mesurant l'audition lors de stimulations à faibles décibels
- La seconde à forts décibels.

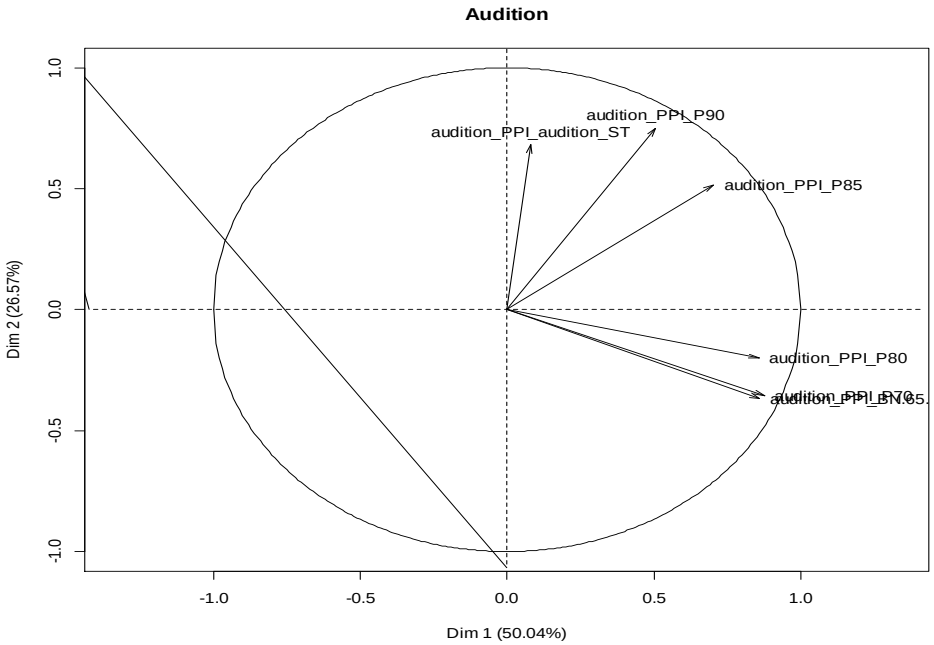


Figure 10: Cercle des Corrélations pour l'audition

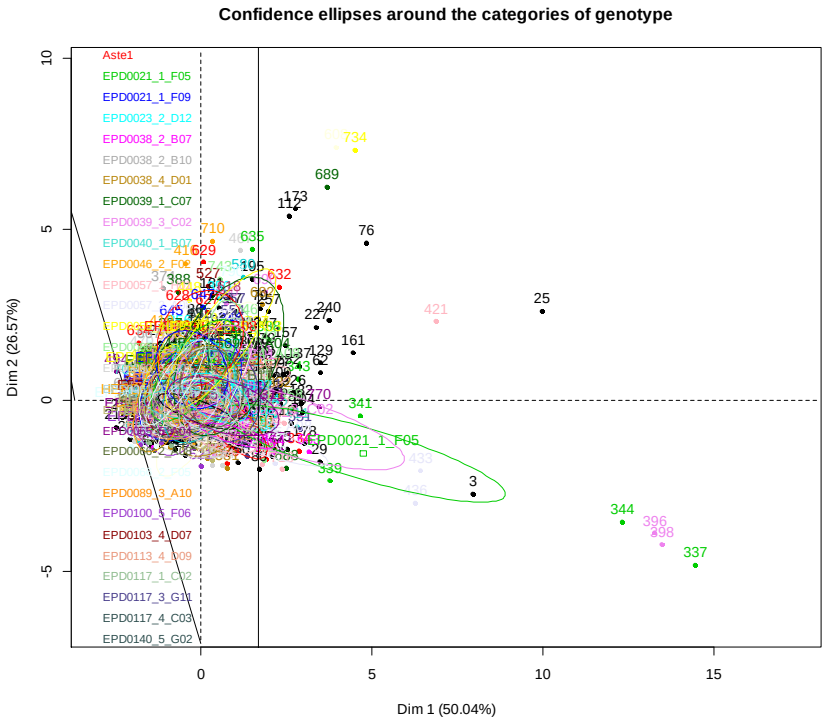


Figure 11: Représentation des individus en fonction des composantes principales, regroupés par génotype.

Comme le montre la figure, certaines lignées mutantes ont l'air de se comporter différemment des autres. Cela est vérifié statistiquement dans la prochaine partie au moyen d'une Analyse de la Variance à un facteur, ou son équivalent non paramétrique le Kruskal-Wallis, suivit par des comparaisons multiples si nécessaire. Il en sera de même pour toutes les composantes principales qui suivent.

3. Capacités motrices

16 paramètres mesurent les capacités motrices de la souris grâce à plusieurs tests mesurant les aptitudes de l'animal à sauter, se déplacer, s'accrocher ou résister à la chaleur.

Capacités motrices	OpenField_Dist.global	OpenField_Number.of.Rearings.global	OpenField_Resting.Time.whole.arena	OpenField_Distance.periphery	OpenField_Resting.Time.periphery	OpenField_Distance.center	SHIRPA_Loc.Activity	Grip_Trial	Grip_Trial.4paws.	Rotarod_Trial	PPI_BN.65.	PPI_P70	PPI_P80	PPI_P85	PPI_P90	PPI_ST
OpenField_Dist.global	1	0,331	-0,891	0,814	-0,862	0,642	0,259	0,043	0,024	0,061	0,058	0,058	0,043	-0,04	-0,009	-0,012
OpenField_Number.of.Rearings.global	0,331	1	-0,385	0,254	-0,382	0,232	0,068	-0,018	-0,01	-0,008	-0,017	0,001	-0,018	-0,007	0,012	-0,108
OpenField_Resting.Time.whole.arena	-0,891	-0,385	1	-0,686	0,985	-0,629	-0,197	-0,005	-2E-04	-0,038	-0,047	-0,048	-0,028	0,038	0,007	0,014
OpenField_Distance.periphery	0,814	0,254	-0,686	1	-0,606	0,081	0,142	0,06	0,098	-0,047	0,069	0,074	0,086	-0,056	-0,041	-0,054
OpenField_Resting.Time.periphery	-0,862	-0,382	0,985	-0,606	1	-0,684	-0,203	0,012	0,025	-0,052	-0,033	-0,029	-0,008	0,038	0,008	0,008
OpenField_Distance.center	0,642	0,232	-0,629	0,081	-0,684	1	0,259	-0,008	-0,087	0,164	0,009	0,004	-0,04	0,007	0,04	0,047
SHIRPA_Loc.Activity	0,259	0,068	-0,197	0,142	-0,203	0,259	1	0,013	-0,036	0,143	0,07	0,068	0,051	0,046	0,02	0,018
Grip_Trial	0,043	-0,018	-0,005	0,06	0,012	-0,008	0,013	1	0,654	-0,028	0,128	0,111	0,132	0,077	0,093	-0,003
Grip_Trial.4paws.	0,024	-0,01	-2E-04	0,098	0,025	-0,087	-0,036	0,654	1	-0,054	0,112	0,117	0,157	-0,009	0,007	-0,035
Rotarod_Trial	0,061	-0,008	-0,038	-0,047	-0,052	0,164	0,143	-0,028	-0,054	1	0,028	0,054	1E-03	0,035	0,021	0,066
PPI_BN.65.	0,058	-0,017	-0,047	0,069	-0,033	0,009	0,07	0,128	0,112	0,028	1	0,912	0,712	0,346	0,137	-0,06
PPI_P70	0,058	0,001	-0,048	0,074	-0,029	0,004	0,068	0,111	0,117	0,054	0,912	1	0,736	0,377	0,153	-0,062
PPI_P80	0,043	-0,018	-0,028	0,086	-0,008	-0,04	0,051	0,132	0,157	1E-03	0,712	0,736	1	0,469	0,259	-0,055
PPI_P85	-0,04	-0,007	0,038	-0,056	0,038	0,007	0,046	0,077	-0,009	0,035	0,346	0,377	0,469	1	0,715	0,178
PPI_P90	-0,009	0,012	0,007	-0,041	0,008	0,04	0,02	0,093	0,007	0,021	0,137	0,153	0,259	0,715	1	0,347
PPI_ST	-0,012	-0,108	0,014	-0,054	0,008	0,047	0,018	-0,003	-0,035	0,066	-0,06	-0,062	-0,055	0,178	0,347	1

Figure 12: Matrice de corrélations des paramètres mesurant les capacités motrices.

Peu de variables sont fortement corrélées entres elles à l'intérieur de cette fonction biologique. Cela influe donc sur le nombre de composantes nécessaires afin de garder 75% de l'information globale contenue dans les données, il en faut ici 6[11] :

- La composante 1 : aptitude à se déplacer
- La composante 2 : aptitude à sauter
- La composante 3 : aptitude faible à s'agripper
- La composante 4 : aptitude forte à s'agripper
- La composante 5 : aptitude à résister à la chaleur
- La composante 6 : aptitude à se redresser

4. Métabolisme hépatique

Les 5 paramètres renseignant du métabolisme hépatique c'est-à-dire relatif au foie de l'animal sont fortement corrélés :

Metabolisme hépatique	BC_T.proteins	BC_Albumin	BC_LDH	BC_ASAT	BC_ALAT
BC_T.proteins	1	0,65	-0	0,002	0,065
BC_Albumin	0,65	1	0,016	0,028	0,13
BC_LDH	-0	0,016	1	0,863	0,72
BC_ASAT	0,002	0,028	0,863	1	0,482
BC_ALAT	0,065	0,13	0,72	0,482	1

Figure 13 : Matrice de corrélations des paramètres mesurant le métabolisme hépatique

[11] Les cercles de corrélations correspondants sont visibles en Annexe, il en sera de même pour toutes les fonctions biologiques qui suivent.

Environ 80% de la variance des données peut être expliquée au moyen de deux composantes principales :

- La première renseigne sur le taux total de Protéine et d'Albumine chez la souris
- La seconde informe sur le taux de LDH, d'ASAT et d'ALAT

5. Schizophrénie

10 paramètres évaluent le reflexe de sursaut de la souris déclenché par un stimulus auditif inattendu précédé à court terme d'un pré-stimulus plus faible et qui n'induit donc pas le reflexe. Ces paramètres donnent des indications sur le trouble de la schizophrénie.

Schizophrénie	PPI_BN.65.	PPI_P70	PPI_P80	PPI_P85	PPI_P90	PPI_ST	PPI_PP70	PPI_PP80	PPI_PP85	PPI_PP90
PPI_BN.65.	1	0,912	0,712	0,347	0,141	-0,055	-0,037	-0,007	0,043	0,162
PPI_P70	0,912	1	0,736	0,377	0,156	-0,058	-0,055	-0,022	0,034	0,158
PPI_P80	0,712	0,736	1	0,47	0,262	-0,051	-0,112	-0,163	-0,058	0,088
PPI_P85	0,347	0,377	0,47	1	0,716	0,182	0,099	0,126	0,106	0,264
PPI_P90	0,141	0,156	0,262	0,716	1	0,347	0,262	0,267	0,256	0,392
PPI_ST	-0,055	-0,058	-0,051	0,182	0,347	1	0,879	0,789	0,75	0,679
PPI_PP70	-0,037	-0,055	-0,112	0,099	0,262	0,879	1	0,857	0,807	0,704
PPI_PP80	-0,007	-0,022	-0,163	0,126	0,267	0,789	0,857	1	0,885	0,792
PPI_PP85	0,043	0,034	-0,058	0,106	0,256	0,75	0,807	0,885	1	0,901
PPI_PP90	0,162	0,158	0,088	0,264	0,392	0,679	0,704	0,792	0,901	1

Figure 14: Matrice de Corrélations des paramètres indiquant sur le trouble de la Schizophrénie.

Deux composantes résument ces données :

- La première rend compte de l'information lorsqu'il y a pré-stimulus
- La seconde lorsqu'il n'y a pas de pré-stimulus

6. Hématologie

Dans l'ensemble des données, les cellules sanguines sont étudiées dans 7 paramètres. Le tout informe donc sur l'hématologie de la souris.

Hématologie	BC_Iron	WBC	RBC	HGB	HCT	MCV	PLT
BC_Iron	1	0,01	0,05	0,09	0,06	0,02	0,03
WBC	0,01	1	0,09	-0,02	0,16	0,2	-0,02
RBC	0,05	0,09	1	0,81	0,92	-0,04	0,16
HGB	0,09	-0,02	0,81	1	0,76	0	0,14
HCT	0,06	0,16	0,92	0,76	1	0,35	0,09
MCV	0,02	0,2	-0,04	0	0,35	1	-0,15
PLT	0,03	-0,02	0,16	0,14	0,09	-0,15	1

Figure 15: Matrice de Corrélations des paramètres renseignant sur l'Hématologie.

Les variables étant peu liées entres elles, pour garder plus de 75% de l'information 4 composantes principales sont nécessaires :

- La composante 1 : Hémoglobine, Hématocrite, Globules rouges
- La composante 2 : Volume moyen de cellules
- La composante 3 : Taux de Fer
- La composante 4 : Nombre de Plaquettes et Globules blancs

7. Immunologie

L'immunologie regroupe 16 paramètres qui s'occupent de l'étude du système immunitaire. Elle renseigne sur le taux de différents anticorps et cellules dans le sang.

Immunologie	FACS_T.cell.CD4.	FACS_T.cell.CD8.	FACS_T.cell.CD3.	FACS_Treg.cell.CD25.	FACS_B.cell.CD19.	FACS_Mature.B.cells.IgD.	FACS_Granulocyte.Gr1.	FACS_NK.cell	FACS_monocyte.gate	Luminex_IgM	Luminex_IgG3	Luminex_IgG1	Luminex_IgG2b	Luminex_IgG2a	Luminex_IgA	Luminex_IgE
FACS_T.cell.CD4.	1	0,814	0,386	-0,22	-0,65	0,43	0,066	0,248	0,015	0,01	0,142	0,191	0,203	-0,18	-0,14	-0,01
FACS_T.cell.CD8.	0,814	1	0,417	-0,3	-0,67	0,35	-0,01	0,358	-0,12	-0,06	0,182	0,239	0,125	-0,18	-0,15	-0,01
FACS_T.cell.CD3.	0,386	0,417	1	-0,19	-0,51	0,187	0,354	0,304	-0,04	-0,07	0,264	0,073	0,094	-0,1	-0,09	0,603
FACS_Treg.cell.CD25.	-0,22	-0,3	-0,19	1	0,07	-0,23	0,046	0,081	0,018	-0,02	-0,1	0,034	-0,08	0,025	-0,05	-0
FACS_B.cell.CD19.	-0,65	-0,67	-0,51	0,07	1	-0,17	-0,31	-0,51	-0,12	0,101	-0,09	-0,36	-0,07	0,197	0,15	-0,21
FACS_Mature.B.cells.IgD.	0,43	0,35	0,187	-0,23	-0,17	1	0,031	0,22	0,069	0,108	0,117	0,196	-0,23	-0,15	-0,03	-0,06
FACS_Granulocyte.Gr1.	0,066	-0,01	0,354	0,046	-0,31	0,031	1	-0	0,493	-0,12	-0,15	0,031	-0,09	-0,07	-0,12	0,226
FACS_NK.cell	0,248	0,358	0,304	0,081	-0,51	0,22	-0	1	-0,13	-0,01	0,295	0,301	-0	-0,13	-0,03	0,124
FACS_monocyte.gate	0,015	-0,12	-0,04	0,018	-0,12	0,069	0,493	-0,13	1	-0,09	-0,12	0,005	0,022	-0,03	-0,03	-0,07
Luminex_IgM	0,01	-0,06	-0,07	-0,02	0,101	0,108	-0,12	-0,01	-0,09	1	0,447	0,23	0,154	0,186	0,506	-0,08
Luminex_IgG3	0,142	0,182	0,264	-0,1	-0,09	0,117	-0,15	0,295	-0,12	0,447	1	0,2	0,344	0,346	0,431	0,111
Luminex_IgG1	0,191	0,239	0,073	0,034	-0,36	0,196	0,031	0,301	0,005	0,23	0,2	1	-0,18	0,165	0,075	-0,06
Luminex_IgG2b	0,203	0,125	0,094	-0,08	-0,07	-0,23	-0,09	-0	0,022	0,154	0,344	-0,18	1	0,07	0,206	0,086
Luminex_IgG2a	-0,18	-0,18	-0,1	0,025	0,197	-0,15	-0,07	-0,13	-0,03	0,186	0,346	0,165	0,07	1	0,215	0,005
Luminex_IgA	-0,14	-0,15	-0,09	-0,05	0,15	-0,03	-0,12	-0,03	-0,03	0,506	0,431	0,075	0,206	0,215	1	-0,08
Luminex_IgE	-0,01	-0,01	0,603	-0	-0,21	-0,06	0,226	0,124	-0,07	-0,08	0,111	-0,06	0,086	0,005	-0,08	1

Figure 16 : Matrice des corrélations des paramètres de la fonction Immunologie

Au vu du peu de corrélation au sein des variables, il ne faut pas moins de 7 composantes pour conserver au minimum 75 % de l'explication de la variance :

- La composante 1 : Pourcentage de lymphocytes T et B
- La composante 2 : Taux d'immunoglobulines IgM, IgG3 et IgA
- La composante 3 : Taux d'immunoglobuline IgE
- La composante 4 : Pourcentage de Monocyte et taux IgG1
- La composante 5 : Pourcentage de Monocyte, de lymphocytes NK et T
- La composante 6 : Taux IgG2b, pourcentage de lymphocytes B matures et T
- La composante 7 : Taux d'immunoglobuline IgG2a

8. Métabolisme Rénal

9 paramètres définissent le métabolisme rénal et donc informent sur les reins de la souris :

Metabolisme Rénal	BC_Urea	BC_Creatinine	BC_Na	BC_K	BC_Cl	BC_T.proteins	BC_Albumin	BC_Ca	BC_P
BC_Urea	1	0,33	0,13	0	0,09	0,02	-0,06	0,23	0,14
BC_Creatinine	0,33	1	0,11	0,2	0,15	-0,17	-0,07	0,01	0,25
BC_Na	0,13	0,11	1	-0,05	0,64	0,25	0,17	0,13	0,25
BC_K	0	0,2	-0,05	1	0,03	0,05	0,02	-0,2	0,07
BC_Cl	0,09	0,15	0,64	0,03	1	-0,04	-0,01	-0,03	0,19
BC_T.proteins	0,02	-0,17	0,25	0,05	-0,04	1	0,66	-0,01	-0,19
BC_Albumin	-0,06	-0,07	0,17	0,02	-0,01	0,66	1	-0,14	-0,21
BC_Ca	0,23	0,01	0,13	-0,2	-0,03	-0,01	-0,14	1	0,33
BC_P	0,14	0,25	0,25	0,07	0,19	-0,19	-0,21	0,33	1

Figure 17: Matrice de corrélations des paramètres relatifs au métabolisme rénal.

Peu de liens entre variables, cela implique qu'un grand nombre de composantes est utile à l'obtention de plus de 75% d'explication de la variance des données. 5 composantes seront nécessaires :

- La composante 1 : Taux de Phosphore, de Chlorure, de Créatinine et de Sodium
- La composante 2 : Taux total de Protéine et d'Albumine
- La composante 3 : Taux de Potassium et de Calcium
- La composante 4 : Taux d'Urée et de Chlorure
- La composante 5 : Taux de Potassium, de Phosphore et d'Urée

9. Métabolisme énergétique

22 paramètres caractérisent le métabolisme énergétique de la souris :

Métabolisme Energétique	Calori_VO2day	Calori_VO2_night	Calori_VCO2day	Calori_VCO2night	Calori_HeatDay	Calori_HeatNight	Calori_XAMBday	Calori_XAMBnight	Calori_XTOTday	Calori_XTOTnight	Calori_Total.Food.intake	Calori_Cumulative.FoodDay	Calori_Cumulative.FoodNight	Calori_RERday	Calori_RERnight	DEXA_Fat.Tissue	DEXA_Lean.Tissue	BC_T.Chol	BC_TG	BC_FFA	BC_HDL.Chol	BC_Glycerol
Calori_VO2day	1	0,94	0,72	0,66	0,99	0,93	0,26	0,22	0,24	0,21	0,17	0,17	0,18	-0,41	-0,44	-0,09	0,12	0,03	-0,25	-0,01	0,05	-0,09
Calori_VO2_night	0,94	1	0,64	0,75	0,93	0,99	0,15	0,32	0,14	0,3	0,11	0,12	0,13	-0,41	-0,42	-0,08	0,12	0,03	-0,24	0,01	0,06	-0,07
Calori_VCO2day	0,72	0,64	1	0,84	0,8	0,7	0,29	0,08	0,29	0,09	0,36	0,37	0,36	0,32	0,21	-0,11	0,19	0,05	-0,15	-0,02	0,05	-0,05
Calori_VCO2night	0,66	0,75	0,84	1	0,72	0,82	0,09	0,23	0,08	0,23	0,22	0,23	0,26	0,21	0,28	-0,1	0,12	0,05	-0,16	0,02	0,06	-0
Calori_HeatDay	0,99	0,93	0,8	0,72	1	0,93	0,27	0,21	0,26	0,2	0,21	0,22	0,22	-0,3	-0,35	-0,1	0,14	0,03	-0,24	-0,01	0,05	-0,09
Calori_HeatNight	0,93	0,99	0,7	0,82	0,93	1	0,15	0,31	0,13	0,3	0,13	0,14	0,15	-0,33	-0,31	-0,09	0,12	0,03	-0,24	0,01	0,06	-0,06
Calori_XAMBday	0,26	0,15	0,29	0,09	0,27	0,15	1	0,28	0,98	0,26	0,27	0,31	0,34	0	-0,11	-0,08	0,09	-0,07	0,01	0,03	-0,09	-0,01
Calori_XAMBnight	0,22	0,32	0,08	0,23	0,21	0,31	0,28	1	0,25	0,97	0	0	-0,02	-0,18	-0,15	-0,19	0	-0,21	-0,13	0	-0,19	-0,07
Calori_XTOTday	0,24	0,14	0,29	0,08	0,26	0,13	0,98	0,25	1	0,28	0,31	0,34	0,35	0,02	-0,1	-0,1	0,09	-0,08	-0,01	0,03	-0,1	-0,01
Calori_XTOTnight	0,21	0,3	0,09	0,23	0,2	0,3	0,26	0,97	0,28	1	0,04	0,03	0	-0,17	-0,13	-0,2	0,01	-0,22	-0,16	0	-0,21	-0,06
Calori_Total.Food.intake	0,17	0,11	0,36	0,22	0,21	0,13	0,27	0	0,31	0,04	1	0,98	0,87	0,22	0,1	-0,14	0,22	-0,11	-0,11	0,11	-0,14	0,01
Calori_Cumulative.FoodDay	0,17	0,12	0,37	0,23	0,22	0,14	0,31	0	0,34	0,03	0,98	1	0,93	0,22	0,11	-0,14	0,2	-0,1	-0,1	0,1	-0,13	0,01
Calori_Cumulative.FoodNight	0,18	0,13	0,36	0,26	0,22	0,15	0,34	-0,02	0,35	0	0,87	0,93	1	0,21	0,15	-0,14	0,1	-0,05	-0,06	0,11	-0,08	0,04
Calori_RERday	-0,41	-0,41	0,32	0,21	-0,3	-0,33	0	-0,18	0,02	-0,17	0,22	0,22	0,21	1	0,9	-0,03	0,06	0,03	0,12	-0,03	0,01	0,03
Calori_RERnight	-0,44	-0,42	0,21	0,28	-0,35	-0,31	-0,11	-0,15	-0,1	-0,13	0,1	0,11	0,15	0,9	1	-0,02	-0,02	0,03	0,12	-0,02	0	0,07
DEXA_Fat.Tissue	-0,09	-0,08	-0,11	-0,1	-0,1	-0,09	-0,08	-0,19	-0,1	-0,2	-0,14	-0,14	-0,14	-0,03	-0,02	1	-0,39	0,28	0,35	0,1	0,34	0,2
DEXA_Lean.Tissue	0,12	0,12	0,19	0,12	0,14	0,12	0,09	0	0,09	0,01	0,22	0,2	0,1	0,06	-0,02	-0,39	1	0,03	-0,07	-0,03	0,03	-0,15
BC_T.Chol	0,03	0,03	0,05	0,05	0,03	0,03	-0,07	-0,21	-0,08	-0,22	-0,11	-0,1	-0,05	0,03	0,03	0,28	0,03	1	0,32	0,04	0,94	0,1
BC_TG	-0,25	-0,24	-0,15	-0,16	-0,24	-0,24	0,01	-0,13	-0,01	-0,16	-0,11	-0,1	-0,06	0,12	0,12	0,35	-0,07	0,32	1	0,36	0,24	0,37
BC_FFA	-0,01	0,01	-0,02	0,02	-0,01	0,01	0,03	0	0,03	0	0,11	0,1	0,11	-0,03	-0,02	0,1	-0,03	0,04	0,36	1	0,05	0,83
BC_HDL.Chol	0,05	0,06	0,05	0,06	0,05	0,06	-0,09	-0,19	-0,1	-0,21	-0,14	-0,13	-0,08	0,01	0	0,34	0,03	0,94	0,24	0,05	1	0,1
BC_Glycerol	-0,09	-0,07	-0,05	-0	-0,09	-0,06	-0,01	-0,07	-0,01	-0,06	0,01	0,01	0,04	0,03	0,07	0,2	-0,15	0,1	0,37	0,83	0,1	1

Figure 18 : Matrice de corrélations des paramètres caractérisant le métabolisme énergétique.

6 composantes sont nécessaires pour expliquer 75% de la variance :

- La composante 1 : Respiration de l'animal (production CO_2, consommation O_2)
- La composante 2 : Masse de nourriture ingérée par l'animal
- La composante 3 : Cholestérol
- La composante 4 : Glycérol et Acides gras
- La composante 5 : Activité ambulatoire la nuit
- La composante 6 : Activité ambulatoire le jour

Les autres fonctions biologiques ne sont pas modifiées car elles comportent trop peu de variables, une ACP n'est donc pas réalisable.

Nous arrivons donc à moins d'une quarantaine de nouvelles variables qui sont des combinaisons linéaires de nos paramètres de départ. En rajoutant les paramètres n'ayant pas été modifiés, nos données sont donc réduites à un peu plus d'une cinquantaine de variables maintenant.

II. Etude des phénotypes

En génétique, le phénotype est l'ensemble des caractères observables d'un individu, il est défini par opposition au génotype. Notre étude se porte donc sur la présence de variations phénotypiques dues aux variations génétiques.

Nous introduisons le modèle :

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}\ i = 1 \dots I, j = 1 \dots J_i$$

Avec la contrainte $\alpha_1 = 0$.

Soit α_i représentant le facteur contrôlé lignée de souris à I modalités.

La contrainte $\alpha_1 = 0$ est utilisée afin que la lignée contrôle serve de lignée de référence.

Soit $Y_{i,j}$ la valeur prise par la variable réponse Y pour la lignée *i* lors de la *j*-ème répétition.
Les hypothèses classiques pour les erreurs sont pour $1 \leq i \leq I, 1 \leq j \leq J_i$:

- Indépendance des $\varepsilon_{i,j}$
- $\mathcal{L}(\varepsilon_{i,j}) = \mathcal{N}(0, \sigma^2)$

Dans chaque cas, on suppose que les erreurs sont indépendantes.

Pour chacune de nos nouvelles variables, après vérification de l'hypothèse de normalité, puis celle d'homoscédasticité, une Analyse de la Variance à un facteur est effectuée si les hypothèses sont vérifiées, dans le cas contraire la méthode non paramétrique de Kruskal-Wallis lui sera préférée.

Grâce à ces méthodes, nous pouvons conclure à l'apparition ou non d'un phénotype dans au moins l'une des lignées mutantes. Les comparaisons multiples permettent de savoir quelles lignées mutantes ont mis en évidence le phénotype en question.

Le test de Shapiro-Wilk est utilisé pour tester l'hypothèse de normalité, puis si cette hypothèse est vérifiée le test de Bartlett servira à tester l'homoscédasticité. Dans le cas où l'hypothèse de normalité n'est pas vérifiée, le test de Levène testera l'égalité des variances.

Commençons par regarder l'effet du génotype sur les variables représentant le poids.
Soit $Y_{i,j}$ la mesure du poids lors du Pipeline 1 de la *j*-ème souris de génotype *i*.

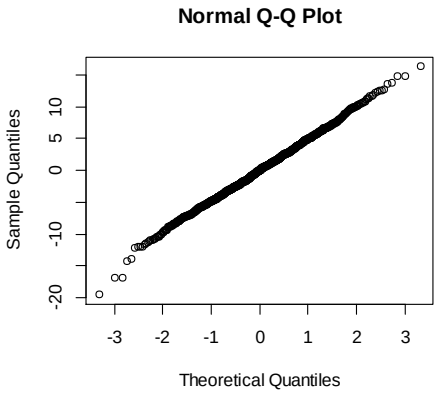


Figure 19 : QQplot des résidus

```
        Shapiro-Wilk normality test

data:  res
W = 0.999, p-value = 0.7985
```

L'hypothèse de normalité des résidus est donc vérifiée. Un test de Bartlett est utilisé pour vérifier l'égalité des variances :

```
                    Bartlett test of homogeneity
of variances
        data:  Poids_Pip1 and genotype
        Bartlett's K-squared = 109.0818, df = 77,
p-value = 0.009481
```

L'hypothèse d'homoscédasticité n'étant pas vérifiée, un test non paramétrique de Kruskal-Wallis est effectué :

```
                Kruskal-Wallis rank sum test

      data:  Poids_Pip1 by genotype
      Kruskal-Wallis chi-squared = 265.2791, df = 77, p-value < 2.2e-16
```

La p-value étant inferieure à $2{,}2 \times 10^{16}$, nous pouvons conclure à la différence significative d'au moins un des échantillons. Les comparaisons multiples au groupe contrôle nous indiquerons le ou lesquelles, pour cela nous utilisons la fonction ***kruskalmc()*** du Package R **`pgirmess`** :

```
Multiple comparison test after Kruskal-Wallis, treatment vs control (two-tailed)
p.value: 0.05
Comparisons
                              obs.dif critical.dif difference
A_baseline-Aste1           149.06867     367.6471      FALSE
                                            …
A_baseline-EPD0023_2_D12   367.96867     329.5266       TRUE
                                            …
A_baseline-EPD0038_4_D01   376.34645     346.9865       TRUE
A_baseline-EPD0039_1_C07   358.31867     367.6471      FALSE
                                            …
A_baseline-EPD0060_2_D02    79.14010     392.6165      FALSE
A_baseline-EPD0060_2_H09   429.80633     367.6471       TRUE
A_baseline-EPD0060_3_G04   135.67978     346.9865      FALSE
                                            …
A_baseline-EPD0116_1_C09   292.85438     392.6165      FALSE
A_baseline-EPD0117_1_C02   531.90200     346.9865       TRUE
A_baseline-EPD0117_3_G11    35.56867     423.6258      FALSE
A_baseline-EPD0117_4_C03   352.01311     346.9865       TRUE
                                            …
A_baseline-EPD0146_4_H09   370.76867     329.5266       TRUE
A_baseline-EPD0155_1_B07   131.94367     367.6471      FALSE
                                            …
A_baseline-EPD0173_1_C12   117.44367     367.6471      FALSE
A_baseline-EPD0173_1_F02   476.06867     367.6471       TRUE
A_baseline-EPD0175_3_D06   211.81867     367.6471      FALSE
                                            …
A_baseline-HEPD0505_2_B06 162.54244      346.9865      FALSE
A_baseline-HEPD0507_7_C03 412.56867      392.6165       TRUE
A_baseline-HEPD0507_8_G08 215.63133      329.5266      FALSE
                                            …
A_baseline-HEPD0537_2_B03 341.85438      392.6165      FALSE
A_baseline-HEPD0540_3_F06 368.44367      367.6471       TRUE
                                            …
```

9 lignées mutantes sortent significativement différentes du groupe contrôle.

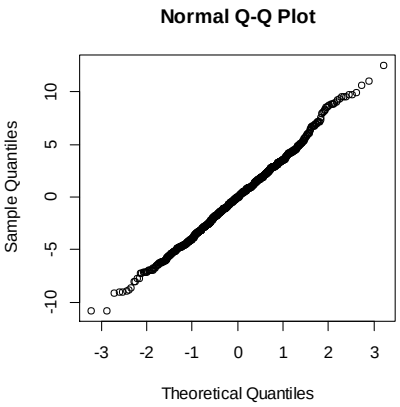


Figure 20 : QQplot des residus

Effectuons maintenant l'analyse avec comme variable réponse le poids de la souris lors du Pipeline 2. Commençons par la vérification des hypothèses :

```
            Shapiro-Wilk normality test

      data:  res
      W = 0.9963, p-value = 0.06503
```

La normalité est vérifiée, testons l'égalité des variances :

```
                Bartlett test of homogeneity of variances
      data:  Poids_Pip2 and genotype
      Bartlett's K-squared = 74.8729, df = 57, p-value = 0.05633
```

Les deux hypothèses nécessaires à l'application de l'Analyse de la Variance sont vérifiées. Les p-valeurs seront ajustées grâce à la méthode de Benjamini-Hochberg pour prendre en compte les comparaisons multiples :

```
Call:
lm(formula = Poids_Pip2 ~ genotype, data = Poids_Pip2_CP)

                            pval_adj
(Intercept)             5.593317e-05
genotypeAste1           1.259391e-05
genotypeEPD0021_1_F05   2.545732e-01
genotypeEPD0021_1_F09   1.300599e-05
…
genotypeEPD0038_2_B10   8.969829e-03
…
genotypeEPD0060_3_G04   1.415539e-09
genotypeEPD0061_1_C10   8.434082e-03
genotypeEPD0065_2_E04   2.628792e-02
genotypeEPD0065_5_A04   1.585959e-01
genotypeEPD0066_2_A08   4.438728e-03
genotypeEPD0066_2_F05   1.931845e-02
genotypeEPD0089_3_A10   4.515072e-01
genotypeEPD0100_5_F06   4.507054e-04
genotypeEPD0103_4_D07   1.686477e-02
genotypeEPD0113_4_D09   4.515072e-01
genotypeEPD0117_1_C02   2.465333e-05
genotypeEPD0117_3_G11   3.060182e-02
…
genotypeEPD0173_1_12   7.052183e-04
genotypeEPD0173_1_F02   8.662198e-01
genotypeEPD0175_3_D06   3.060182e-02
…
genotypeGpt2            3.485394e-03
genotypeHEPD0501_1_F05 4.515072e-01
genotypeHEPD0507_7_C03 4.296013e-03
genotypeHEPD0509_6_A04 1.984105e-02
…
genotypeHEPD0516_3_C04 4.507054e-04
genotypeHEPD0522_1_A11 1.243226e-04
…
genotypeHEPD0547_1_C09 9.645246e-04
genotypeHEPD0547_4_B11 2.406706e-02
…

Residual standard error: 3.933 on 712 degrees of freedom
Multiple R-squared: 0.3049,      Adjusted R-squared: 0.2492
F-statistic: 5.478 on 57 and 712 DF,  p-value: < 2.2e-16
```

22 lignées mutantes ont un poids lors du Pipeline 2 significativement différent de la lignée contrôle.

Des analyses identiques ont été faites pour toutes les composantes, afin d'éviter de répéter la description de la méthode, les résultats trouvés sont résumés dans le tableau qui suit.
Les variables possédant un nombre de données manquantes supérieur à 50% ont été supprimées.

	Poids_Pip1	Poids_Pip2	Aptitude à se déplacer	Aptitude à se sauter	Aptitude faible à s'aggriper	Aptitude forte à s'aggriper	Aptitude à résister à la chaleur	Aptitude à se redresser	audition Faibles dB	audition Fortes dB	Réaction AVEC pré-stimulus	Réaction SANS pré-stimulus	Hémoglobine, Hématocrite, Globules rouges	Volume moyen de cellules	Taux d'Iron	Nombre de Plaquettes et Globules blancs	LDH, d'ASAT et d'ALAT	taux total de Protéine et d'Albumine	Taux de Phosphore, de Chlorure, de Créatinine et de Sodium	Taux total de Protéine et d'Albumine2	Taux de Potassium et de Calcium	Taux d'Urée et de Chlorure	Taux de Potassium, de Phosphore et d'Urée	% de lymphocytes T et B	Taux d'anticorps IgM, IgG3 et IgA	Taux d'anticorps IgE	% Monocyte Gate et taux IgG1	%de Monocyte Gate, de lymphocytes NK et T régl	Taux d'anticorps IgG2b, % de lymphocytes B matures et Treg	Taux d'anticorps IgG2a
Aste1	FALSE	1,26E-05	FALSE	FALSE	FALSE	FALSE	0,900770907	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,278730415	FALSE	FALSE	FALSE	0,000919191	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Agxt2l2	FALSE	0,254573171	FALSE	TRUE	FALSE	FALSE	0,914602671	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	0,88374646	FALSE	FALSE	TRUE	0,134543792	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE
Prpsap2	FALSE	1,30E-05	FALSE	FALSE	FALSE	FALSE	0,914602671	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,987187227	FALSE	FALSE	FALSE	0,973082934	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Tcf7	TRUE	0,056794534	FALSE	FALSE	FALSE	FALSE	0,92025082	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,03805242	FALSE	FALSE	TRUE	0,088416624	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
C1qbpt	FALSE	0,154284713	FALSE	FALSE	FALSE	FALSE	0,961900688	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Xbp1	FALSE	0,008969829	FALSE	FALSE	FALSE	FALSE	0,577917404	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	0,355226062	FALSE	FALSE	TRUE	0,048443798	FALSE	NA	NA	NA	NA	NA	NA	NA
Ube2b	TRUE	0,154284713	FALSE	FALSE	FALSE	FALSE	0,577917404	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	4,76E-10	FALSE	TRUE	FALSE	0,934528086	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
Srsf4	FALSE	0,631259865	FALSE	FALSE	FALSE	FALSE	0,001027204	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,603983102	FALSE	FALSE	FALSE	0,670665099	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE
Grhl3	FALSE	0,702913843	FALSE	FALSE	FALSE	FALSE	0,801805491	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	0,258586041	FALSE	FALSE	FALSE	0,934528086	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
Limk2	FALSE	0,45150719	FALSE	FALSE	FALSE	FALSE	0,070446573	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,104524181	FALSE	FALSE	FALSE	0,945076479	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
Yipf1	FALSE	0,789048179	FALSE	FALSE	FALSE	FALSE	0,801805491	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,314971245	FALSE	FALSE	FALSE	0,093429769	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Fam63a	FALSE	0,545206892	FALSE	FALSE	FALSE	FALSE	0,451403876	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	0,000146117	FALSE	TRUE	TRUE	0,388921193	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
Ndufs3	FALSE	0,545206892	FALSE	FALSE	TRUE	FALSE	0,541595829	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	0,114064477	TRUE	FALSE	TRUE	0,596079896	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Ngfr	FALSE	0,465189356	FALSE	FALSE	FALSE	FALSE	9,54E-06	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	0,298554786	FALSE	FALSE	FALSE	0,934528086	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE
Aldh3a2	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,577917404	FALSE	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,050661295	FALSE	FALSE	FALSE	0,091352587	FALSE	NA	NA	NA	NA	NA	NA	NA
B9d1	TRUE	0,545206892	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,314971245	TRUE	FALSE	FALSE	0,953525078	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
Sept08"	FALSE	1,42E-09	FALSE	FALSE	TRUE	FALSE	0,32121927	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	0,114064477	FALSE	FALSE	FALSE	0,137559796	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
Fabp3	FALSE	0,008434082	FALSE	FALSE	FALSE	TRUE	0,993525008	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,59473794	FALSE	FALSE	FALSE	0,130307659	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE
Clk1	FALSE	0,026287919	FALSE	FALSE	FALSE	FALSE	0,858940359	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	2,50E-08	TRUE	FALSE	TRUE	0,000919191	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
Jmjd5	FALSE	0,158595905	FALSE	TRUE	FALSE	FALSE	0,718858768	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	0,123574597	TRUE	FALSE	TRUE	0,012697477	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE
Ttll1	FALSE	0,004438728	FALSE	FALSE	FALSE	FALSE	0,914602671	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	0,836118648	FALSE	FALSE	FALSE	0,953525078	FALSE	NA	NA	NA	NA	NA	NA	NA
Ift20	FALSE	0,019318454	FALSE	TRUE	FALSE	TRUE	0,577917404	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	0,603983102	FALSE	FALSE	FALSE	0,05504036	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
Orc4	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,314971245	FALSE	FALSE	FALSE	0,589092352	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
Ruvbl1	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	TRUE	FALSE	FALSE	FALSE	0,314971245	FALSE	FALSE	TRUE	0,953525078	FALSE	NA	NA	NA	NA	NA	NA	NA
Mrps35	FALSE	0,45150719	FALSE	FALSE	FALSE	FALSE	0,914602671	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,114064477	FALSE	TRUE	FALSE	0,596079896	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
Elk4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,258586041	FALSE	FALSE	FALSE	0,861671028	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Nfxl1	FALSE	0,000450705	FALSE	FALSE	FALSE	FALSE	0,577917404	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0,941578493	FALSE	FALSE	FALSE	0,924581921	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Ddx52	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,340957625	FALSE	FALSE	FALSE	0,945076479	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
Pik3c3	FALSE	0,016864768	FALSE	FALSE	FALSE	FALSE	0,801805491	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,627321354	FALSE	FALSE	FALSE	0,934528086	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE
Flot2	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,906895149	FALSE	FALSE	FALSE	0,714190604	FALSE	NA	NA	NA	NA	NA	NA	NA
Bspry	FALSE	0,45150719	FALSE	FALSE	FALSE	FALSE	0,674353088	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,839344627	TRUE	FALSE	FALSE	0,020427083	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
Itgb5	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	TRUE	FALSE	FALSE	FALSE	FALSE	0,048927124	FALSE	FALSE	FALSE	0,571945904	FALSE	NA	NA	NA	NA	NA	NA	NA
Eif2b5	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	TRUE	FALSE	FALSE	FALSE	0,9114511	FALSE	FALSE	FALSE	0,953525078	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
Tmem68	TRUE	2,47E-05	FALSE	FALSE	FALSE	FALSE	0,577917404	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	0,03805242	TRUE	FALSE	FALSE	0,263893657	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Gar1	FALSE	0,030601817	FALSE	FALSE	FALSE	FALSE	0,577917404	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,019715937	FALSE	TRUE	FALSE	0,827935512	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
Pgd	TRUE	0,545206892	FALSE	FALSE	FALSE	FALSE	0,577917404	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,499931827	FALSE	FALSE	FALSE	0,827935512	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
B4galnt3	FALSE	0,097671388	FALSE	FALSE	FALSE	FALSE	0,891033261	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,896913756	FALSE	FALSE	FALSE	0,861671028	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Fkbp10	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,314971245	FALSE	FALSE	FALSE	0,83995121	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
Crtc1	FALSE	0,078222806	FALSE	FALSE	FALSE	FALSE	0,142442726	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	0,88374646	FALSE	FALSE	FALSE	0,688116676	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE
Fabp12	TRUE	0,955992887	FALSE	FALSE	FALSE	FALSE	0,674353088	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,019715937	FALSE	FALSE	FALSE	0,299974884	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Tnfrsf1b	FALSE	0,826890087	FALSE	FALSE	FALSE	TRUE	0,336483797	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,014227354	FALSE	TRUE	FALSE	0,845641953	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Ift122	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	0,285121073	FALSE	FALSE	FALSE	1,24E-08	FALSE	NA	NA	NA	NA	NA	NA	NA
H6pd	FALSE	0,000705218	TRUE	FALSE	FALSE	TRUE	0,993525008	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0,083853551	FALSE	FALSE	FALSE	0,83995121	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Mib2	TRUE	0,86621981	FALSE	FALSE	TRUE	TRUE	0,175207971	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,101990345	FALSE	FALSE	FALSE	5,98E-05	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
1700052N19Rik	FALSE	0,030601817	FALSE	FALSE	FALSE	FALSE	0,674353088	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,88374646	FALSE	FALSE	FALSE	0,953525078	FALSE	NA	NA	NA	NA	NA	NA	NA
Mrpl54	FALSE	0,096093982	FALSE	FALSE	FALSE	FALSE	0,577917404	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,000486585	FALSE	TRUE	FALSE	0,140847299	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE
Mtmr14	NA	0,833861201	FALSE	FALSE	TRUE	FALSE	0,718858768	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,430291491	FALSE	FALSE	FALSE	0,051439491	FALSE	NA	NA	NA	NA	NA	NA	NA
Otop3	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,083853551	FALSE	FALSE	FALSE	0,861671028	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2310067B10Rik	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	TRUE	FALSE	FALSE	FALSE	0,952492813	FALSE	FALSE	FALSE	0,783062691	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE
Drg2	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,340957625	FALSE	FALSE	FALSE	0,973082934	FALSE	NA	NA	NA	NA	NA	NA	NA
Sox13	FALSE	0,8577565	FALSE	FALSE	FALSE	FALSE	0,961900688	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,431881408	FALSE	FALSE	FALSE	0,113524901	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Slc25a38	FALSE	0,833861201	FALSE	FALSE	TRUE	TRUE	0,891033261	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,590426485	FALSE	FALSE	FALSE	0,237450129	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
Lhx1	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,273832563	FALSE	FALSE	FALSE	0,134543792	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
Ephx1	FALSE	0,230256841	FALSE	FALSE	FALSE	FALSE	0,577917404	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	0,059036142	FALSE	FALSE	FALSE	0,62269619	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Gpt2	FALSE	0,003485304	FALSE	FALSE	FALSE	TRUE	0,707876157	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,350203928	FALSE	FALSE	FALSE	0,140847299	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
Mrpl10	FALSE	0,45150719	FALSE	FALSE	FALSE	FALSE	0,17926641	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,3807525	FALSE	FALSE	FALSE	0,134543792	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
Bag1	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,88374646	FALSE	FALSE	FALSE	0,374392115	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Fpgs	TRUE	0,004296013	FALSE	FALSE	FALSE	FALSE	0,993525008	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,083853551	FALSE	FALSE	FALSE	0,091352587	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Cacnb4	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,340957625	FALSE	FALSE	FALSE	0,388921193	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
Cops4	FALSE	0,019841048	FALSE	FALSE	FALSE	FALSE	0,993525008	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,88374646	FALSE	FALSE	FALSE	0,943368856	TRUE	NA	NA	NA	NA	NA	NA	NA
Col5a1	FALSE	0,45150719	FALSE	FALSE	FALSE	FALSE	0,577917404	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,312590685	FALSE	FALSE	FALSE	0,029061429	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
Arcn1	FALSE	0,179839561	FALSE	TRUE	FALSE	FALSE	0,914602671	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	0,101990345	FALSE	FALSE	FALSE	0,688116676	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
E330016A19Rik	FALSE	0,000450705	FALSE	FALSE	FALSE	TRUE	0,577917404	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,007498036	FALSE	TRUE	FALSE	0,000919191	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
Btk	FALSE	0,001124323	FALSE	FALSE	FALSE	FALSE	0,993525008	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	0,941578493	TRUE	FALSE	FALSE	0,032003438	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
Xiap	FALSE	0,68785403	FALSE	FALSE	FALSE	FALSE	0,577917404	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,087321451	FALSE	TRUE	FALSE	0,973082934	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE
2900026A02Rik	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,101990345	FALSE	FALSE	FALSE	0,945076479	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Zfp13	FALSE	0,542591509	FALSE	FALSE	FALSE	FALSE	0,916503107	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,88374646	FALSE	FALSE	FALSE	0,386618209	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Rnase10	FALSE	0,791289524	FALSE	FALSE	FALSE	FALSE	0,55339796	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,029526588	TRUE	FALSE	FALSE	0,62269619	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Pacs2	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	0,079903367	FALSE	TRUE	FALSE	0,827935512	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Fgl1	TRUE	0,463875759	FALSE	FALSE	FALSE	FALSE	0,993525008	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,981033531	FALSE	FALSE	FALSE	0,137559796	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Mosc2	FALSE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	FALSE	TRUE	FALSE	0,361570836	FALSE	FALSE	FALSE	0,62269619	FALSE	NA	NA	NA	NA	NA	NA	NA
Asxc3	FALSE	0,000964525	FALSE	FALSE	FALSE	FALSE	0,900770907	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,314971245	FALSE	FALSE	FALSE	0,12704704	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
Bloc1s1	FALSE	0,545206892	FALSE	FALSE	FALSE	FALSE	0,142442726	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,000201438	TRUE	TRUE	FALSE	0,973082934	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
Snx5	NA	0,230256841	FALSE	FALSE	FALSE	FALSE	0,92025082	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	0,314971245	FALSE	FALSE	FALSE	0,130307659	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Figure 21 : Tableau récapitulatif des phénotypes par lignée mutante.

Chaque ligne représente une lignée mutante identifiée maintenant par le nom du gène muté par exemple « Tcf7 », alors qu'elle était identifiée par son code « EPD0023_2_D12 » tout au long de l'analyse. La correspondance de chaque code est disponible en Annexe.

Une case rouge est le croisement d'une lignée mutante et d'un paramètre pour lequel la mutation en question entraine un phénotype.

Les cases roses représentent les cas où la comparaison avec le groupe contrôle n'a pas pu être effectuée. Cela peut être dû au fait que les valeurs du paramètre sont manquantes pour la lignée.

Seuls les résultats des variables modifiées sont visibles dans ce tableau.

Conclusion :

Le fait de réduire le nombre de paramètres a permis une analyse plus rapide de toutes nos données et de garder un maximum de l'information qui y est contenue.

L'une des étapes importante de l'Analyse en Composantes Principales est l'interprétation de chaque composante ainsi produite. Cette étape mériterait une étude plus approfondit de la part des biologistes afin de récolter une information plus précise de ces nouvelles variables.

Classification non supervisée des gènes

Lors des études précédentes, seules les données complètes étant prises en compte, pour cette nouvelle étude, il est nécessaire de mettre en place une méthode d'imputation. La suppression des données incomplètes réduit considérablement le nombre de lignées mutantes à étudier, il deviendrait alors délicat d'effectuer une classification.

Nous supposons que nos données manquantes sont de type MAR (Missing At Random).

La méthode d'imputation multiple est plus performante et donne des résultats plus cohérents que les méthodes d'imputations simples, c'est pour cela qu'elle a été choisie. Avec environ 22,5% de données manquantes, nous utiliserons 5 imputations pour obtenir un tableau de données complet et satisfaisant.

Après les différentes Analyses en Composantes Principales effectuées précédemment, le nombre de paramètres a été réduit de plus de moitié, il reste tout de même une cinquantaine de nouvelles variables.

I. Etude du tableau complet

1. Classification ascendante hiérarchique

Les données sont centrées-réduites préalablement. La distance entre deux individus est caractérisée par la distance euclidienne. La méthode de Ward est utilisée afin garder une distance inter-classe maximum, ce qui est équivalent, d'après le théorème de Huygens, à ce que l'augmentation de l'inertie intra-classe soit minimum.
Pour une visualisation simple de l'organisation des données, le dendrogramme constitue la représentation graphique la plus parlante. La hauteur du cluster dans le dendrogramme représente la similarité entre les deux clusters avant l'agrégation.

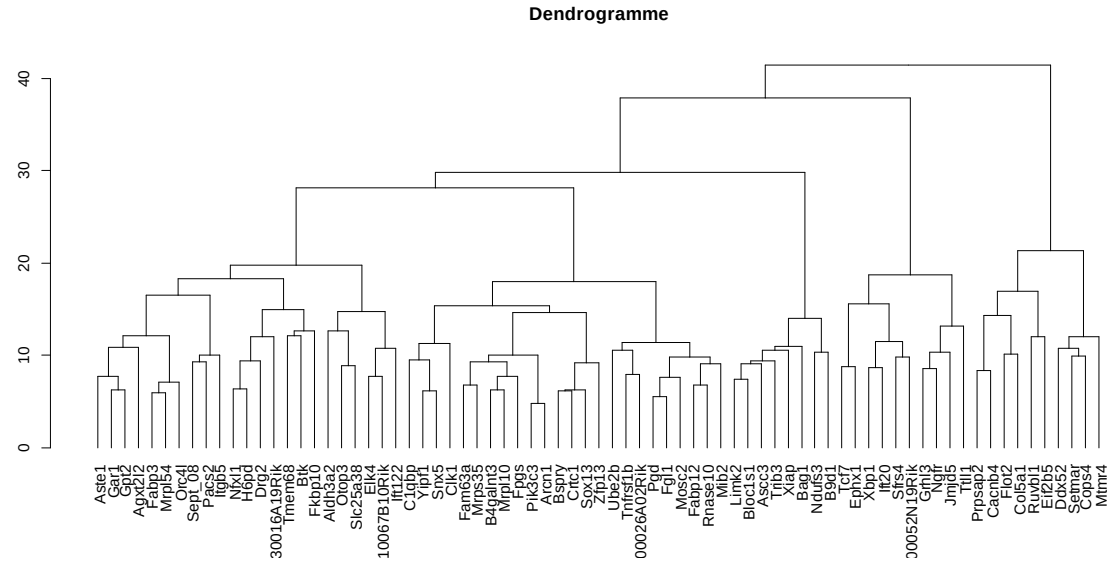
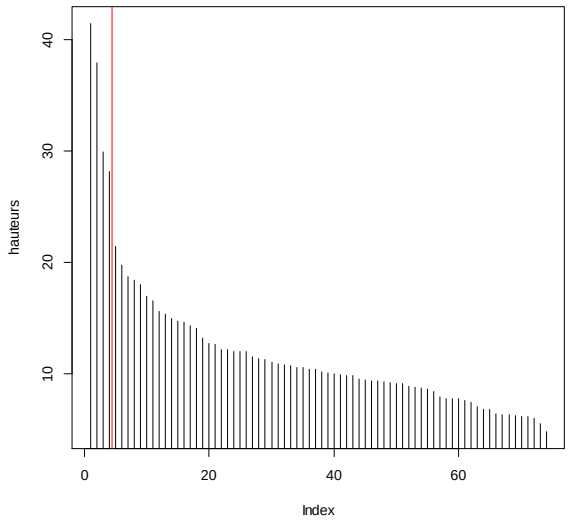


Figure 22 : Dendrogramme



Ce graphique permet de visualiser l'évolution des similarités entre classes et donc de déterminer combien de clusters il faut retenir.

Une forte dissimilarité est visible entre les deux clusters agrégés pour passer de 5 à 4 clusters. Il parait donc logique de ne pas fusionner ces deux classes.

Le nombre de classes à retenir est donc de 5, la figure ci-dessous représente le dendrogramme ainsi obtenu.

Figure 23 : Evolution des similarités entre classes.

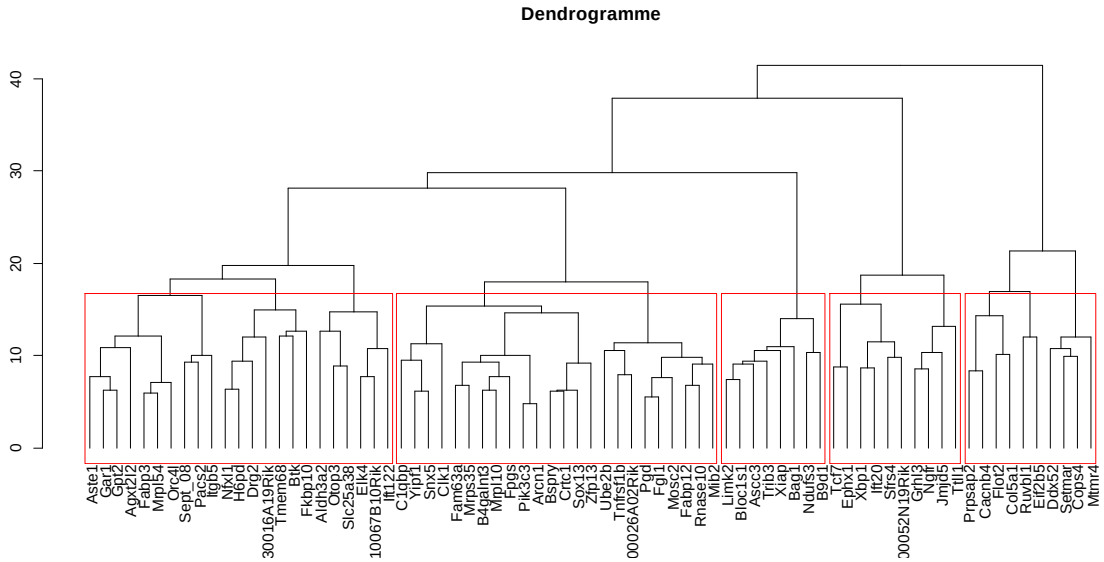


Figure 24 : Dendrogramme séparé en 5 classes.

Classe 1		Classe 2	Classe 3	Classe 4		Classe 5
Aste1	Ift122	Prpsap2	Tcf7	C1qbp	Tnfrsf1b	Limk2
Agxt2l2	H6pd	Ruvbl1	Xbp1	Ube2b	Mib2	Ndufs3
Aldh3a2	Mrpl54	Ddx52	Sfrs4	Yipf1	Sox13	B9d1
Sept_08	Otop3	Flot2	Grhl3	Fam63a	Mrpl10	Bag1
Fabp3	2310067B10Rik	Eif2b5	Ngfr	Clk1	Fpgs	Trib3
Orc4l	Drg2	Mtmr4	Jmjd5	Mrps35	Arcn1	Xiap
Elk4	Slc25a38	Setmar	Ttll1	Pik3c3	2900026A02Rik	Ascc3
Nfxl1	Gpt2	Cacnb4	Ift20	Bspry	Zfp13	Bloc1s1
Itgb5	E330016A19Rik	Cops4	1700052N19Rik	Pgd	Rnase10	
Tmem68	Btk	Col5a1	Ephx1	B4galnt3	Fgl1	
Gar1	Pacs2			Crtc1	Mosc2	
Fkbp10				Fabp12	Snx5	

Figure 25 : Composition des 5 classes.

Chaque classe est caractérisée par des variables, pour lesquelles les individus de ces classes ont des valeurs plus élevées ou au contraire plus faibles que les individus des autres groupes. Ces informations sont récoltées grâce à la fonction ***catdes ()*** du package `FactoMineR`.

```
Classe `1`
                              v.test Mean in category Overall mean sd in category Overall sd      p.value
audition_FortesDb          -2.854474       -0.7023150   -0.3473491      0.5729693  0.7114405 4.310816e-03
AptitudeForteAgripper      -3.121495       -0.8349289   -0.2456429      1.1002039  1.0800442 1.799351e-03
PoidsPip2                  -5.497027       -4.5769710   -0.4532481      2.1537958  4.2918020 3.862468e-08

Classe `2`
                              v.test Mean in category Overall mean sd in category Overall sd      p.value
PoidsPip2                   5.747902        6.8579716 -0.453248116      3.5536159  4.2918020 9.035773e-09
AptitudeRedresser           3.204362        0.5888150  0.081427775      0.7189914  0.5342655 1.353622e-03
audition_FortesDb           2.812797        0.2457370 -0.347349147      0.8076275  0.7114405 4.911265e-03
%lymphoT&B                  2.773763        3.2371050  0.176818858      2.7235325  3.7226471 5.541204e-03
NbPlaquettes&GloBlancs      2.671844        0.5942546  0.079650742      0.4248392  0.6498619 7.543569e-03
AptitudeFaibleAgripper      2.490891        1.2385430  0.356075503      1.2935211  1.1953725 1.274234e-02
ReactionAVECpréstimulus     2.300152        1.8745286  0.304983893      2.8463778  2.3023761 2.143959e-02
Anxiety_OF_Latency.center   2.293039        0.3369973  0.019302232      0.6020068  0.4674748 2.184575e-02
VolumeMoyCellules           2.190634        0.6094049  0.009723143      0.9709631  0.9236557 2.847826e-02
Bone.Metabo_BC_Ca          -2.055684       -0.5159435 -0.094410268      0.7745686  0.6918863 3.981301e-02
AptitudeDeplacer           -2.656163       -2.8938651 -0.866615919      3.2631700  2.5752040 7.903545e-03
AptitudeResChaleur         -3.440505       -0.8852765 -0.073486460      0.8107070  0.7961240 5.806301e-04

Classe`3`
                              v.test Mean in category Overall mean sd in category Overall sd      p.value
TauxIgM.G3.A                3.427740        0.4585119  -0.71837318      1.0916122  1.1584715 6.086278e-04
TauxIgE                     3.264634        0.8107513  -0.35848796      1.2471052  1.2084482 1.096055e-03
ReactionSANSpréstimulus    -2.268824       -1.2219619  -0.38516974      0.8767569  1.2444454 2.327900e-02
ReactionAVECpréstimulus    -2.826517       -0.4647435   0.07965074      0.6790737  0.6498619 4.705722e-03
%lymphoT&B                 -6.903382       -7.4396659   0.17681886      2.3114433  3.7226471 5.077894e-12

Classe`4`
                                v.test Mean in category Overall mean sd in category Overall sd     p.value
AptitudeDeplacer              3.124301        0.4967990 -0.866615919      1.4076475  2.5752040 0.001782283
Anxiety_OF_NEntries.center    2.920273        0.1700172 -0.108363239      0.3929285  0.5625377 0.003497244
Anxiety_OF_Perm.Time.center   2.235203        0.1362815 -0.061352111      0.4960701  0.5217714 0.025404009
Taux_LDH_ASAT_ALAT           -2.068408       -0.4698508 -0.009159621      0.9413056  1.3143470 0.038601655
Anxiety_OF_Latency.center    -2.390441       -0.1700628  0.019302232      0.3480662  0.4674748 0.016828156
PoidsPip1                    -2.758742       -2.5385277 -0.913717493      2.1228472  3.4755854 0.005802436

Classe`5`
                              v.test Mean in category Overall mean sd in category Overall sd      p.value
PoidsPip1                   5.294226        5.2765157   -0.9137175      2.4538910  3.4755854 1.195214e-07
Bone.Metabo_DEXA_BMC        3.748629        0.5652826   -0.1072086      0.1803385  0.5332583 1.778040e-04
HeartWeight                 2.706020        0.5697585    0.1166610      0.7763944  0.4977188 6.809499e-03
Taux_K_Ca                  -2.092446       -0.5284983    0.0514555      0.5383499  0.8238772 3.639860e-02
```

Nous remarquons que le Poids est une variable caractéristique de presque chacune des classes :

- Les individus de la classe 1 ont un poids moyen mesuré dans le Pipeline 2 inferieur à la moyenne
- Les individus de la classe 2 ont un poids moyen mesuré dans le Pipeline 2 supérieur à la moyenne
- Les individus de la classe 4 ont un poids moyen mesuré dans le Pipeline 1 inferieur à la moyenne
- Les individus de la classe 5 ont un poids moyen mesuré dans le Pipeline 1 supérieur à la moyenne

Les souris de la classe 2 ont des résultats dans les paramètres résumant leurs capacités motrices visiblement distincts des autres souris.

La classe 3 est identifiée par des valeurs plus faibles que la moyenne dans les variables caractéristique de la Schizophrénie, et des taux d'immunoglobulines bien différents de la moyenne.

La classe 4 est déterminée par des variables expliquant l'anxiété de la souris.

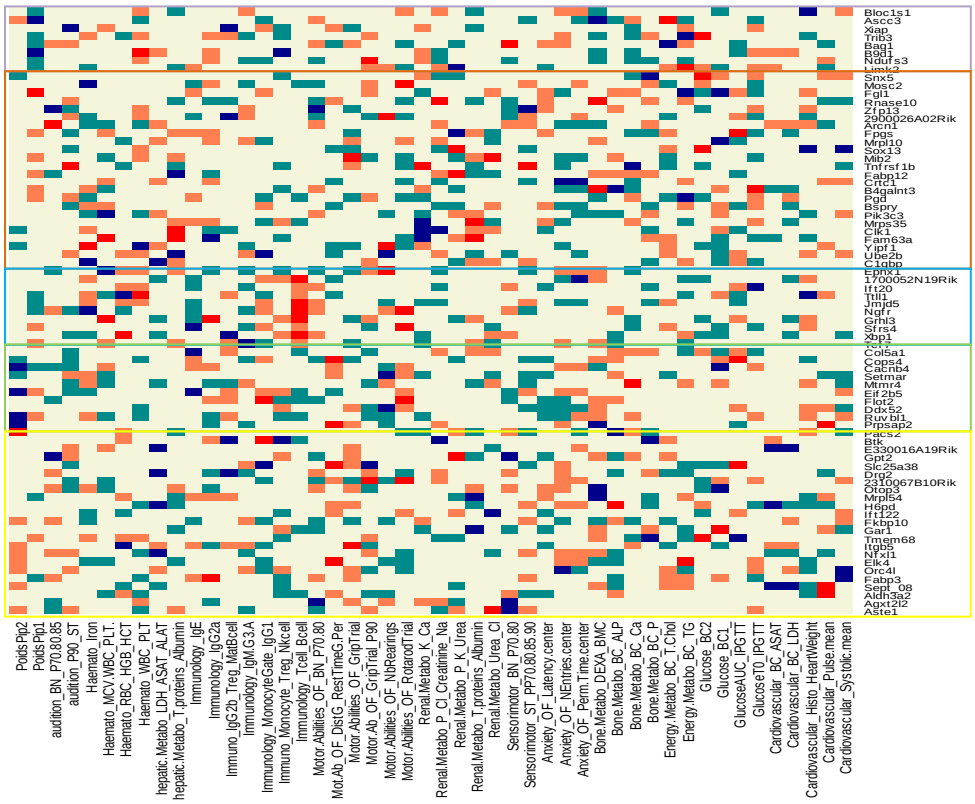


Figure 26 : Heat-Map des écarts entre la valeur de la lignée et la valeur moyenne générale exprimée en nombre d'Ecarts-types.

Aucun profil ne se dégage réellement de cette classification.

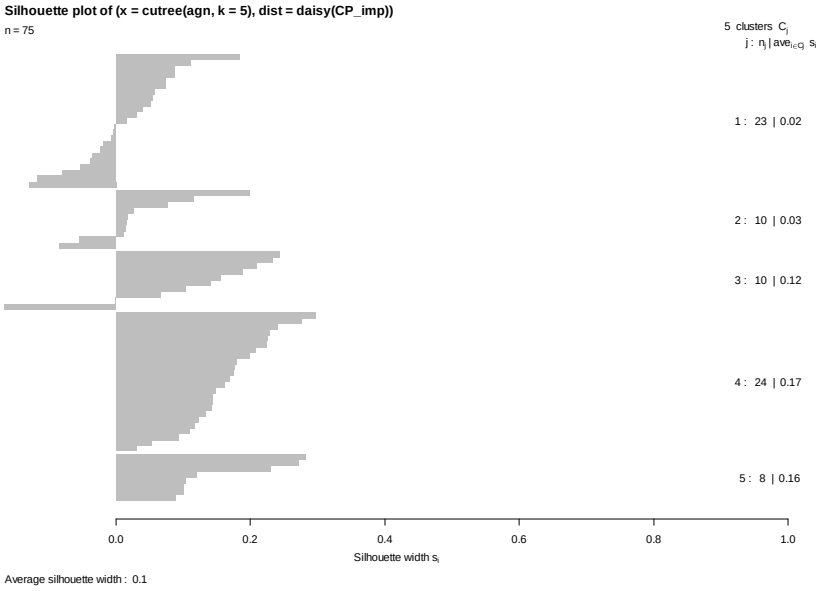


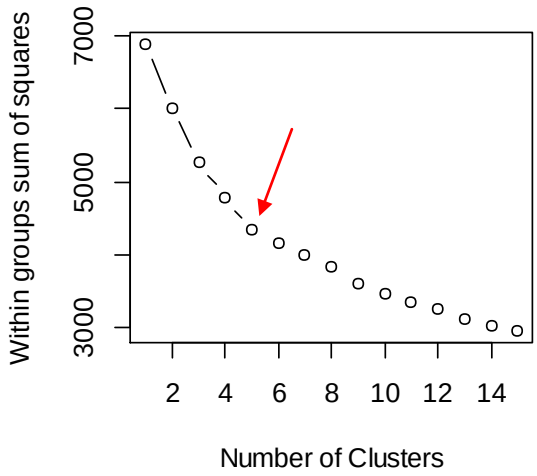
Figure 27 : Graphique des silhouettes, Classification hiérarchique

Avec un indice de silhouette moyen de *0.1*, nous remarquons donc que les individus ne sont pas particulièrement bien classés.

Une classification n'est jamais unique, elle dépend toujours de la méthode de classification utilisée. Regardons les résultats obtenus avec des méthodes de classification à partitionnement.

2. Méthode K-means

L'algorithme des K-means a pour but de diviser les individus en *K* classes dans lesquelles chaque individu appartient à la classe avec la moyenne la plus proche.



La difficulté est de pouvoir déterminer ce nombre de classe a priori. Une bonne « classe » regroupe des individus très rapprochés, avec donc peu de dispersion.

Nous choisissons *K* au point d'inflexion de la courbe (flèche rouge).

On définit donc $K = 5$.

Figure 28 : Aide à la détermination du nombre de clusters

Classe 1		Classe 2		Classe 3	Classe 4	Classe 5
C1qbp	Mtmr4	Aste1	Ift122	Agxt2l2	Prpsap2	Tcf7
Ube2b	Sox13	Aldh3a2	H6pd	Limk2	Clk1	Xbp1
Yipf1	Slc25a38	Sept_08	Otop3	Ndufs3	Ruvbl1	Sfrs4
Fam63a	Ephx1	Orc4l	2310067B10Rik	B9d1	Ddx52	Grhl3
Mrps35	Mrpl10	Elk4	Drg2	Fabp3	Flot2	Ngfr
Pik3c3	Fpgs	Nfxl1	Gpt2	Bspry	Eif2b5	Jmjd5
Pgd	Arcn1	Itgb5	E330016A19Rik	Mrpl54	Setmar	Ttll1
B4galnt3	2900026A02Rik	Tmem68	Btk	Bag1	Cacnb4	Ift20
Crtc1	Zfp13	Gar1	Pacs2	Trib3	Cops4	1700052N19Rik
Fabp12	Rnase10	Fkbp10		Xiap	Col5a1	
Tnfrsf1b	Fgl1			Ascc3		
Mib2	Mosc2			Bloc1s1		
Mtmr4	Snx5					

Figure 29 : Composition des 5 classes.

On retrouve des ressemblances avec les classes formées par la classification hiérarchique. Regardons les variables qui caractérisent ces différentes classes :

```
Classe `1`
                              v.test Mean in category Overall mean sd in category Overall sd     p.value
Anxiety_O_NEntries.center     2.864661        0.1647159 -0.108363239      0.4044908  0.5625377 0.004174556
AptitudeDeplacer              2.776904        0.3451984 -0.866615919      1.8301967  2.5752040 0.005487941
Anxiety_OF_Perm.Time.center   2.171982        0.1306916 -0.061352111      0.5027989  0.5217714 0.029856998
TauxIgM.G3.A                 -2.158501       -1.1421150 -0.718373178      0.8852287  1.1584715 0.030888878
Taux_LDH_ASAT_ALAT           -2.274010       -0.5156439 -0.009159621      0.9511151  1.3143470 0.022965388
Anxiety_OF_Latency.center    -2.771593       -0.2002568  0.019302232      0.2737188  0.4674748 0.005578278
PoidsPip1                    -2.923278       -2.6354345 -0.913717493      2.1487780  3.4755854 0.003463666

Classe `2`
                              v.test Mean in category Overall mean sd in category Overall sd     p.value
TauxIgM.G3.A                 3.179688        0.3733455 -0.718373178      1.2936647  1.1584715 1.474335e-03
Bone.Metabo_BC_ALP           2.821668        0.7220614  0.035282319      0.5139966  0.8212404 4.777458e-03
TauxIgG2a                    2.508058        0.4768830  0.080672565      0.3083385  0.5330248 1.213967e-02
TauxIgE                      2.076756        0.3853090 -0.358487959      1.1498267  1.2084482 3.782406e-02
Taux_LDH_ASAT_ALAT           1.963747        0.7557964 -0.009159621      1.6543344  1.3143470 4.955940e-02
NbPlaquettes&GloBlancs      -2.082856       -0.3215126  0.079650742      0.6429255  0.6498619 3.726433e-02
HeartWeight                 -2.132020       -0.1978357  0.116661047      0.3702342  0.4977188 3.300520e-02
PoidsPip2                   -2.208694       -3.2626642 -0.453248116      2.4874610  4.2918020 2.719593e-02
TauxTotalProteines_Albumin  -2.284145       -0.7343689  0.123498736      0.7196028  1.2672308 2.236300e-02
ReactionSANSprestimulus     -2.327784       -1.2437074 -0.385169743      0.8615293  1.2444454 1.992358e-02
PoidsPip1                   -2.537249       -3.5272743 -0.913717493      2.2338144  3.4755854 1.117273e-02
%lymphoT&B                  -4.557579       -4.8515472  0.176818858      3.1467777  3.7226471 5.174665e-06

Classe `3`
                              v.test Mean in category Overall mean sd in category Overall sd     p.value
```

TauxTotalProteines_Albumin	3.409930	0.87880501	0.12349874	1.1320535	1.2672308	6.497956e-04
HeartWeight	2.611446	0.34384987	0.11666105	0.5725022	0.4977188	9.016033e-03
%lymphoT&B	2.425797	1.75525970	0.17681886	1.5140295	3.7226471	1.527479e-02
ReactionSANSpréstimulus	2.000948	0.05007508	-0.38516974	1.4681339	1.2444454	4.539797e-02
PoidsPip1	1.996179	0.29897106	-0.91371749	2.9918154	3.4755854	4.591448e-02
TauxAnticorpsIgE	-2.031855	-0.78767116	-0.35848796	0.7714993	1.2084482	4.216830e-02
PoidsPip2	-4.536004	-3.85603722	-0.45324812	2.4749631	4.2918020	5.733001e-06

Classe '4'

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
PoidsPip2	6.043676	6.45558414	-0.45324812	3.4039805	4.2918020	1.506419e-09
NbPlaquettes&GloBlancs	3.025527	0.60335486	0.07965074	0.3884762	0.6498619	2.481999e-03
AptitudeRedresser	2.959299	0.50255152	0.08142778	0.6846302	0.5342655	3.083400e-03
%lymphoT&B	2.828698	2.98162097	0.17681886	2.5740644	3.7226471	4.673780e-03
Anxiety_OF_Latency.center	2.460344	0.32565227	0.01930223	0.6150711	0.4674748	1.388037e-02
audition_FortesDb	2.315055	0.09134687	-0.34734915	0.8542421	0.7114405	2.060992e-02
Taux_K_Ca	2.310789	0.55854732	0.05145550	0.7059868	0.8238772	2.084450e-02
ReactionAVECpréstimulus	2.218544	1.66551381	0.30498389	2.7298807	2.3023761	2.651777e-02
AptitudeFaibleAgripper	1.960622	0.98032875	0.35607550	1.3156444	1.1953725	4.992319e-02
AptitudeDéplacer	-2.240419	-2.40337075	-0.86661592	3.3127806	2.5752040	2.506376e-02
Bone.Metabo_BC_Ca	-2.272804	-0.51326220	-0.09441027	0.7265237	0.6918863	2.303798e-02
AptitudeResChaleur	-2.766433	-0.66011706	-0.07348646	0.9421094	0.7961240	5.667316e-03

Classe '5'

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
PoidsPip1	3.942609	4.4881597	-0.91371749	2.9237140	3.4755854	8.060009e-05
Bone.Metabo_BC_Ca	2.399368	0.5600218	-0.09441027	0.5216867	0.6918863	1.642341e-02
Taux_P_Cl_Creatinine_Na	2.354596	1.2371602	-0.02091397	0.9099552	1.3553664	1.854284e-02
Bone.Metabo_BC_P	2.067763	0.5654082	0.01519449	0.3880890	0.6749905	3.866234e-02
AptitudeFaibleAgripper	-2.112859	-0.6395750	0.35607550	0.8371817	1.1953725	3.461281e-02
NbPlaquettes&GloBlancs	-3.360277	-0.7812027	0.07965074	0.5566105	0.6498619	7.786437e-04
%lymphoT&B	-4.592767	-6.5631723	0.17681886	2.7747424	3.7226471	4.374078e-06

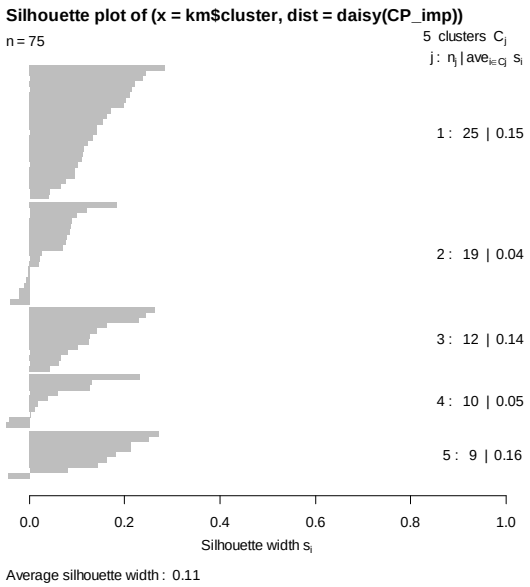


Figure 30 : Graphique des silhouettes, méthode K-means

Pour cette méthode des K-means, les indices de silhouettes sont à nouveau proches de 0. Cette méthode n'améliore donc pas considérablement la qualité des classes formées.

3. Méthode K-medoids

Tout comme celui des K-means, l'algorithme des K-medoids a pour but de diviser les individus en K classes. Chaque individu appartient à la classe avec le medoids le plus proche, c'est-à-dire le membre le plus central de la classe.

La fonction R ***pamk ()*** du package `fpc` nous aide à déterminer le nombre de classes optimal pour cette méthode, grâce au critère de Calinski et Harabasz[12] :

```
     1  clusters  0           6  clusters  7.572298
     2  clusters  10.02111    7  clusters  6.622169
     3  clusters  10.76416    8  clusters  6.379947
     4  clusters  8.730219    9  clusters  6.049781
     5  clusters  7.298097   10  clusters  5.709483
```

Le critère est maximisé pour 3 classes, donc $K = 3$.

Classe 1			Classe 2				Classe 3
Aste1	Tmem68	Gpt2	Agxt2l2	Mrps35	Mib2	Cops4	Xbp1
Aldh3a2	Gar1	Arcn1	Prpsap2	Ddx52	Mtmr4	Col5a1	Sfrs4
Sept_08	Fkbp10	E330016A19Rik	Tcf7	Flot2	Sox13	Trib3	Grhl3
Fabp3	Ift122	Btk	C1qbp	Bspry	Slc25a38	Xiap	Ndufs3
Orc4l	H6pd	2900026A02Rik	Ube2b	Eif2b5	Setmar	Zfp13	Ngfr
Elk4	Mrpl54	Pacs2	Limk2	Pgd	Ephx1	Rnase10	B9d1
Nfxl1	Otop3	Ascc3	Yipf1	B4galnt3	Mrpl10	Fgl1	Jmjd5
Pik3c3	2310067B10Rik	Bloc1s1	Fam63a	Crtc1	Bag1	Mosc2	Ttll1
Itgb5	Drg2		Clk1	Fabp12	Fpgs	Snx5	Ift20
			Ruvbl1	Tnfrsf1b	Cacnb4		1700052N19Rik

Figure 31 : Composition des 3 classes.

Regardons les variables qui caractérisent ces différentes classes :

```
Classe `1`
                          v.test Mean in category Overall mean sd in category Overall sd      p.value
Bone.Metabo_BC_ALP         2.725708        0.3925101   0.035282319      0.9142703  0.8212404 6.416365e-03
TauxTotalProteines_Albumin 2.515788        0.6322731   0.123498736      1.2598172  1.2672308 1.187668e-02
NbPlaquettes&GloBlancs     2.153186        0.3029556   0.079650742      0.5900894  0.6498619 3.130408e-02
VolumeMoyCellules          2.102504        0.3196380   0.009723143      0.8385970  0.9236557 3.550915e-02
HeartWeight                2.048317        0.2793570   0.116661047      0.5907675  0.4977188 4.052897e-02
ReactionAVECpréstimulus   -2.213532       -0.5083290   0.304983893      2.0886725  2.3023761 2.686097e-02
AptitudeFaibleAgripper    -3.159397       -0.7901973  -0.245642903      1.0175291  1.0800442 1.580959e-03
audition_FortesDb         -3.370854       -0.7300629  -0.347349147      0.5159045  0.7114405 7.493545e-04
PoidsPip2                 -5.778694       -4.4111487  -0.453248116      2.1173997  4.2918020 7.528273e-09

Classe `2`
                          v.test Mean in category Overall mean sd in category Overall sd      p.value
PoidsPip2                  5.338843       2.10586758  -0.45324812      3.7053795  4.2918020 9.354159e-08
%lymphoT&B                 3.080655       1.45766832   0.17681886      2.4026976  3.7226471 2.065455e-03
audition_FortesDb          2.670784      -0.13513202  -0.34734915      0.7492291  0.7114405 7.567432e-03
audition_FaiblessDb        2.302037      -0.23195291  -0.56236459      1.1202386  1.2851092 2.133310e-02
ReactionAVECpréstimulus    2.289452       0.89370672   0.30498389      2.3780451  2.3023761 2.205311e-02
AptitudeFaibleAgripper     2.232302       0.02363306  -0.24564290      1.0512754  1.0800442 2.559500e-02
Bone.Metabo_BC_ALP        -2.583044      -0.20163956   0.03528232      0.7046791  0.8212404 9.793291e-03

Classe `3`
                          v.test Mean in category Overall mean sd in category Overall sd      p.value
TauxIgE                    3.324227       0.83209471  -0.35848796      1.2392508  1.2084482 8.866390e-04
PoidsPip1                  2.391683       1.54989519  -0.91371749      4.4023975  3.4755854 1.677131e-02
TauxIgM.G3.A               2.333459       0.08279997  -0.71837318      0.5911800  1.1584715 1.962406e-02
TauxP_Cl_Creatinine_Na     2.299828       0.90291790  -0.02091397      1.0232423  1.3553664 2.145795e-02
TauxCa                     2.178315       0.35226947  -0.09441027      0.5079585  0.6918863 2.938258e-02
TauxBC_P                   1.987244       0.41274255   0.01519449      0.4849957  0.6749905 4.689540e-02
NbPlaquettes&GloBlancs    -3.056776      -0.50909188   0.07965074      0.6810579  0.6498619 2.237317e-03
%lymphoT&B                -6.831489      -7.36034709   0.17681886      2.3844256  3.7226471 8.403762e-12
```

[12] Calinski, T., Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics* **3**: 1-27

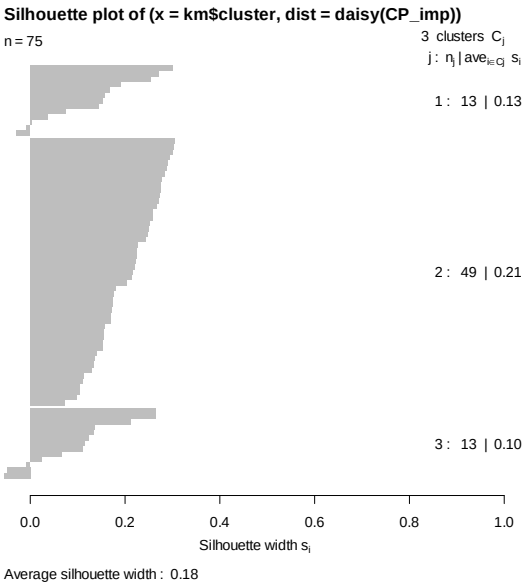


Figure 32 : Graphique des Silhouettes, méthode K-medoids

L'indice de silhouette moyen atteint 0.18, il reste toujours relativement faible.

La mauvaise qualité des différentes classes obtenues avec différentes méthodes peut être due aux grand nombre de paramètres pour un nombre de lignées faible. Il serait intéressant d'étudier les résultats sur un nombre limité de variables.

II. Etude de tableaux partiels

Afin de réduire encore le nombre de paramètres, nous séparons les variables caractérisant le comportement de l'animal par rapport aux variables résumant le métabolisme de celui-ci. Les différentes méthodes de classifications sont ensuite appliquées à ces nouveaux tableaux de données. Commençons par **le comportement**, une quinzaine de variables sont ici retenues :

Méthode K-means

Classe 1	Classe 2				Classe 3		
Prpsap2	Agxt2l2	Ngfr	Mib2	Xiap	Aste1	Pik3c3	2310067B10Rik
Clk1	Tcf7	B9d1	Mtmr4	2900026A02Rik	Xbp1	Itgb5	Drg2
Ruvbl1	C1qbp	Jmjd5	Sox13	Zfp13	Aldh3a2	Tmem68	Gpt2
Ddx52	Ube2b	Mrps35	Slc25a38	Rnase10	Sept_08	Gar1	Fpgs
Flot2	Sfrs4	Bspry	Ephx1	Fgl1	Fabp3	Fkbp10	E330016A19Rik
Eif2b5	Grhl3	Pgd	Mrpl10	Mosc2	Ttll1	Ift122	Btk
Setmar	Limk2	B4galnt3	Bag1	Bloc1s1	Ift20	H6pd	Pacs2
Cacnb4	Yipf1	Crtc1	Col5a1	Snx5	Orc4l	1700052N19Rik	Ascc3
Cops4	Fam63a	Fabp12	Trib3		Elk4	Mrpl54	
	Ndufs3	Tnfrsf1b	Arcn1		Nfxl1	Otop3	

Figure 33 : Composition des 3 classes.

Ces classes sont maintenant caractérisées uniquement par des variables expliquant le comportement de la souris :

```
Classe '1'
                          v.test Mean in category Overall mean sd in category Overall sd      p.value
PoidsPip2                6.109247         7.8007098   -0.45324812      2.8327505  4.2918020 1.001020e-09
AptitudeRedresser        3.456257         0.6627246    0.08142778      0.6701885  0.5342655 5.477338e-04
Anxiety_OF_Latency.center 3.439957        0.5255301    0.01930223      0.5772324  0.4674748 5.818069e-04
AptitudeDeplacer        -2.202405        -2.6520492   -0.86661592      3.4148294  2.5752040 2.763673e-02
AptitudeResChaleur      -3.386119        -0.9221152   -0.07348646      0.8488364  0.7961240 7.088867e-04

Classe '2'
                          v.test Mean in category Overall mean sd in category Overall sd    p.value
AptitudeDeplacer         2.414757       -0.15330592   -0.86661592      2.0921428  2.5752040 0.01574570
PoidsPip2                2.223418        0.64134814   -0.45324812      1.5424658  4.2918020 0.02618764
```

```
Anxiety_OF_Latency.center -2.065108      -0.09143515   0.01930223      0.3361591  0.4674748 0.03891276
AptitudeRedresser         -2.163472      -0.05115943   0.08142778      0.5024575  0.5342655 0.03050489

Classe `3`
                             v.test Mean in category Overall mean sd in category Overall sd      p.value
audition_FortesDb         -2.153930        -0.578143   -0.3473491      0.5803063  0.7114405 3.124565e-02
AptitudeForteAgripper     -2.155374        -0.596248   -0.2456429      1.1089518  1.0800442 3.113256e-02
PoidsPip2                 -6.402621        -4.591830   -0.4532481      1.6964008  4.2918020 1.527324e-10
```

Nous remarquons que la classe 1 et la classe 2 ont l'air de regrouper les individus aux caractéristiques opposées, dans le groupe 1 les souris se déplacent moins que la moyenne, c'est le contraire dans la classe 2, de même pour le redressement et l'anxiété.

Une nouvelle fois le poids joue un rôle principal dans la classification des différentes lignées mutantes. Ce qui peut paraitre logique, le poids a un effet non négligeable sur les aptitudes physiques de chacun.

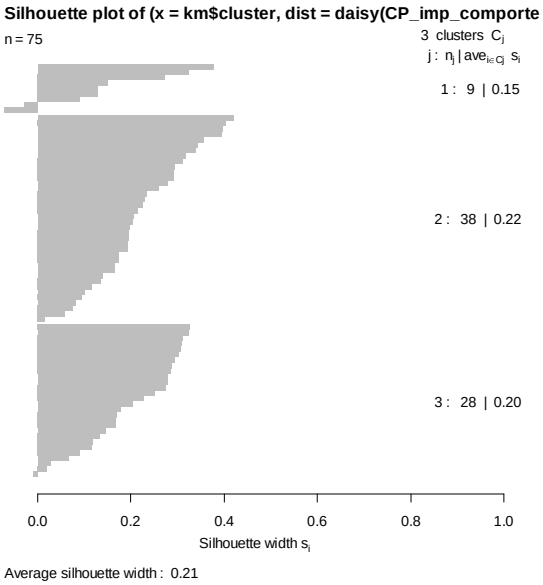


Figure 34 : Graphique des silhouettes, méthode K-means sur les variables du comportement.

L'indice de silhouette moyen augmente légèrement pour atteindre *0.21*, cela ne suffit pas à assurer la qualité des classes.

Pour **le métabolisme**, nous développerons uniquement la méthode K-means :

Méthode K-means

Classe 1	Classe 2		Classe 3				
Tcf7	Prpsap2	Setmar	Aste1	Bloc1s1	Fkbp10	2310067B10Rik	E330016A19Rik
Xbp1	C1qbp	Bag1	Agxt2l2	Mrps35	Crtc1	Drg2	Btk
Sfrs4	Ube2b	Cacnb4	Limk2	Elk4	Fabp12	Sox13	Xiap
Grhl3	Clk1	Cops4	Yipf1	Nfxl1	Tnfrsf1b	Slc25a38	2900026A02Rik
Ndufs3	Ruvbl1	Col5a1	Fam63a	Pik3c3	Ift122	Ephx1	Zfp13
Ngfr	Ddx52	Fgl1	Aldh3a2	Bspry	H6pd	Gpt2	Rnase10
B9d1	Flot2	Snx5	Sept_08	Itgb5	Mib2	Mrpl10	Pacs2
Jmjd5	Eif2b5		Fabp3	Gar1	Mrpl54	Fpgs	Mosc2
Ttll1			Tmem68	Pgd	Otop3	Trib3	Ascc3
Ift20			Orc4l	B4galnt3	Mtmr4	Arcn1	
1700052N19Rik							

Figure 35 : Composition des 3 classes

```
Classe `1`
                             v.test Mean in category Overall mean sd in category Overall sd      p.value
%lymphoT&B                3.276996        1.2098196   0.17681886      1.7739462  3.7226471 1.049177e-03
TauxTotalProteines_Albumin 2.234244        0.2471265   0.03217449      1.0852136  1.1361567 2.546702e-02
HeartWeight               2.033360        0.2023591   0.11666105      0.5037417  0.4977188 4.201619e-02
Taux_IgM.G3.A            -2.483192       -0.9619683  -0.71837318      1.1069174  1.1584715 1.302107e-02
Bone.Metabo_BC_P         -2.556809       -0.1309454   0.01519449      0.6501877  0.6749905 1.056373e-02
```

```
Taux_IgE                  -3.239576       -0.6899921  -0.35848796       0.7864469  1.2084482 1.197075e-03
PoidsPip2                 -5.435846       -2.4287600  -0.45324812       2.7169576  4.2918020 5.453690e-08

Classe `2`
                            v.test Mean in category Overall mean sd in category Overall sd      p.value
PoidsPip2                 6.363900          5.896802   -0.4532481       3.209804   4.291802 1.966939e-10
%lymphoT&B                2.274656          2.145529    0.1768189       2.909543   3.722647 2.292657e-02

Classe `3`
                              v.test Mean in category Overall mean sd in category Overall sd      p.value
Taux_IgM.G3.A              3.348095        0.3692060 -0.718373178      1.0667734  1.1584715 8.136924e-04
Taux_IgE                   3.221065        0.7329655 -0.358487959      1.2224547  1.2084482 1.277154e-03
Taux_P_Cl_Creatinine_Na    2.301342        0.8536987 -0.020913973      0.9879606  1.3553664 2.137229e-02
Bone.Metabo_BC_Ca          2.222293        0.3367253 -0.094410268      0.4868075  0.6918863 2.626350e-02
VolumeMoyCellules         -2.164048       -0.5507498  0.009723143      0.7077972  0.9236557 3.046070e-02
NbPlaquettes&GloBlancs    -3.184992       -0.5007222  0.079650742      0.6499024  0.6498619 1.447582e-03
%lymphoT&B                -6.980212       -7.1093342  0.176818858      2.4080485  3.7226471 2.947342e-12
```

Les classes ainsi formées sont légèrement différentes des précédentes à l'exception des 4 lignées suivantes qui restent toujours regroupées : Prpsap2, Ruvbl1, Eif2b5, Cacnb4.
Le pourcentage de lymphocytes T et B influe sur chacune des classes créées. La classe 3 représente ainsi les souris ayant un pourcentage de lymphocytes T et B bien inferieur à la moyenne en opposition aux classes 1 et 2.

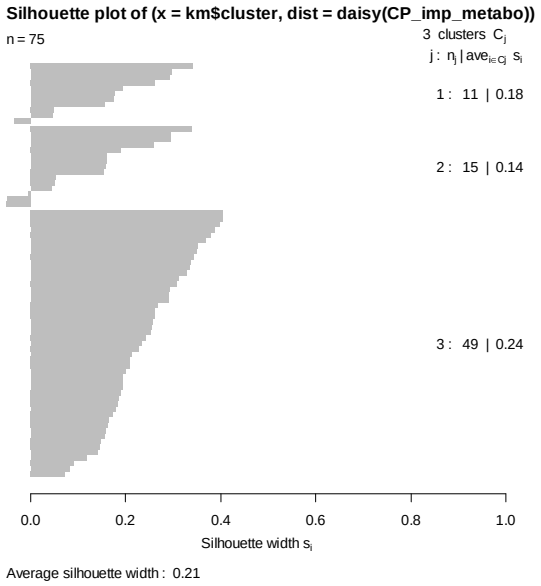


Figure 36 : Graphique des silhouettes, méthode K-means pour le métabolisme.

L'indice de silhouette n'atteint toujours pas un niveau acceptable pour conclure à une classification de qualité.

Conclusion :

75 lignées de souris ont ici été étudiées sur une cinquantaine de variables au total. Les résultats ainsi obtenus ne sont donc pas de qualité optimale. Comme on peut le voir sur la Heat-Map, aucun profil ne se distingue réellement et les différents graphiques de l'indice de silhouette ne concluent pas à des classes homogènes, bien séparées les unes des autres. On voit cependant que lorsque l'on réduit le nombre de variable, l'indice des silhouettes augmente. Un nombre plus élevé de lignées apporterait bien plus d'information et permettrait de créer des classes plus robustes. Il est donc difficile de conclure sur la méthode de classification la plus appropriée dans notre cas.

Conclusion

Lorsque l'on dispose d'un très grand tableau de données, les méthodes d'analyses classiques sont difficiles à mettre en œuvre. L'analyse des données, et plus précisément l'Analyse en Composantes Principales, a donc permis de réduire le nombre de paramètres tout en gardant un maximum de l'information contenue dans les données.

Les analyses qui ont suivies ont permis de décrire les phénotypes de chaque lignée mutante étudiée.

Enfin les classifications finales ont mis en évidence des similarités entre les différentes lignées. Il est cependant nécessaire d'améliorer la qualité de ces classifications pour conclure à une réelle classification. Un nombre de lignées plus important aurait certainement amélioré nos résultats.

Cette première étude a permis de développer une méthode d'analyse qui reste cependant à valider grâce à des interprétations biologiques bien plus poussées.

Discussions et Perspectives

Pour le moment, seules les données de l'Institut Clinique de la Souris ont été étudiées. Trois autres centres de phénotypages ont également récolté des résultats sur un bon nombre d'autres lignées mutantes. Une étude globale est envisagée afin d'améliorer la qualité des résultats. Il faudra cependant adapter la méthode afin de prendre en compte un effet groupe non négligeable. Les souris ne sont en effet pas exactement dans le même environnement selon le centre où elles se trouvent. La nourriture peut avoir un effet significatif sur le poids de ces souris et il a été montré que le poids a un rôle significatif sur l'état global de l'animal.

Dans ce projet, les données qualitatives n'ont pas été analysées. Il pourrait être intéressant de les intégrer dans l'étude au moyen d'une analyse de Hill et Smith en remplacement de l'Analyse en Composantes Principales qui prend en compte uniquement les variables quantitatives.

Les données manquantes présentes dans notre tableau de données ont certainement eu un impact sur la qualité de nos résultats malgré une méthode d'imputation multiple connue pour ses résultats convaincants. Il faudrait donc réduire celles-ci au maximum dès la récolte des données et trouver une méthode d'imputation plus efficace pour ce type de données.

Les souris contrôles sont récupérées par groupe de 7, un effet groupe peut donc exister. Cet effet n'ayant pas été pris en compte dans notre étude, il serait intéressant de l'inclure lors des prochaines analyses.

Bibliographie

Benjamini, I., Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society*, **57**, 289-300.

Bertrand, M., Bertrand F. (2006/2007). Mesures d'associations.

Calinski, T., Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics* **3**, 1-27

Duby, C., Robin, S. (2006). Analyse en Composantes Principales.

Fisher, R. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance, P*hilosophical Transactions of the Royal Society of Edinburgh*, **52**, 399-433.

Fisher, R.A. (1924). The distribution of the partial correlation coefficient*, Metron* **3** (3–4), 329–332.

Hotteling, H. (1933). Analysis of a Complex of Statistical Variables with Principal Components, *Journal of Educational Psychology,* **24**(6) 417-441.

Kruskal, W., Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association*, **47** (260), 583–621.

Mallon, A-M., Blake, A., Hancock, J. M. (2008). EuroPhenome and EMPReSS: online mouse phenotyping resource, *Nucleid Acids Research*. **36**, D715-D718.

Rousseeuw, P. J., (1987). Silhouettes: A graphical aid to the interpretation and validation of clusters analysis, *Journal of Computational and Applied Mathematics*, **20**, (53-65).

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, Wiley, New-York.

Rubin, D. B. (1988). An overview of multiple imputation, *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, 79-84.

Steel, R. (1959). A multiple comparison rank sum test: treatment versus control, *Biometrics*, **15** (4), 560-572.

Annexes

Liste des gènes et leurs identifiants

Gène	Identifiant	Gène	Identifiant	Gène	Identifiant	Gène	Identifiant
Aste1	Aste1	Gar1	EPD0117_3_G11	Ddx52	EPD0102_1_B10	Rc3h1	HEPD0546_1_F01
Snx5	Snx5	1700052N19Rik	EPD0175_3_D06	E330016A19Rik	HEPD0516_3_C04	Nfya	EUCJ004_F10
Yipf1	Yipf1	Cacnb4	HEPD0507_8_G08	Orc4l	EPD0087_3_G06	Mtmr4	EPD0211_4_B08
Gpt2	Gpt2	B4galnt3	EPD0140_5_G02	Zfp13	HEPD0528_4_A08	Pex16	HEPD0537_2_B03
Xbp1	EPD0038_2_B10	2310067B10Rik	EPD0215_1_D09	Eif2b5	EPD0116_1_C09	Aldh2	EPD0089_4_F11
Sfrs4	EPD0039_1_C07	H6pd	EPD0173_1_C12	Slc25a38	EPD0227_7_D02	Setmar	EPD0242_4_B03
Clk1	EPD0065_2_E04	Mrpl10	HEPD0501_1_F05	Stk30	HEPD0530_7_A11	Tmc8	EPD0227_3_A12
Agxt2l2	EPD0021_1_F05	Mrps35	EPD0089_3_A10	Pacs2	HEPD0529_7_C09	Pla2g4f	HEPD0527_2_D03
Ngfr	EPD0059_4_D03	Limk2	EPD0040_1_B07	Mrpl54	EPD0176_5_A09	Dpy19l3	EPD0227_3_B02
Jmjd5	EPD0065_5_A04	Ruvbl1	EPD0089_2_G11	Trib3	HEPD0510_8_G06	Hcn4	HEPD0547_4_B11
Fabp3	EPD0061_1_C10	Fpgs	HEPD0507_7_C03	Bloc1s1	HEPD0550_3_B11	Spns1	EPD0227_1_A03
Ift20	EPD0066_2_F05	Itgb5	EPD0114_3_B05	Ascc3	HEPD0547_1_C09	Gsta3	HEPD0539_2_G05
Grhl3	EPD0039_3_C02	Ttll1	EPD0066_2_A08	Col5a1	HEPD0509_7_C01	Gmfg	HEPD0535_6_B06
C1qbp	EPD0038_2_B07	Btk	HEPD0522_1_A11	Bag1	HEPD0505_2_B06	Slc1a7	HEPD0537_5_G04
Fam63a	EPD0057_1_H01	Flot2	EPD0113_4_B07	Mat2a	HEPD0547_3_D11	Patl2	HEPD0535_1_C03
Prpsap2	EPD0021_1_F09	Fabp12	EPD0146_4_H09	Nfasc	HEPD0529_1_B03	Ttll4	EPD0057_2_C02
Ndufs3	EPD0057_2_A08	Cops4	HEPD0509_6_A04	Fkbp10	EPD0142_1_A09	Slc38a10	EPD0023_1_F07
Ube2b	EPD0038_4_D01	Xiap	HEPD0523_5_A03	Vat1l	HEPD0518_7_E08	Arhgef4	EPD0135_1_A05
Tmem68	EPD0117_1_C02	Rnase10	HEPD0528_6_A02	Mosc2	HEPD0546_2_F03	Kdm4c	EPD0033_3_F04
Tcf7	EPD0023_2_D12	Drg2	EPD0215_4_A11	Xbp1	EPD0038_2_B10	Fam134c	EPD0037_1_B03
Crtc1	EPD0142_4_B09	Sox13	EPD0226_6_B08	Entpd1	EPD0156_1_B01	Fkbp9	HEPD0527_4_D08
Pik3c3	EPD0103_4_D07	Fgl1	HEPD0540_3_F06	Mtrf1	HEPD0530_4_B02	Acap3	HEPD0508_4_G09
Aldh3a2	EPD0060_2_D02	Mib2	EPD0173_1_F02	Otop3	EPD0197_3_D09	Heatr3	HEPD0527_5_A04
Sept_08	EPD0060_3_G04	Ift122	EPD0155_2_E05	Mapre3	EPD0144_4_G05	Tmod3	HEPD0547_3_H08
Ephx1	EPD0518_6_B02	Arcn1	HEPD0516_2_B11	Elk4	EPD0100_4_A06	cd9	HEPD0538_5_H11
B9d1	EPD0060_2_H09	Pgd	EPD0117_4_C03	2900026A02Rik	HEPD0525_3_B05		
Bspry	EPD0113_4_D09	Nfxl1	EPD0100_5_F06	Tnfaip1	EPD0037_1_E07		
Tnfrsf1b	EPD0155_1_B07	Lhx1	EPD0243_4_E03	Afm	EPD0288_3_A08		

Figure 1: Gène / Identifiant

Description du déroulement des différents Pipelines

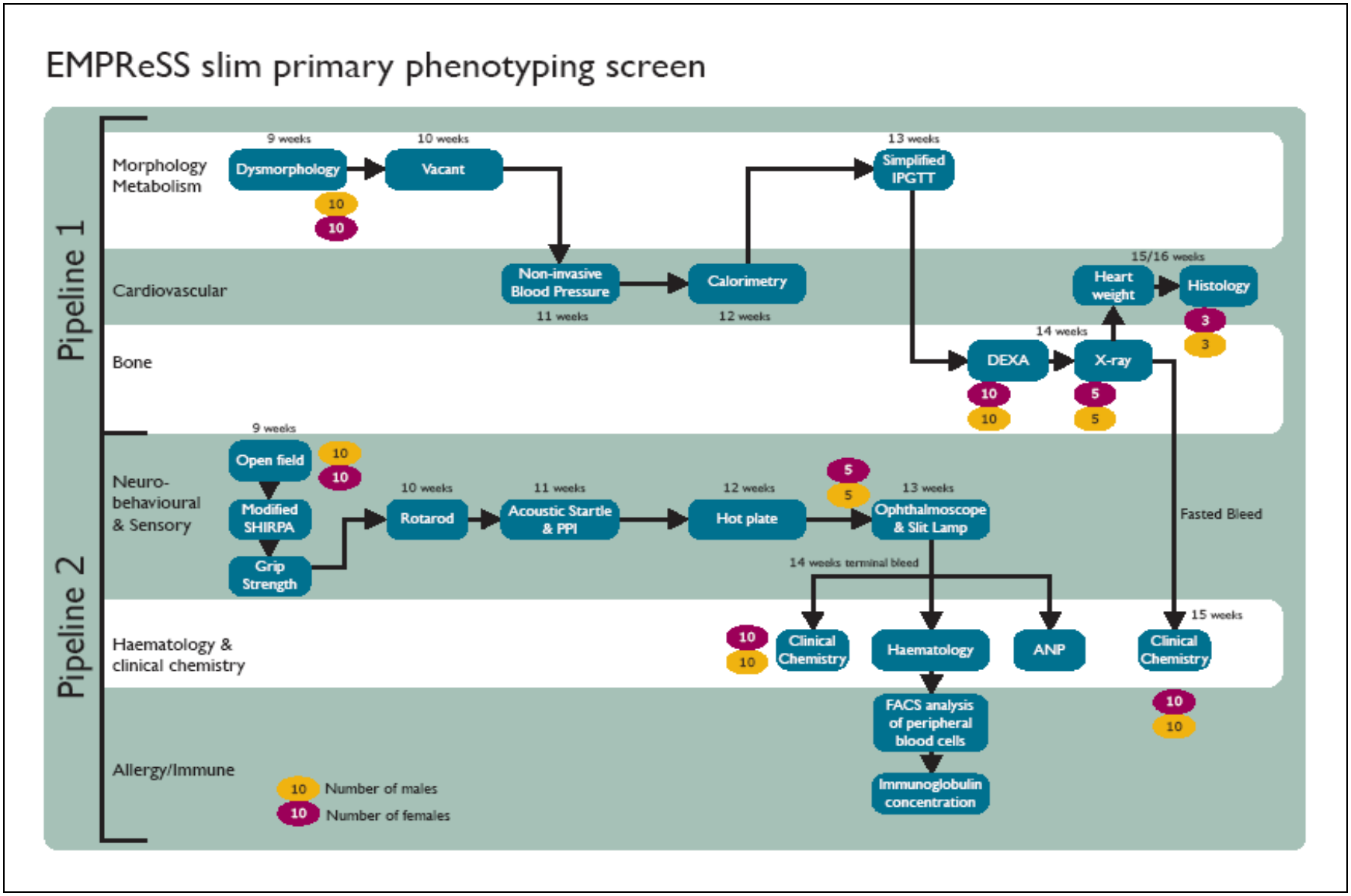


Figure 2: Déroulement Pipeline 1&2

ACP : Cercles de corrélations

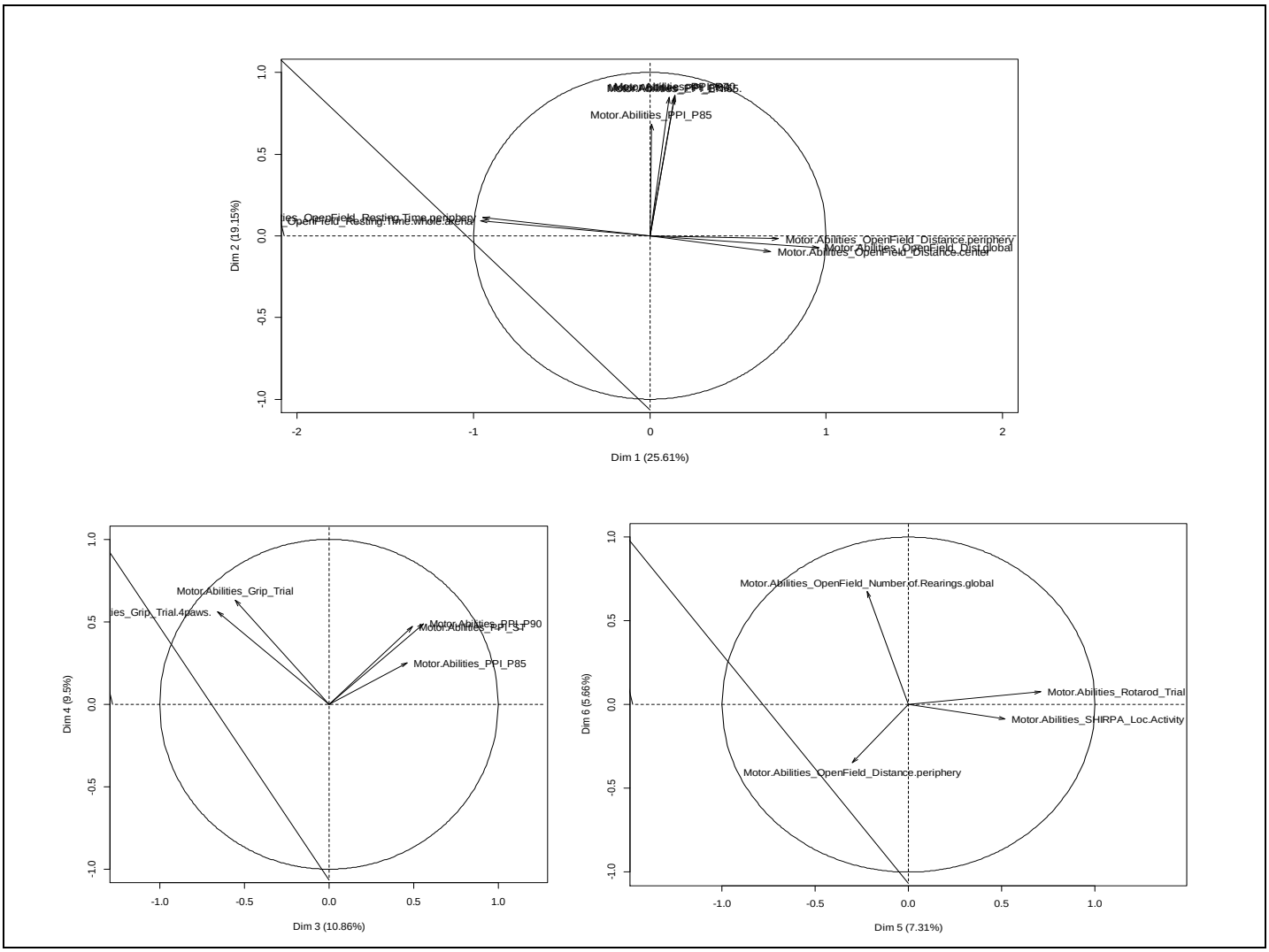


Figure 3: Cercles de corrélations correspondants aux capacités motrices

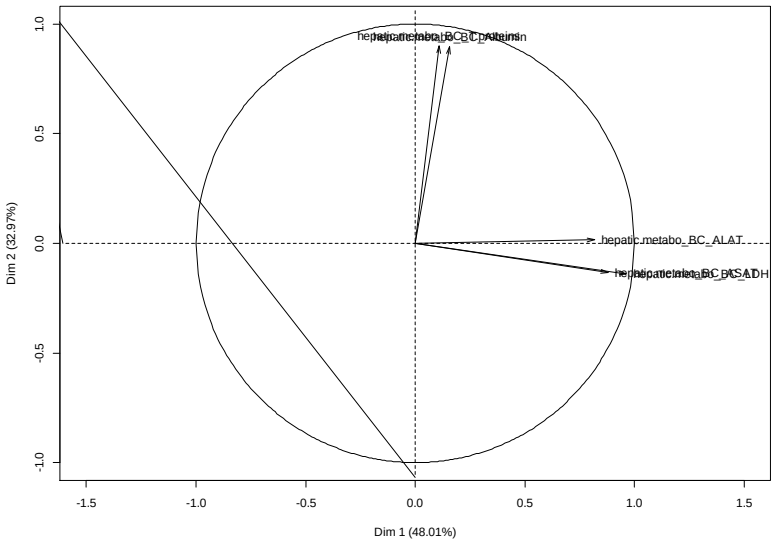


Figure 4 : Cercle de corrélations correspondant au métabolisme hépatique.

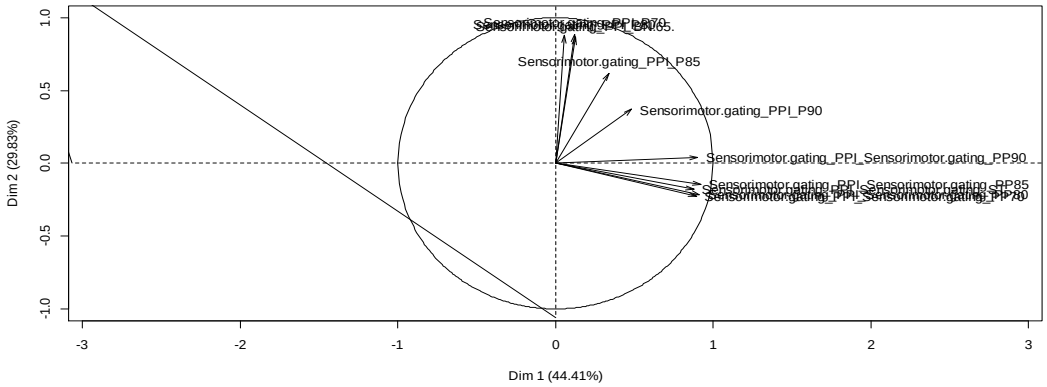


Figure 5 : Cercle de corrélations correspondant à la Schizophrénie.

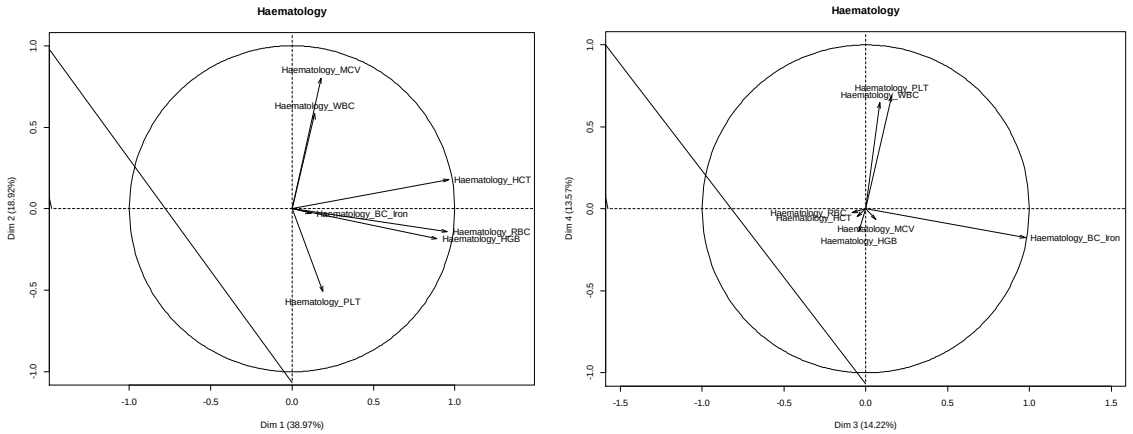


Figure 6 : Cercles de corrélations correspondants à l'Hématologie.

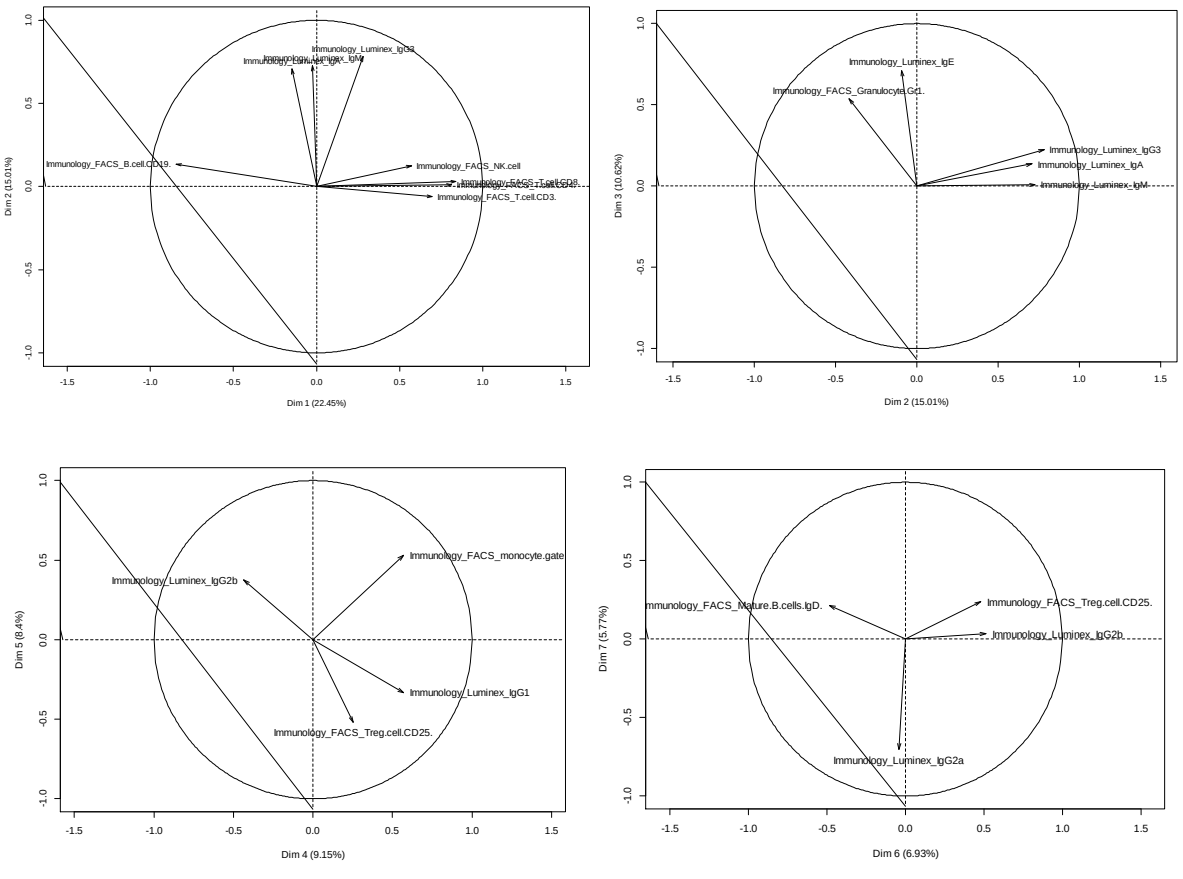


Figure 7 : Cercles de corrélations correspondants à l'immunologie.

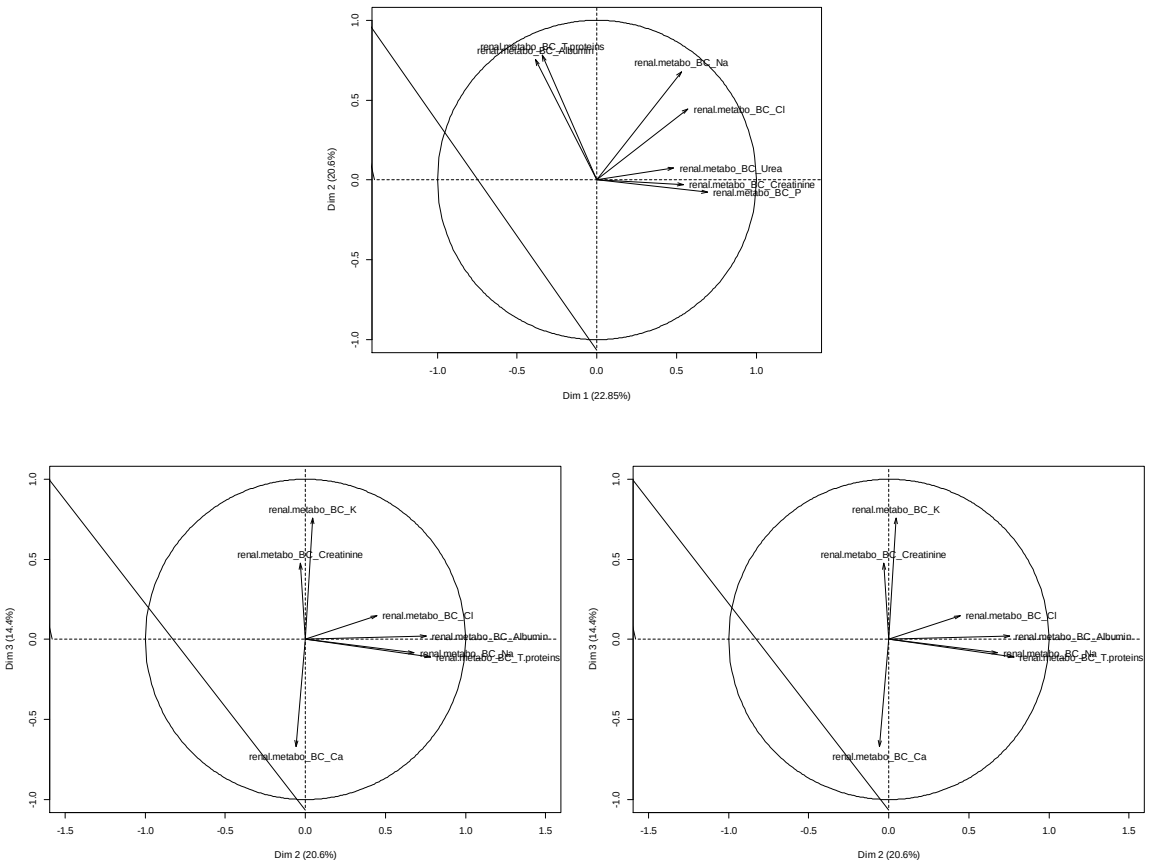


Figure 8 : Cercles de corrélations correspondants au métabolisme rénal.

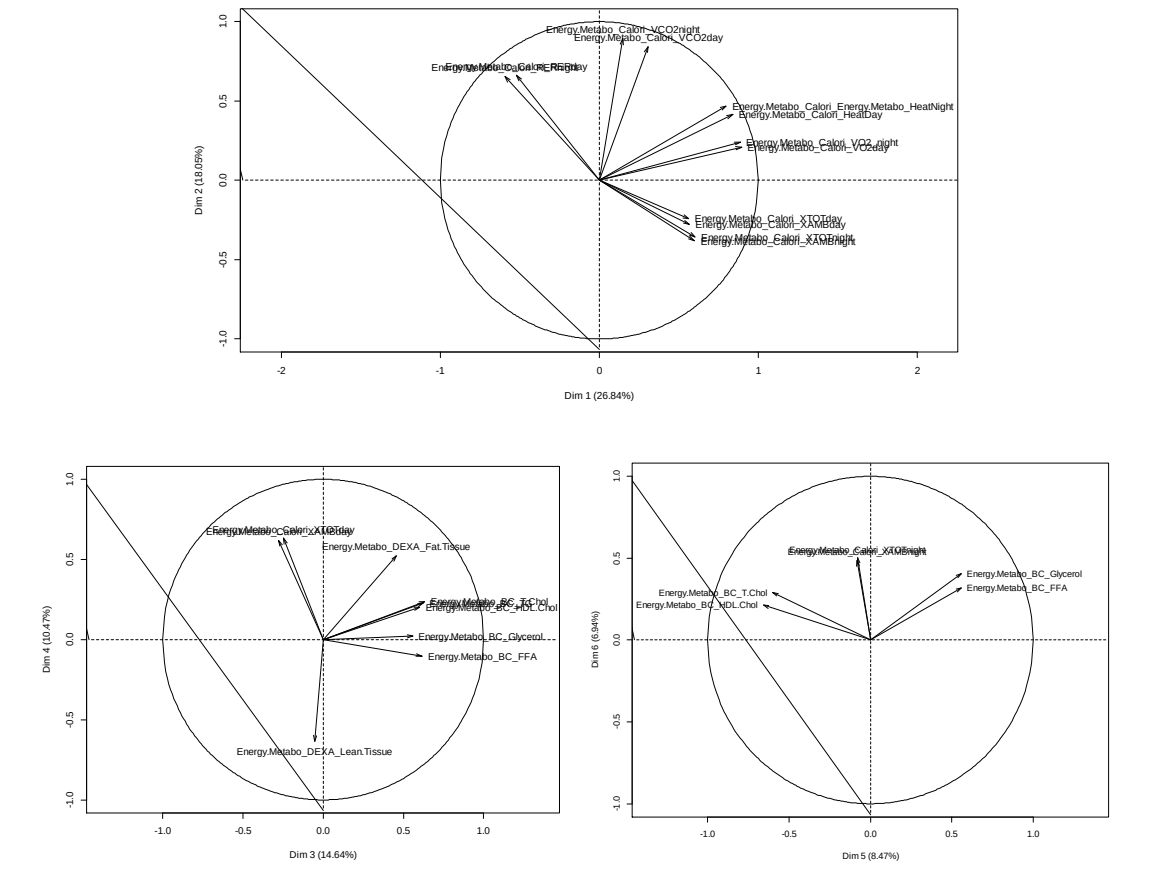


Figure 9 : Cercles de corrélations correspondants au métabolisme énergétique