INSTITUT de MATHEMATIQUES de TOULOUSE

UNIVERSITÉ TOULOUSE III PAUL SABATIER — Université de Toulouse

Rapport de stage

Master 1 MAPI$^3$

---

# Intégration de données omiques multi-niveaux pour expliquer l'efficacité alimentaire chez les agneaux

---

INRAE

la science pour la vie, l'humain, la terre


24 Chemin Borde Rouge 31320 Auzeville Tolosane

Tuteurs : Nathalie VIALANEIX et Jérôme MARIETTE

Professeur encadrant : Luca AMODEI

**Annaleah JOHANNY**


03 Mai 2021 — 31 Août 2021

# Remerciements

Je tiens à remercier mes encadrants de stage Nathalie VIALANEIX et Jérôme MARIETTE, pour leur disponibilité et la confiance qu'ils m'ont accordée tout au long de mon stage. Leur implication m'a permis d'approfondir mes connaissances et de développer mes compétences en analyses de données.

Je remercie l'université Toulouse III - Paul Sabatier de m'avoir donné l'opportunité de réaliser ce stage.

Enfin je remercie l'ensemble de l'unité MIAT pour son accueil chaleureux pendant ces 4 mois de stage.

# Table des matières

# 1 INRAe et l'unité MIAT

**L'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE)**

L'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE) est un organisme public de recherche scientifique, placé sous la double tutelle du ministère de l'Enseignement supérieur et de la Recherche, et du ministère de l'Alimentation, de l'Agriculture et de la Pêche. Il est constitué de 14 départements scientifiques, répartis sur 18 centres de recherche régionaux. INRAE organise ses recherches selon 5 grandes orientations scientifiques :

- répondre aux enjeux environnementaux et gérer les risques associés ;
- accélérer les transitions vers des systèmes agricoles et alimentaires agroécologiques en tenant compte des enjeux économiques et sociaux ;
- une bioéconomie basée sur une utilisation sobre et circulaire des ressources ;
- favoriser une approche globale de la santé ;
- mobiliser la science des données et les technologies du numérique au service des transitions.

Trois orientations de politique générale fournissent un cadre pour la réalisation des recherches selon les orientations scientifiques et pour la vie collective :

- Placer la science, l'innovation et l'expertise au cœur de nos relations avec la société pour renforcer notre culture de l'impact ;
- être un acteur engagé dans les sites universitaires en France et un leader dans les partenariats européens et internationaux ;
- la stratégie « Responsabilité Sociale et Environnementale » (RSE) : une priorité collective.

Il mène des recherches sur les thèmes de l'agriculture, l'alimentation, la sécurité des aliments, l'environnement et la gestion des territoires. Toutes ces recherches sont effectuées dans une perspective de développement durable.

**L'Unite de Mathematiques et Informatique Appliquées de Toulouse (MIAT)**

L'unité MIAT est chargée de mettre au point des méthodes mathématiques et informatiques et de les mettre à disposition d'INRAE, favorisant ainsi les collaborations entre départements. Le domaine de compétence de l'unité s'étend de la statistique, aux probabilités, à l'algorithmique, l'intelligence artificielle et aux sciences de la décision. L'unité comporte, depuis janvier 2011, deux équipes de recherches thématiques :

- SCIDYN (Simulation, Contrôle et Inférence de Dynamiques Agro-environnementales et Biologiques) : modélisation des systèmes complexes dans les champs de l'agriculture, de l'environnement, de l'analyse des risques alimentaires et des procédés industriels.
- SaAB (Statistique et Algorithmique pour la Biologie) : développement de méthodes relevant des mathématiques, de la statistique et de l'informatique destinées à l'exploitation de données de génomique et de post-génomique. Mon activité est rattachée à cette équipe.

L'unité s'appuie aussi sur l'activité de trois plateformes :

- Plateforme GENOTOUL : Plateforme bioinformatique du GIS GENOTOUL, dont l'activité est centrée sur la bioinformatique et l'analyse de séquences.
- Plateforme RECORD (RÉnovation et COORDination de la modélisation des cultures pour la gestion des agro-écosystèmes) : Plateforme issue du partenariat des départements Environnement et Agronomie (EA) et MIA. Elle vise à offrir un cadre et des outils informatiques communs aux modélisateurs des différentes disciplines (agronomie, bioclimatologie, sciences de gestion, mathématiques, ...) pour la modélisation et la simulation des systèmes de culture.
- Plateforme SIGENAE (Système d'Information des GENomes des Animaux d'Elevage). Elle se compose d'ingénieurs en bio-informatique qui accompagnent les biologistes des départements « animaux » (Génétique Animale, Physiologie Animale et Système d'Elevage, Santé Animale) d'INRAE dans le traitement de leurs données à haut débit.

Ce stage a été encadré par Mme Nathalie VIALANEIX, Directrice de Recherche au sein de l'unité MIAT, dans l'équipe SaAB, ainsi que par Mr Jérôme MARIETTE, Ingénieur d'Étude en bioinformatique dans l'équipe GENOTOUL Bioinfo.

# 2 Contexte et objectif du stage

## 2.1 Projet MILAGE

Ce stage s'inscrit dans le cadre du projet MILAGE (financement métaprogramme « Méta-omiques et écosystèmes microbiens (MEM) » d'INRAE). L'unité portant le projet est GenPhyse avec laquelle mon unité (MIAT) collabore. Les deux biologistes de l'unité GenPhyse avec qui j'ai été en contact pendant mon stage sont : Flavie TORTEREAU et Christel MARIE-ETANCELIN.
L'objectif du projet est de comprendre et prédire l'efficacité alimentaire chez les agneaux. Celle-ci se définit comme la capacité à bien grandir/grossir avec une alimentation réduite. En effet, l'élevage d'agneaux dont l'efficacité alimentaire est bonne permet de réduire le coût de l'alimentation des animaux tout en ayant la même production finale. Par conséquent, cela réduit également l'impact environnemental de la production de nourriture.

## 2.2 Plan expérimental

Un dispositif a été mis en place permettant de collecter plusieurs types de données sur deux lignées divergentes : RFI+ et RFI-. Les animaux de lignée RFI- ont une bonne efficacité alimentaire contrairement aux animaux de la lignée RFI+ : un agneau de lignée RFI- consommera moins de nourriture qu'un agneau de lignée RFI+ tout en ayant la même vitesse de croissance ainsi que la même production finale. Pour obtenir les deux lignées, les parents et grands-parents des animaux concernées avaient été préalablement sélectionnés : ceux de la lignée RFI- étaient parmi les animaux les plus efficaces parmi un ensemble d'animaux sur lesquels la valeur génétique rfi avait été calculée (et inversement pour deux de la lignée RFI+). Au fur et à mesure des générations sur lequel ce processus est répété, on obtient des animaux (descendants des animaux de la génération précédente, sélectionnés sur le caractère d'intérêt qui est ici l'efficacité alimentaire) de plus en plus extrêmes et différents entre les deux lignées. On appelle ce type d'expériences des « lignées divergentes ».

Les mesures ont été réalisées sur des animaux de deux années différentes (2018 et 2019). Les animaux de 2019 sont, en grande partie, des descendants des animaux de 2018. Quelques semaines après leur naissance les agneaux suivent un premier régime alimentaire contrôlé (DAC) pendant lequel une première série de mesures est effectuée. Durant cette période, les animaux sont nourris à base de concentré. À l'issue de ce régime alimentaire, seuls les animaux extrêmes de chaque lignée sont gardés pour la deuxième phase d'alimentation : le DAF. Une nouvelle série de mesures a été réalisée lors de ce second régime alimentaire (*voir figure 2.1* ).

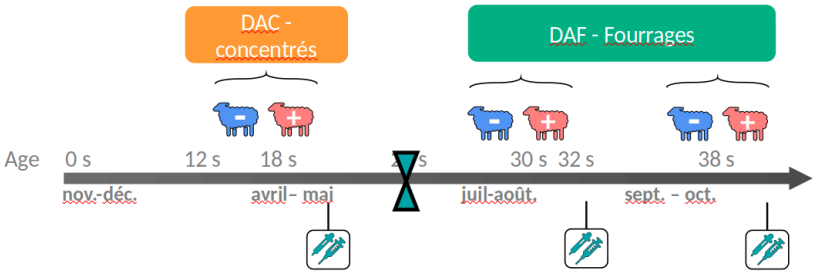



FIGURE 2.1 – Plan d'expérience.

Les données collectées représentent des informations génotypiques, métabolomiques, métagénomiques et phénotypiques pour les deux types de régimes alimentaires (DAC et DAF). De manière plus précise, les données correspondent à des mesures réalisées sur les mêmes animaux :

- performances en élevage,
- métabolites du sang,
- métabolites du rumen (toutes obtenues par technique haut débit de résonance magnétique nucléaire $^{1}$H RMN),
- microbiote du rumen (par séquençage haut-débit basé sur le gène 16S),
- mesures infra-rouge des fèces,
- génotypes (obtenus par une technique de puces).

## 2.3 Problématique du stage

Le but du stage est de comparer les deux lignées (les animaux d'une même lignée seront analysés ensemble mais les animaux en DAC et en DAF seront analysés séparément) en intégrant les informations provenant de toutes les données.

Les objectifs de ce stage sont de sélectionner des variables explicatives des deux lignées, pour cela des modèles prédictifs intégrant ces différents omiques ont été utilisés et pour finir de trouver les relations entre les différents jeux de données.

## 2.4 Organisation

Mon travail s'est organisé selon un cycle hebdomadaire de 5 jours et un volume horaire de 35h. Une réunion avec mes encadrants de stage avait lieu chaque semaine le Mardi matin. Cela me permettait d'exposer l'avancée de mes analyses et de discuter des différents résultats obtenus. À l'issue de cette réunion nous définissions une direction à suivre pour la suite de mes travaux.

L'ensemble des analyses a été réalisé à l'aide du langage R et de l'interface graphique RStudio. Plus précisément, j'ai rédigé un ensemble de scripts RMarkdown permettant de générer des rapports de manière dynamique en mélangeant textes mis en forme et résultats produits par du code R, voir un exemple de rapport en Annexe 6. Le partage de mes travaux avec mes encadrants s'effectuait via Git, un système de contrôle de version, auquel j'ai du me former pendant mes premières semaines de stage.

Au cours de mon stage, il m'a également été demandé de présenter mes analyses aux biologistes du projet MILAGE (Flavie TORTEREAU et Christel MARIE-ETANCELIN) afin de discuter avec elles des résultats obtenus.

# 3 Méthodes

Dans cette section, je vais décrire et présenter l'ensemble des données ainsi que les méthodes de pré-traitements et d'analyses qui ont été effectuées tout au long de ce stage.

## 3.1 Présentation des données

L'ensemble des données correspond à des mesures réalisées à plusieurs niveaux de l'échelle du vivant : microbiote du rumen (par séquençage haut-débit basé sur le gène 16S), performances en élevage, métabolites du sang, métabolites du rumen (toutes obtenues par technique haut débit de résonance magnétique nucléaire $^{1}$H RMN), mesures infra-rouges des fèces, génotypes (obtenus par une technique de puces). Madame Vialaneix et Monsieur Mariette ont nettoyé et mis au propre les données en amont de mon stage. Ainsi, seuls les animaux pour lesquels nous avions l'ensemble des mesures ont été gardés pour les analyses. Au total 196 animaux sont conservés pour le premier régime alimentaire (DAC) et 120 pour le deuxième régime (DAF).

Durant ces quatre mois de stage, j'ai pu analyser quatre jeux de données en DAC et en DAF : les performances en élevage, les métabolites du plasma, les métabolites du rumen ainsi que le microbiote du rumen.

Les performances en élevage correspondent à des mesures qui évaluent le physique des animaux : poids en début milieu et fin de régime, poids à âge type (145 jours ici), épaisseur de muscle et de gras en milieu et fin de régime, consommation moyenne et résiduelle de nourriture ainsi que le gain de poids moyen par jour.

Les deux jeux de données contenant les informations sur le métabolome du rumen et du plasma sont des tables donnant la concentration des métabolites présents pour chaque individu. Les métabolites sont soit produits à partir d'autres métabolites grâce à des réactions chimiques, soit ils proviennent de l'extérieur en étant, par exemple, ingérés. L'ensemble des métabolites constitue le métabolome. Nous avons les concentrations de 40 métabolites dans le rumen en DAC et 19 en DAF, contre 26 métabolites dans le plasma en DAC et 19 en DAF.

Le jeu de données du micriobiote est une table de comptages contenant le nombre de fois qu'un OTU est observé pour chaque échantillon. Un OTU (Unité Taxonomique Opérationnelle) peut être défini par un regroupement de micro-organismes d'une même espèce dont les séquences d'ARNr 16S présentent une similitude de plus de 97,5%.

Un fichier *design* contenant le plan d'expérience m'a également été mis à disposition. Les informations présentes dans ce fichier sont les suivantes :

— l'année,
— le lot,
— la lignée,
— la valeur génétique rfi,
— la date, l'heure et l'ordre des prélèvements dans le rumen,
— l'âge au jour des prélèvements,
— la plaque, la profondeur et le run de séquençage (pour les données de microbiote).

Afin d'optimiser leur accès à la nourriture, les agneaux ont été séparés en plusieurs lots, élevés dans des lieux différents : 6 en 2018, 5 en 2019 en DAC (*voir figure 3.1* ) et 4 en 2018, 4 en 2019 en DAF (*voir figure 3.2* ). Les lots ont été réalisés en fonction du poids des animaux pour garantir la bonne nutrition des plus petits. Lors du second régime alimentaire (DAF) la valeur génétique rfi a également été prise en compte pour la séparation en lots : les extrêmes de chaque lignée sont séparés des animaux « intermédiaires » (*voir figure 3.3* ).
Nous avons au total 11 lots différents en DAC et 8 en DAF directement liés au poids et à l'efficacité alimentaire des animaux.

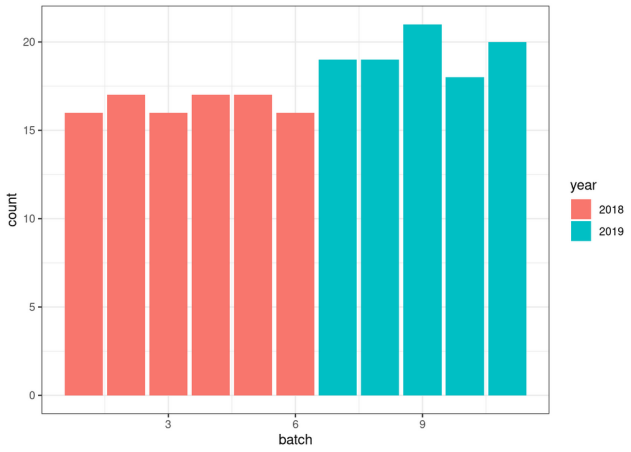



FIGURE 3.1 – Lots DAC en fonction de l'année.

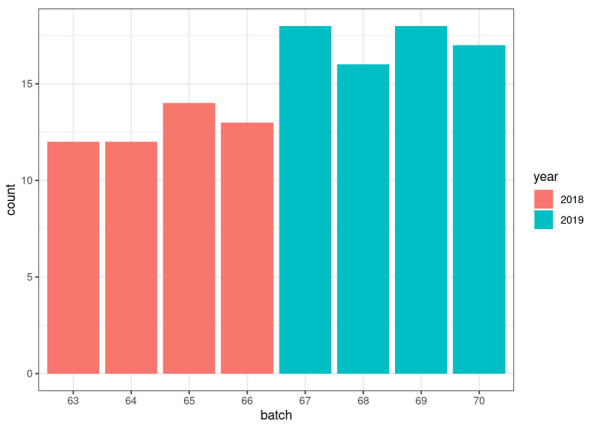



FIGURE 3.2 – Lots DAF en fonction de l'année.

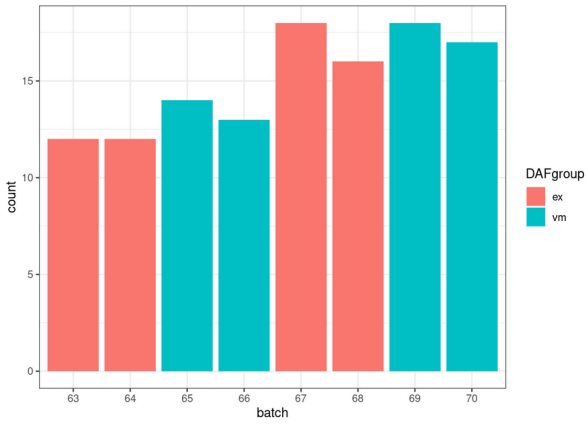



FIGURE 3.3 – Lots DAF en fonction du groupe.

## 3.2 Méthodes de pré-traitement de données

### 3.2.1 Performances en élevage et métabolome : données numériques

Afin d'observer au mieux les effets de la lignée sur les différentes données il faut s'assurer qu'aucun effet technique lié au plan d'expérience n'influence les données. Pour cela, j'ai utilisé une Analyse en Composantes Principales (ACP) qui m'a permis d'observer les éventuels effets indésirables.

L'ACP est une méthode de réduction de dimension qui va permettre de transformer des variables très corrélées en nouvelles variables décorrélées les unes des autres. Autrement dit, on cherche à définir de nouvelles variables, combinaisons linéaires des variables initiales, qui feront perdre un minimum d'informations. Il s'agit en fait de résumer l'information qui est contenue dans un jeu de données en variables synthétiques appelées composantes principales. Cela va permettre de représenter le jeu de données, de manière simple, sur 2 ou 3 axes. La réduction de dimension entraîne obligatoirement une perte d'information et celle ci est minimisée par optimisation d'un critère d'inertie reproduite.

Les ACP ont été exécutées avec la fonction *PCA* du package *FactoMineR*. Cette analyse multi-variée a été réalisée sur les 6 jeux de données : performances DAC et DAF, métabolome du rumen DAC et DAF et métabolome du plasma DAC et DAF, croisée avec les variables du plan d'expérience.

Si un effet technique lié au plan d'expérience est détecté grâce à l'ACP, celui-ci est corrigé avec la fonction *ComBat* du package *sva*. Cette correction se fait par modèle linaire multivarié : les résidus sont récupérés une fois les effets techniques supprimés. Par la suite, les analyses sont réalisées sur les résidus.

### 3.2.2 Microbiote : données de comptages

Les données du microbiote sont des données de comptage d'OTUs qu'on peut assimiler à une espèce bactérienne. Le nombre total de lectures à l'issue du séquençage, aussi appelé profondeur de séquençage, peut varier d'un échantillon à l'autre. Cette différence empêche la comparaison directe de plusieurs échantillons. Un pré-traitement supplémentaire est obligatoire afin d'analyser correctement ce type de données.

La première étape a été de filtrer les OTUs et de supprimer celles dont le nombre d'apparitions total est inférieur à un certain seuil. Ce dernier est calculé par rapport au nombre total d'OTUs comptés. Ce filtre permet d'éviter des analyses erronées par des variables peu présentes dans le jeu de données. Dans un deuxième temps, j'ai appliqué une transformation CLR (Log-Ratio Centrée). Cette stratégie consiste à normaliser et projeter les comptages dans un espace euclidien :

$$y'_{ij} = \log \frac{y_{ij}}{\sqrt[p]{\prod_{k=1}^{p} y_{ik}}}.$$

Cela permet de traiter le problème des données compositionelles. Le données compositionelles sont obtenues après la normalisation. Ce sont des données positives ou nulles dont la somme des valeurs pour chaque individu est égale à 1 (ou 100, ou un nombre $\mathbf{k}$).
Pour appliquer cette transformation j'ai utilisé la fonction *logratio.transfo* avec le paramètre *logratio = "CLR"* du package *MixOmics*.

Après ce premier pré-traitement, j'applique des ACP comme précédemment afin d'identifier et corriger les éventuels effets techniques. Les corrections ont également été réalisées avec la fonction *ComBat* du package *sva*.

## 3.3 Méthodes d'analyse des données

Les deux mêmes analyses ont été réalisées sur chaque jeu de données : une PLS-DA ainsi que deux tests statistiques (test de Wilcoxon et test de Student).

**La PLS-DA**

La régression des moindres carrés partiels (PLS) généralise et combine les caractéristiques de l'ACP et de la régression multiple. Elle est notamment utilisée lorsque l'on veut prédire un ensemble de variables dépendant d'un nombre très grand de variables explicatives (prédicteurs) qui peuvent être fortement corrélées entre elles. La PLS-DA (Partial Least Squares Discriminant Analysis) est une extension de la régression PLS dédiée à la prédiction d'une variable catégorielle. C'est une analyse multivariée descriptive et prédictive. Je l'ai exécutée avec la fonction *plsda* du package *MixOmics*.

**Les tests statistiques**

Les tests d'hypothèses sont des procédures de décision entre deux hypothèses. Il s'agit d'une technique consistant à rejeter ou à ne pas rejeter une hypothèse statistique, appelée hypothèse nulle, en fonction d'un échantillon de données. Les tests appartiennent à la statistique inférentielle : à partir de calculs réalisés sur des données observées, on généralise les conclusions sur toute la population, en leur rattachant un risque de rejet à tord de l'hypothèse nulle, appelé p-valeur.
J'ai appliqué le test de Wilcoxon (non paramétrique) et le test de Student (paramétrique) avec un seuil de confiance à 95$\%$. Un test paramétrique est un test pour lequel une hypothèse paramétrique sur la distribution des données sous $H_0$ est faite. Un test non paramétrique, lui, ne nécessite pas d'hypothèse sur la distribution des données. Lorsque l'hypothèse de distribution est vérifiée, les tests non-paramétriques ne sont pas aussi puissants que les tests paramétriques. La puissance statistique d'un test est la probabilité de rejeter l'hypothèse nulle sachant que l'hypothèse nulle est incorrecte.

Comme de nombreux tests sont réalisés (un pour chaque variable) une correction des p-valeurs est nécessaire. Effectuer un test en boucle un grand nombre de fois augmente le nombre de p-valeurs calculées et par conséquent le risque de détecter des effets significatifs à tort. En effet, plus on réalise de tests plus la probabilité d'avoir un faux positif est élevée :

$$\mathbf{P}(\text{avoir au moins 1 faux positif si } H_0 \text{ est toujours vraie}) = 1 - (1 - \alpha)^N,$$

avec $N$ le nombre de tests réalisés et $\alpha$ le risque.
Pour $N$ supérieur à 75 cette probabilité est très proche de 1. J'utilise la méthode *Benjamini & Hochberg* pour corriger/pénaliser les p-valeurs à la fin de chaque série de tests. Cette procédure permet de contrôler le FDR (False Discovery Rate ou « taux de fausses découvertes »).
La correction est la suivante :

$$p' = \min(\frac{p \times N}{j}, 1),$$

où p est la p-valeur d'origine, $N$ le nombre de p-valeurs calculée au total, et $j$ le rang de la p-valeurs lorsque les p-valeurs sont rangées par ordre croissant.

# 4 Résultats

## 4.1 Pré-traitement des données

### 4.1.1 Performances en élevage et métabolome

Dans un premier temps, les effets techniques ont dû être identifiés et corrigés afin d'observer au mieux l'effet de la lignée sur les données.

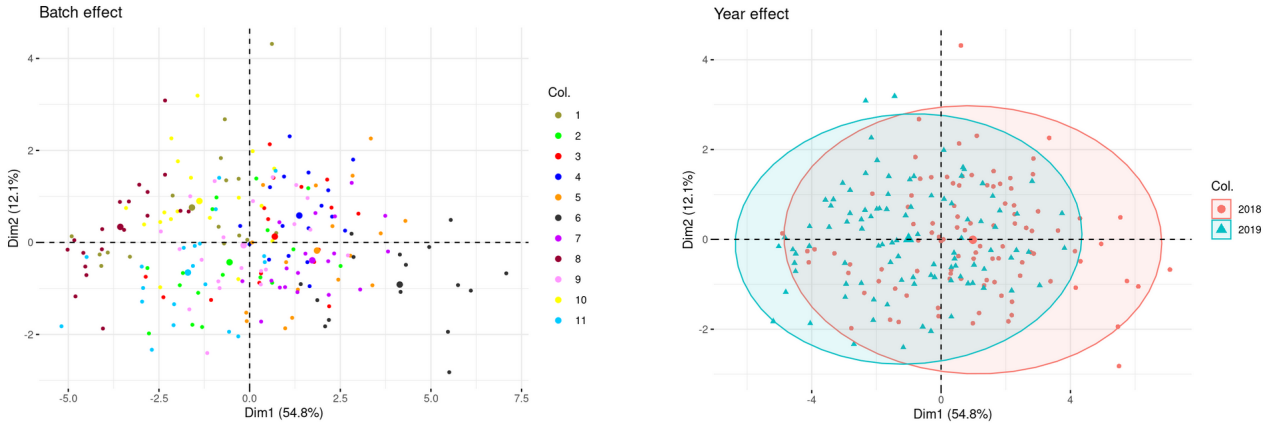



FIGURE 4.1 – Effet lot et effet année sur les performances DAC.

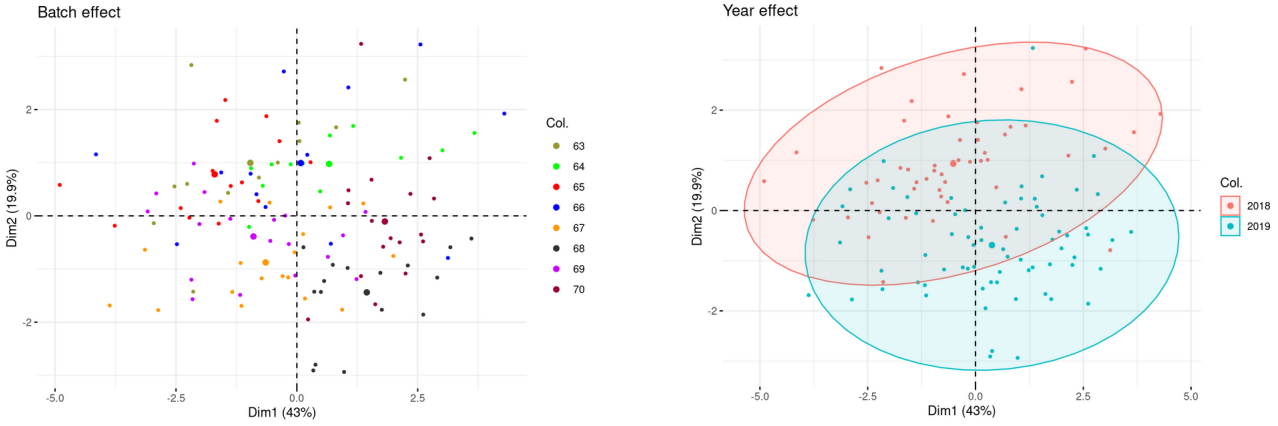



FIGURE 4.2 – Effet lot et effet année sur les performances DAF.

*(Seules les ACP sur les performances sont affichées. On obtient des résultats similaires pour le métabolome.)*

On observe un effet lot et un effet année marqués (*voir figure 4.1 et figure 4.2* ) quel que soit le régime alimentaire analysé (DAF et DAC). Il est nécessaire que ces effets soient corrigés afin d'analyser au mieux l'effet biologique étudié : la lignée.
Comme expliqué précédemment, les lots ont été réalisés en fonction du poids des animaux et de la valeur génétique rfi en DAF et seulement en fonction du poids en DAC.
Aussi, en DAC, c'est l'effet lot qui a été corrigé. En effet, la correction de cet effet permet également de corriger l'effet année. En revanche, pour les données DAF, la correction de l'effet lot entraînerait la perte d'une partie de la variabilité biologique qui nous intéresse (car les définitions des lots sont liées à la valeur génétique rfi qui est en lien avec la lignée). J'ai donc seulement corrigé l'effet année pour les données DAF.

La correction des données a été réalisée par modèle linéaire avec la fonction *ComBat* du package *sva*. Pour chaque jeu de données des ACP ont été faites afin de vérifier l'efficacité de la correction. L'ACP sur les données corrigées (*voir figure 4.3* ) montre que la correction avec la fonction *ComBat* a fonctionné et les effets indésirables ne sont plus visibles.

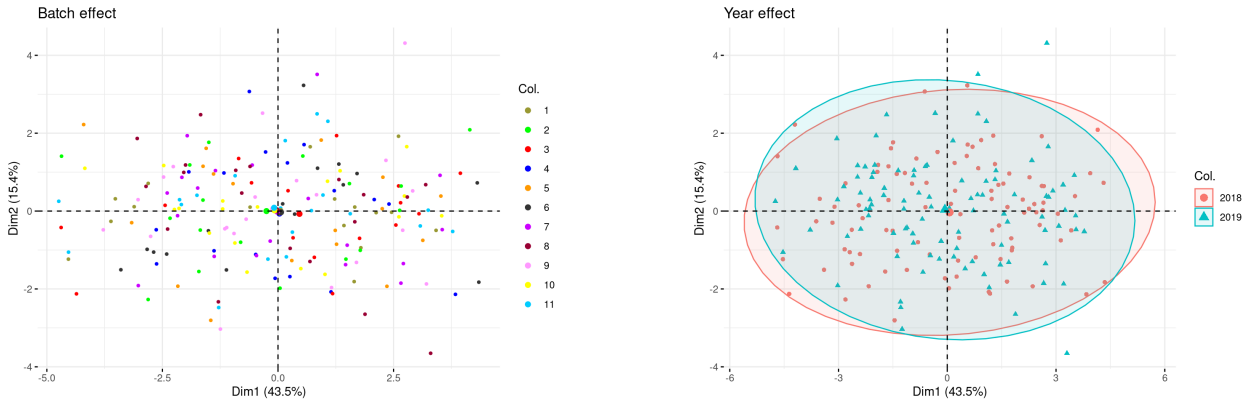



FIGURE 4.3 – Effet lot et effet année sur les performances DAC après correction de l'effet groupe.

*(Seules les ACP sur les performances corrigées DAC sont affichées. On obtient des résultats similaires pour les perfomances DAF et le métabolome)*

### 4.1.2 Microbiote

Une fois le prétraitement des données de comptage fait, j'ai réalisé des ACP comme pour les autres jeux de données.

#### En DAC

Un groupe d'individus se détache nettement des autres à gauche de l'axe 1 (*voir figure 4.4* ). Cependant aucune des variables du plan expérimental n'explique cette différence.

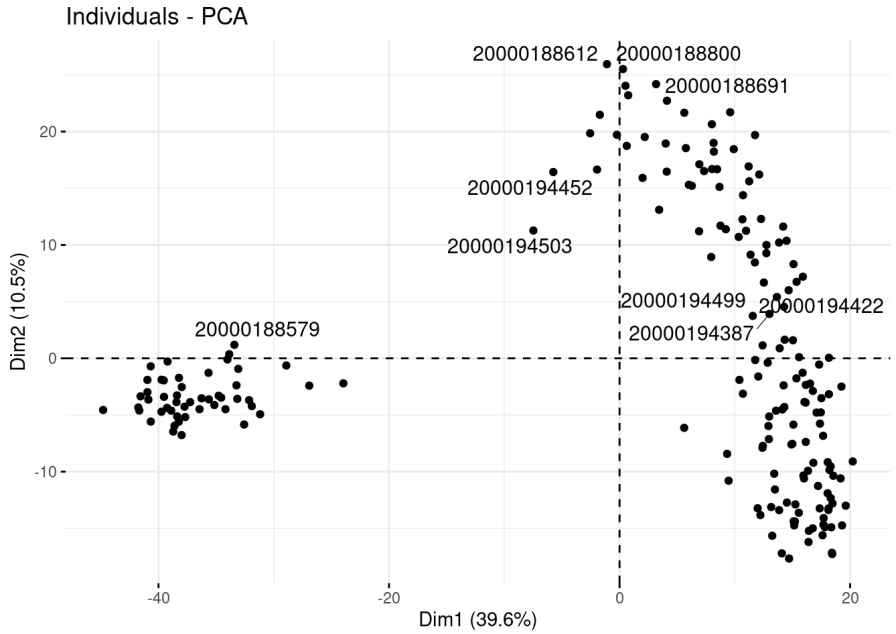



FIGURE 4.4 – ACP sur le microbiote en DAC.

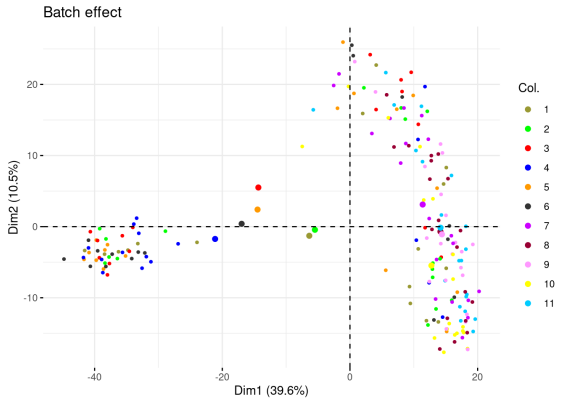



FIGURE 4.5 – Effet lot.

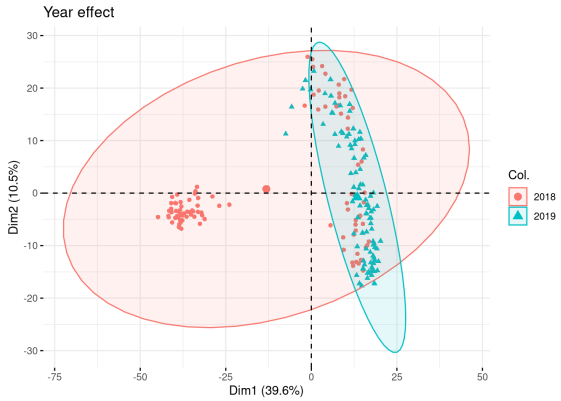



FIGURE 4.6 – Effet année.

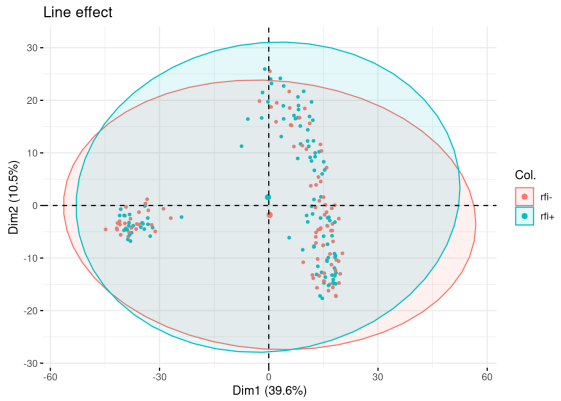



FIGURE 4.7 – Effet lignée.

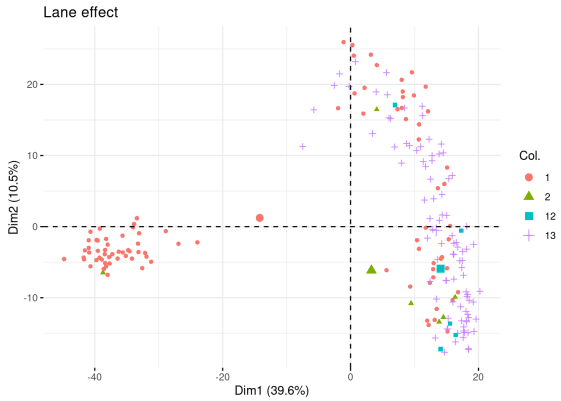



FIGURE 4.8 – Effet plaque de séquençage.

Aucune des variables lot (*voir figure 4.5* ), année (*voir figure 4.6* ), lignée (*voir figure 4.7* ) ou plaque de séquençage (*voir figure 4.8* ) ne semble être la cause de ce groupe d'individus.

Les animaux composant ce groupe ont été identifiés et supprimés de l'analyse du microbiote. Après suppression de ces individus, l'ACP montre un effet lot marqué, de la même manière que les analyses précédentes. Une correction de l'effet lot a donc de nouveau été faite avec la fonction *ComBat*. Comme voulu, les effets indésirables ne sont plus visibles (*voir figures 4.9 et 4.10* ).

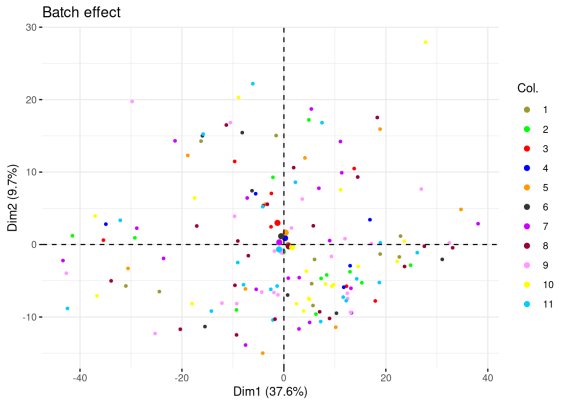



FIGURE 4.9 – Effet lot après correction.

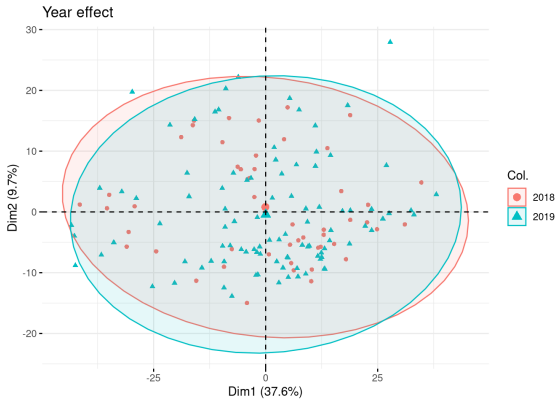



FIGURE 4.10 – Effet année après correction.

**En DAF**

Sur l'ACP nous remarquons 4 animaux atypiques (*voir figures 4.12 et 4.13* ), comme précédemment ces individus sont supprimés.

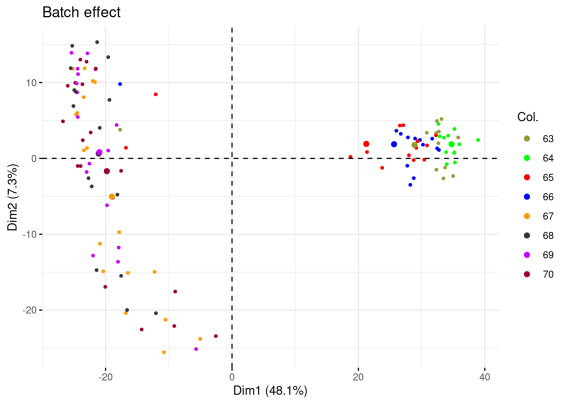



FIGURE 4.11 – Effet lot.

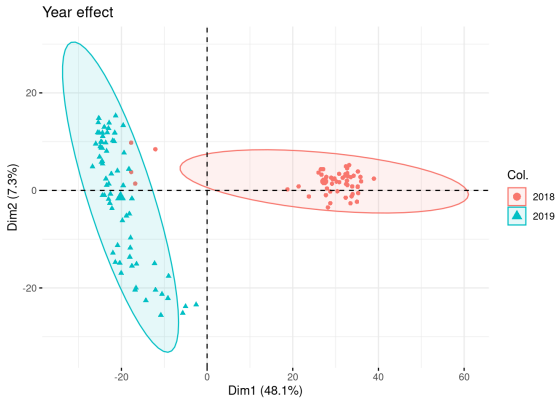



FIGURE 4.12 – Effet année.

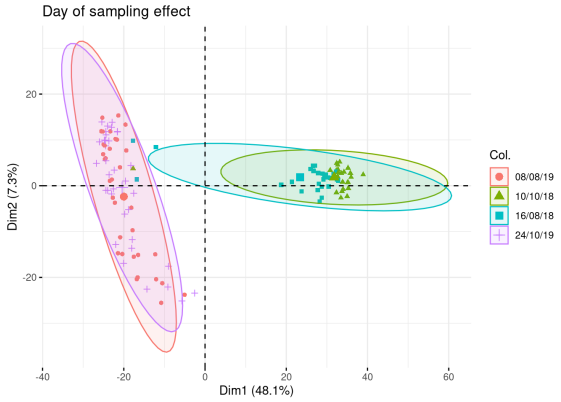



FIGURE 4.13 – Effet jour de prélèvement.

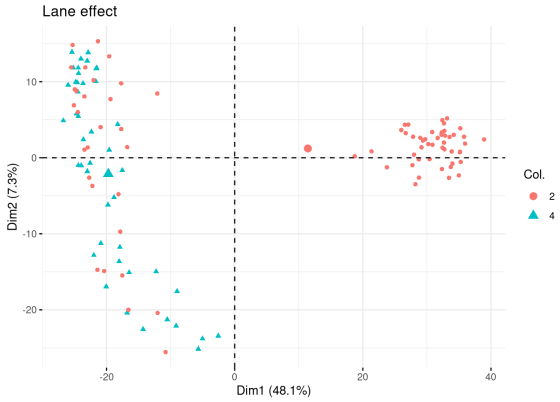



FIGURE 4.14 – Effet plaque de séquençage.

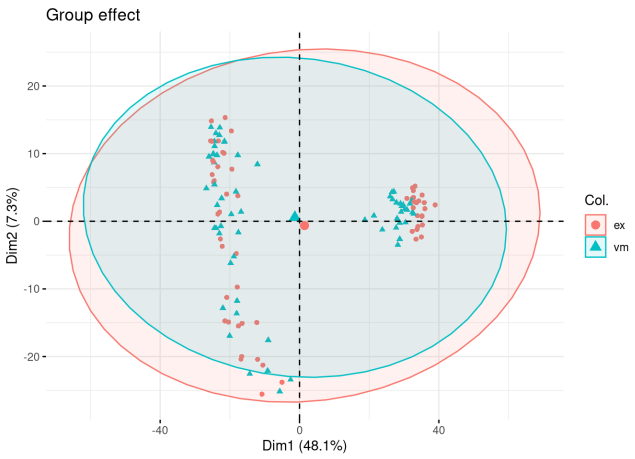



FIGURE 4.15 – Effet groupe.

En plus de ces 4 animaux atypiques, on observe un effet année marqué (*voir figure 4.12* ) ainsi qu'un effet lot (*voir figure 4.11* ). On observe un effet jour de prélèvement en 2018 (*voir figure 4.13* ), en plus de l'effet année qui est également visible sur cette figure. De plus on voit que l'effet jour de prélèvement correspond à l'effet groupe (*voir figure 4.15* ).
La correction de l'effet jour de prélèvement, de l'effet groupe ou de l'effet lot entrainerait une perte de la variabilité biologique que nous souhaitons observer (les groupes et les lots étant liés directement à la valeur génétique rfi qui est en lien avec la lignée). En revanche, l'ACP obtenue après la correction de l'effet année sur l'ensemble des données montre une importante différence des valeurs de variance entre 2018 et 2019 (*voir figure 4.16* ).

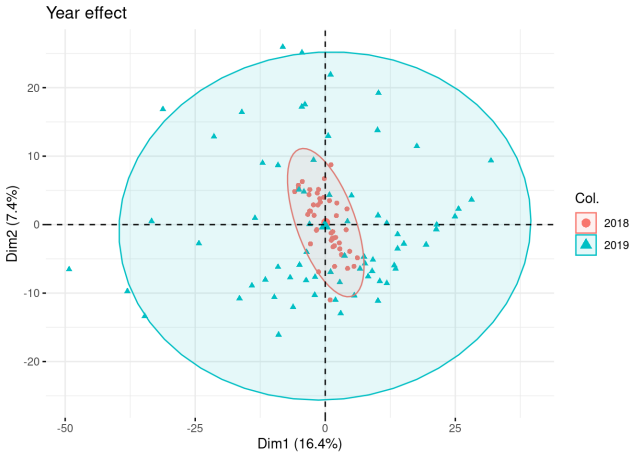



FIGURE 4.16 – Effet année après correction.

Des biais dans les analyses postérieures pourraient apparaitre si l'effet année était corrigé. C'est pourquoi pour ce régime alimentaire je n'ai corrigé aucun effet. Les données de 2018 et 2019 devront être analysées séparément.

## 4.2 Analyses des performances en élevage et du métabolome

### 4.2.1 Analyse détaillée du métabolome en DAF

#### Métabolome du rumen

Dans un premier temps je vais présenter les résultats obtenus avec la PLS-DA.

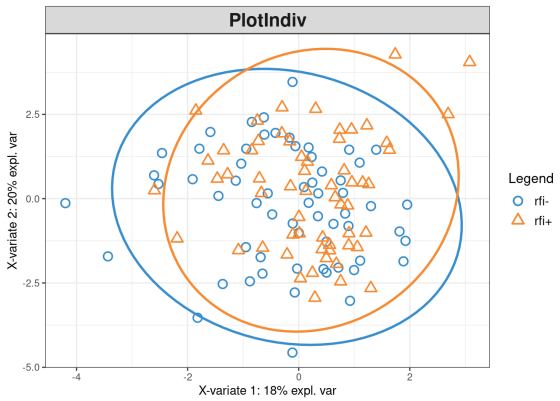



FIGURE 4.17 – Affichage des individus.

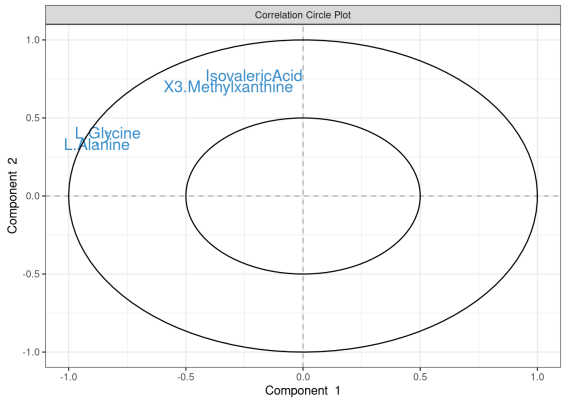



FIGURE 4.18 – Affichage des variables les plus corrélées avec l'axe 1 ou l'axe 2.

On n'observe aucune séparation claire des deux lignées. Les deux nuages de points sont presque semblables (*voir figure 4.17*). Seul un petit nombre d'individus différentie le premier nuage de points du deuxième.
Les deux tests (Student et Wilcoxon) confirment le résultat : aucun des métabolites présents dans le rumen ne présente de différences significatives entre les deux lignées.

#### Métabolome du plasma

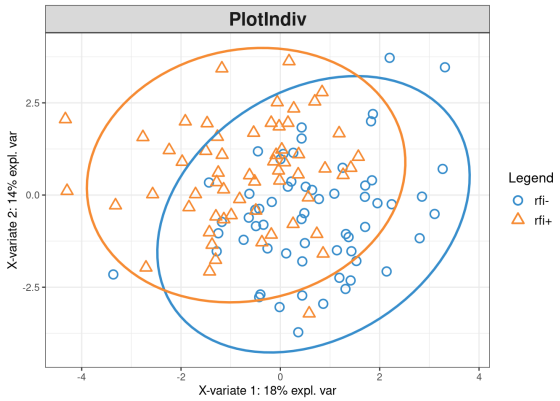



FIGURE 4.19 – Affichage des individus.

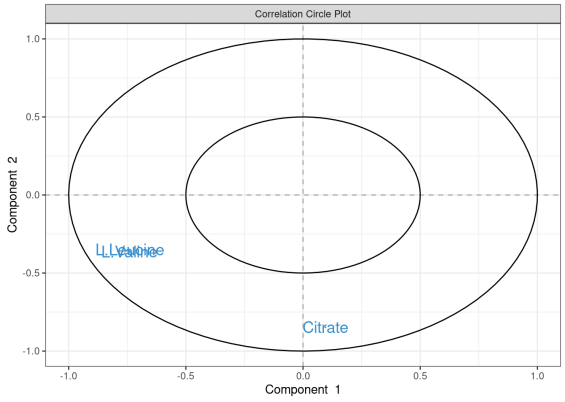



FIGURE 4.20 – Affichage des variables les plus corrélées à l'axe 1 ou l'axe 2.

On observe un effet lignée sur les deux axes (*figure 4.19* ) : le nuage de points correspondant à la lignée rfi+ est plus à gauche sur l'axe 1 que le nuage de points de la lignée rfi-. Sur l'axe 2 c'est le nuage de points de la lignée rfi- qui est plus bas que le deuxième nuage de points.

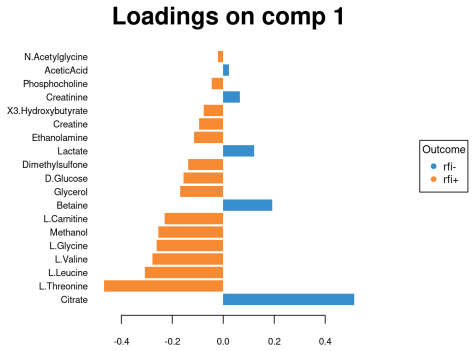



FIGURE 4.21 – Corrélation variables/axe1.

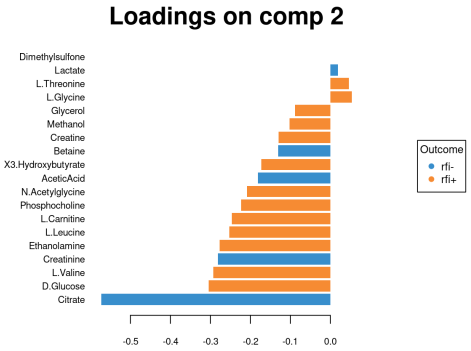



FIGURE 4.22 – Corrélation variables/axe2.

Les concentrations en citrate et en thréonine sont les variables les plus corrélées à l'axe 1 (*voir figure 4.21* ), seule la concentration en citrate est fortement corrélée à l'axe 2 (*voir figure 4.22* ). La concentration en citrate est plus élevée chez les animaux de lignée rfi-. A contrario les animaux de lignée rfi+ ont une concentration en thréonine supérieure à celle des animaux de lignée rfi-.
Les métabolites du plasma présentant des différences significatives entre les deux lignées sont : le citrate et la thréonine d'après les tests de Student et Wilcoxon. À nouveau, les résultats des tests concordent avec les résultats observés avec la PLS-DA.

Ci dessous (*voir figure 4.23 et 4.24* ) les boxplots des deux métabolites les plus discriminants pour la lignée. Ces graphiques confirment les résultats de la PLS-DA : le citrate est plus présent dans le plasma des animaux de lignée rfi- et inversement pour le thréonine qui est en quantité plus important dans le plasma des animaux de lignée rfi+.

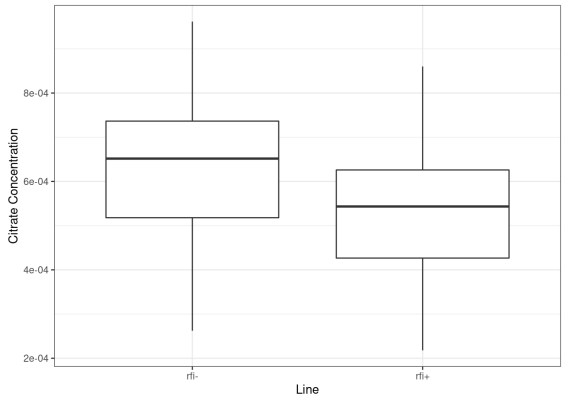



FIGURE 4.23 – Boxplot de la concentration en Citrate.

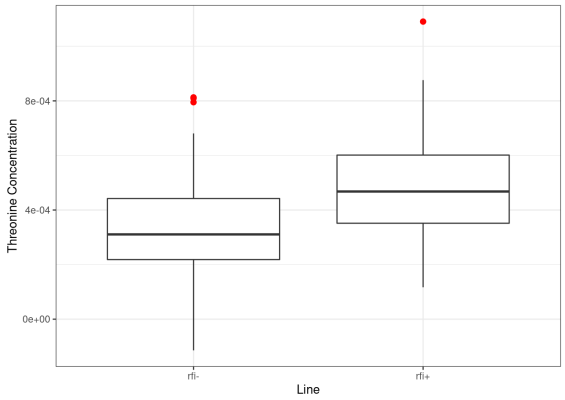



FIGURE 4.24 – Boxplot de la concentration en Thréonine.

### 4.2.2 Synthèses des autres analyses

#### Analyses des performances

En DAC : la consommation moyenne et résiduelle, le poids à âge type, l'épaisseur du muscle en milieu et fin de contrôle DAC, le poids en début, en milieu et fin de contrôle sont les variables qui présentent une différence significative entre les deux lignées. Les variables qui intéressent le plus les biologistes sont les variables liées à la consommation des animaux. Ci-dessous (*voir figure 4.25*

*et 4.26* ) sont représentés les boxplots de la consommation moyenne et résiduelle en fonction de la lignée des animaux.

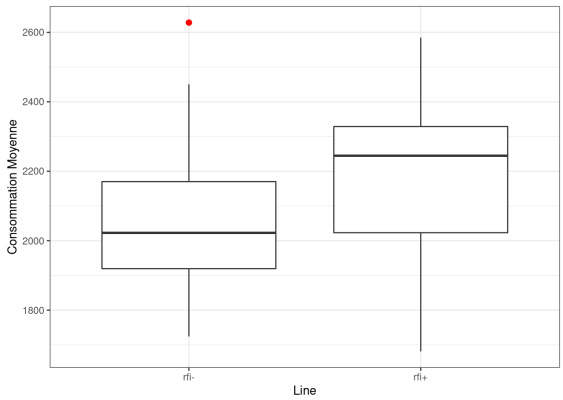



FIGURE 4.25 – Boxplot de la consommation moyenne des animaux.

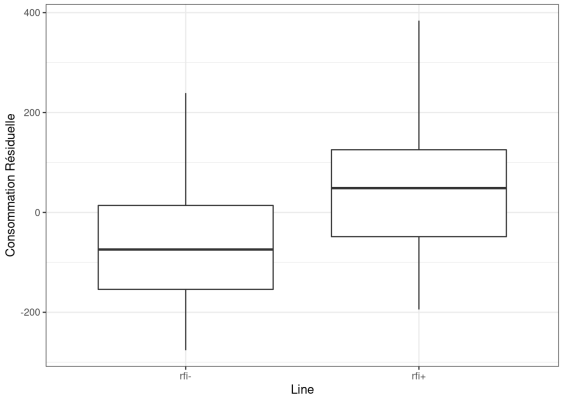



FIGURE 4.26 – Boxplot de la consommation résiduelle des animaux.

En DAF : nous n'observons pas de fortes différences entre les deux lignées sur la base de leurs performances.

### Autres analyses du métabolome en DAC

Les analyses réalisées sur les animaux en DAC donnent les résultats suivants : aucun des métabolites présents dans le rumen ne présente de différence significative entre les deux lignées. En revanche, le citrate est le seul métabolite dans le plasma qui présente une différence significative entre les deux lignées (*voir figure 4.27* ). Comme en DAF, la concentration en citrate dans le plasma est plus élevée chez les animaux de lignée rfi-.

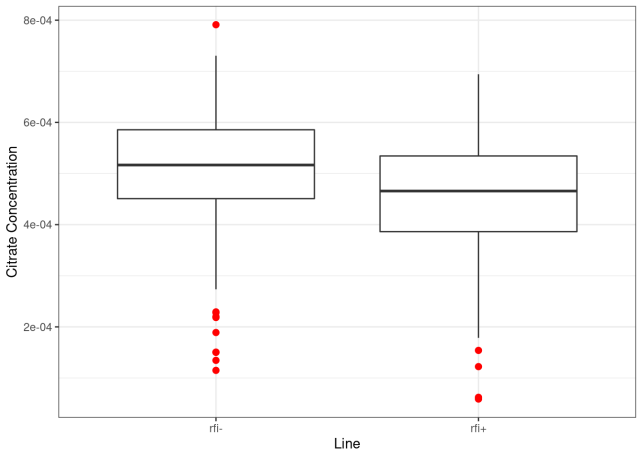



FIGURE 4.27 – Boxplot de la concentration en citrate dans le plasma en phase DAC.

## 4.3 Analyses du microbiote

### 4.3.1 Analyses en DAC

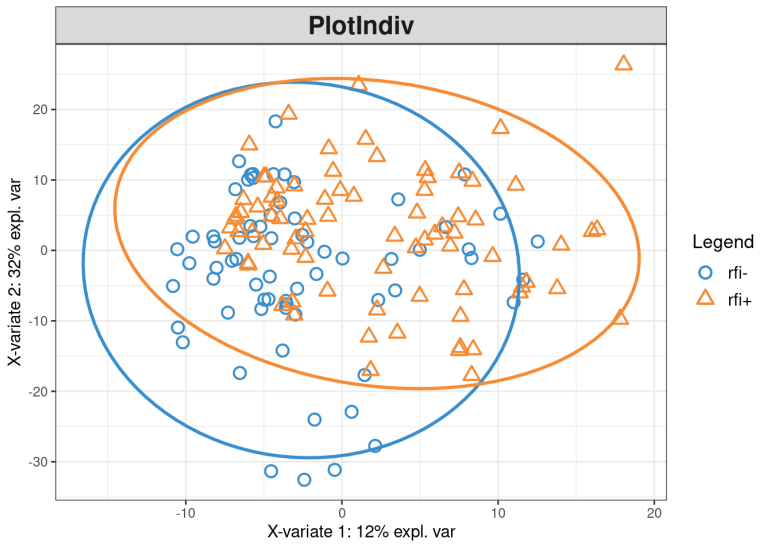



FIGURE 4.28 – Affichage des individus.

On observe un effet lignée sur les deux axes (*figure 4.28* ) : le nuage de points correspondant à la lignée rfi+ est plus à droite sur l'axe 1 que le nuage de points de la lignée rfi-. Sur l'axe 2 c'est le nuage de points de la lignée rfi- qui est plus bas que le deuxième nuage de points.

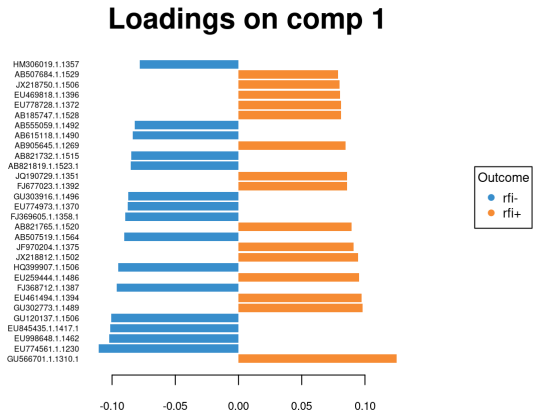



FIGURE 4.29 – Corrélation variables/axe1.

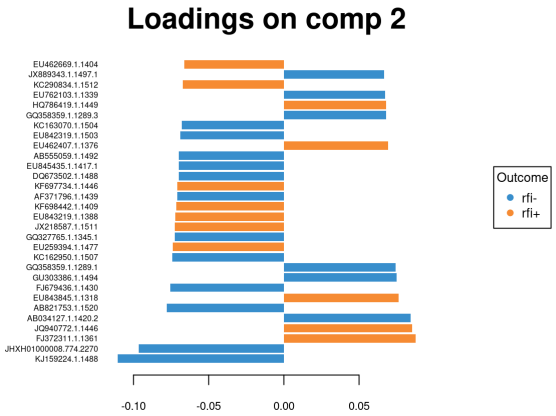



FIGURE 4.30 – Corrélation variables/axe2.

GU566701.1.1310.1 est la variable la plus corrélée à l'axe 1, tandis que KJ159224.1.1488 est la plus corrélée à l'axe 2. L'OTU GU566701.1.1310.1 est plus abondant chez les animaux rfi+ que chez les animaux de lignée rfi-. A l'inverse l'OTU KJ159224.1.1488 est plus abondant chez les animaux de lignée rfi- (*voir figure 4.29 et 4.30* ).

L'OTU présentant des différences significatives entre les deux lignées est : GU566701.1.1310.1 d'après le test de Wilcoxon. Le test de Student, lui, ne ressort aucun OTU discriminant. Les résultats des tests concordent avec les résultats observés avec la PLS-DA. Ci-dessous le boxplot de l'OTU présentant des différences significatives :

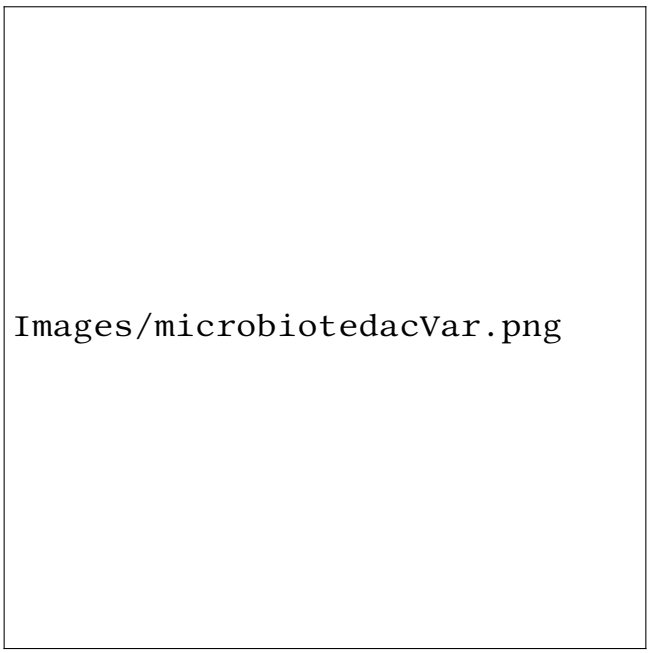



FIGURE 4.31 – Boxplot de la quantité de l'OTU discriminante en phase DAC.

### 4.3.2 Synthèse des analyses en DAF

#### 2018

La pls-da montre un fort effet lignée sur le premier axe en 2018. Les variables les plus corrélée à l'axe 1 sont les suivantes : FJ028792.1.1490, KF697956.1.1425, EU842667.1.1498, AF371914.1.1431.1, AY445602.1.1501 (plus abondant chez les animaux rfi-) et HQ399968.1.1499 (plus abondant chez les animaux rfi+). En revanche les deux tests statistiques ne donnent pas de résultats : aucun OTU ne présente de différence significative entre les deux lignées. Ci-dessous (*voir figure 4.32* ) les boxplots des OTUs les plus corrélés à l'axe 1 :

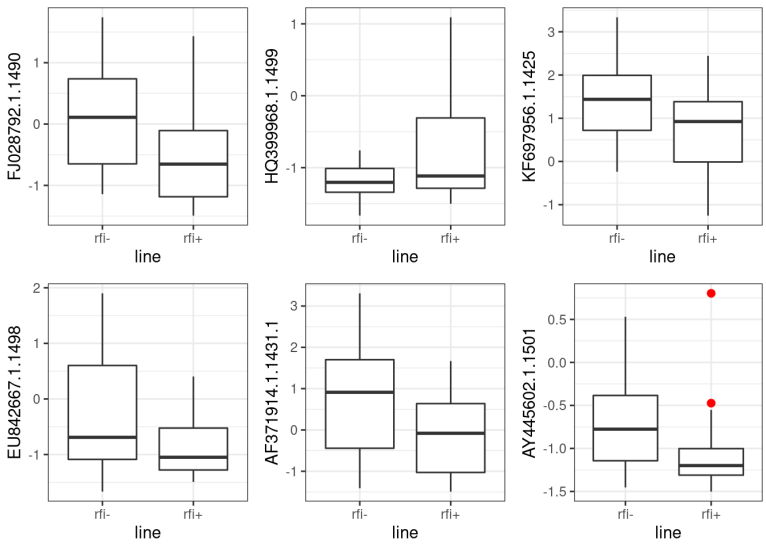



FIGURE 4.32 – Boxplot de la quantité de l'OTU en phase DAC en 2018.

**2019**

La pls-da montre également un fort effet lignée en 2019 : les animaux rfi+ étant plus à droite sur l'axe 1 et plus en haut sur l'axe 2. Les variables les plus corrélées à l'axe 1 sont les suivantes : HQ399907.1.1506, EU774561.1.1230 (plus présents chez les animaux rfi-) et AY445602.1.1501 (plus présents chez les animaux rfi+). Les variables les plus corrélées à l'axe 2 sont les suivantes : EU843219.1.1388 (plus présents chez les animaux rfi-) et AB034127.1.1420.2, GQ358359.1.1289.1 (plus présents chez les animaux rfi+). À nouveau les deux tests statistiques ne donnent pas de résultats : aucun OTU ne présente de différence significative entre les deux lignées. Ci-dessous les boxplots des OTUs les plus corrélés à l'axe 1 (*voir figure 4.33 première ligne*) et les plus corrélés à l'axe 2 (*voir figure 4.33 deuxième ligne*) :

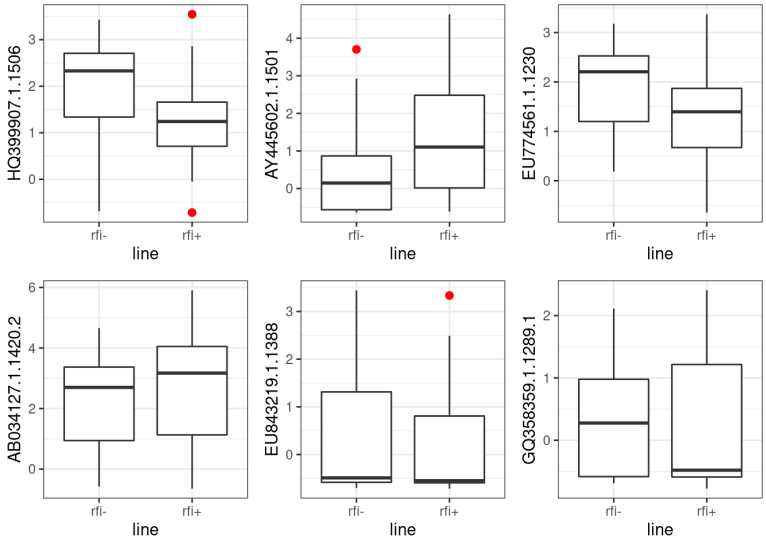



FIGURE 4.33 – Boxplot de la quantité de l'OTU en phase DAC en 2019.

# 5 Conclusions

J'ai pu analyser quatre des six jeux de données omiques que j'avais à disposition. L'objectif était de sélectionner les variables discriminantes pour la lignée. De manière générale, j'ai obtenu plus de variables explicatives pour le premier régime alimentaire (DAC) que pour le deuxième (DAF) :

- 8 variables pour les perfomances en DAC (consommation moyenne et résiduelle, poids à âge type, épaisseur du muscle en milieu et fin de régime, poids en début et en milieu et fin régime) contre zéro en DAF,
- aucune variable discriminante pour le métabolome du rumen que ce soit en DAC ou en DAF,
- une variable pour le métabolome du plasma en DAC (le citrate) contre deux en DAF (le citrate et la thréonine),
- et un OTU discriminant pour le microbiote du rumen en DAC (GU566701.1.1310.1) contre zéro en DAF.

Ceci est dû à une alimentation beaucoup plus contrôlée en DAC : en effet il est plus facile de gérer la quantité de nourriture consommée par les animaux lors qu'elle est sous forme de concentrés (régime DAC) que lorsque c'est du fourrage (régime DAF). Par conséquent, nous avons une différence d'alimentation entre les animaux en DAF plus importante et de ce fait une alimentation différente au sein de chaque lignée. La discrimination des deux lignées est alors moins marquée.

Le génotype ainsi que les mesures infra-rouge des fèces n'ont pas été traités. En effet un certain nombre de difficultés ont été rencontrées lors de l'analyse du microbiote : effets techniques liés à la lignée qui ne pouvaient par conséquent pas être corrigés, groupe d'individus ne pouvant être expliqué par aucune variable, ou encore une différence de variance trop importante entre les animaux de 2018 et 2019. Les travaux sur les données du microbiote ont pris plus de temps que prévu.

Suite aux résultats présentés aux biologistes, il a été décidé qu'il serait intéressant de mettre en lien le microbiote du rumen et le métabolome du rumen afin de voir si certains métabolites sont liés avec certains OTUs. Malheureusement je n'ai pas pu terminer correctement cette étude par manque de temps.

D'un point de vue personnel, ce stage a été une réelle opportunité de développer mes connaissances en statistiques ainsi que mes compétences en R. J'ai trouvé l'échange avec les biologistes très intéressant, cela m'a permis de lier mathématiques et biologie afin de répondre à des problématiques concrètes. J'ai ainsi pu mesurer l'apport que peuvent avoir les statistiques dans d'autres disciplines. Ces quatre mois m'ont également permis de me familiariser avec le monde du travail. Ce stage a été une expérience enrichissante que j'ai beaucoup appréciée.

# 6 Annexes

# Dac data analysis 3

**Annaleah JOHANNY**

**20 août, 2021**

Necessary R packages are loaded with:

```
library("ggplot2")
library("GGally")
library("FactoMineR")
library("factoextra")
library("corrplot")
library("sva")
library("mixOmics")
```

A palette for batches is also defined:

```
palette_batches <- c("#999933", "#00FF00", "#FF0000", "#0000FF", "#FF9900",
                     "#333333", "#CC00FF", "#990033", "#FF99FF", "#FFFF00",
                     "#00CCFF")
```

## Data loading

Data are loaded with:

```
design <- read.table("../data/clean/dac_design.tsv", header = TRUE)
dim(design)
```

```
## [1] 196  16
```

```
head(design)
```

| | id <chr> | sample <chr> | condition <chr> | year <int> | batch <int> | line <chr> |
|---|---|---|---|---|---|---|
| 20000188510 | AMbRFIJR78_ACCAGG-CKWP3_L001_R | AMbRFIJR78 | DAC2018 | 2018 | 5 | rfi+ |
| 20000188513 | AMbRFIJR157_GACTCG-CKWP3_L001_R | AMbRFIJR157 | DAF2018 | 2018 | 5 | rfi- |
| 20000188519 | AMbRFIJR88_TTGTGA-CKWP3_L001_R | AMbRFIJR88 | DAC2018 | 2018 | 5 | rfi- |
| 20000188521 | AMbRFIJR121_CAGAAA-CKWP3_L001_R | AMbRFIJR121 | DAF2018 | 2018 | 4 | rfi+ |
| 20000188523 | AMbRFIJR23_CACCTC-CKWP3_L001_R | AMbRFIJR23 | DAC2018 | 2018 | 6 | rfi- |
| 20000188526 | AMbRFIJR94_AAAGCG-CKWP3_L001_R | AMbRFIJR94 | DAC2018 | 2018 | 5 | rfi+ |

6 rows | 1-7 of 17 columns

```
PlasmaRMN_dac <- read.table("../data/clean/dac_plasmarmn.tsv", header = TRUE)
dim(PlasmaRMN_dac)
```

```
## [1] 196  26
```

FIGURE 6.1 – Rapport sur l'analyse du métabolome du plasma en DAC.

```
head(PlasmaRMN_dac)
```

| | D.Glucose <dbl> | Betaine <dbl> | AceticAcid <dbl> | Lactate <dbl> | L.Glycine <dbl> | L.GlutamicAcid <dbl> |
|---|---|---|---|---|---|---|
| 20000188510 | 0.02247645 | 0.002882615 | 0.0004217906 | 0.001978554 | 0.001618902 | 0.001603716 |
| 20000188513 | 0.02387919 | 0.002602636 | 0.0005055885 | 0.001469148 | 0.001991228 | 0.001537500 |
| 20000188519 | 0.03009591 | 0.002770140 | 0.0005965981 | 0.003638930 | 0.002737173 | 0.001443538 |
| 20000188521 | 0.02337896 | 0.002395278 | 0.0008544613 | 0.001147137 | 0.002146223 | 0.001462438 |
| 20000188523 | 0.02970752 | 0.003538596 | 0.0006938635 | 0.005608077 | 0.002008784 | 0.001866717 |
| 20000188526 | 0.02781356 | 0.001933607 | 0.0006950263 | 0.003214698 | 0.002336103 | 0.001470476 |

6 rows | 1-7 of 27 columns

The dataset contains information on 196 animals (sheeps). In `design` we have 16 variables:

- `id` animal identifier (long form),
- `sample` animal identifier (short form),
- `condition` unrelevant, except for the year,
- `year` year of sampling,
- `batch` batch where the animal has been raised,
- `line` genetic selection line of the animal: rfi+ or rfi-,
- `generation` generation of the animal (2 or 3),
- `spir_num` number of sampling,
- `rfi_gvalue` genetic rfi value (used for the selection),
- `rumen_sample_day` day of rumen sampling,
- `rumen_sample_hour` hour of rumen sampling,
- `order_number` order number of rumen sampling,
- `seq_depth` sequencing depth,
- `plaque` sequencing lane,
- `run` sequencing run,
- `age_sampling` age of the animal at the sampling day.

In `PlasmaRMN_dac` there are 26 variables. Each variables correspond to specific metabolite. The table gives the concentration of each metabolite in the plasma of each animal during DAC control.

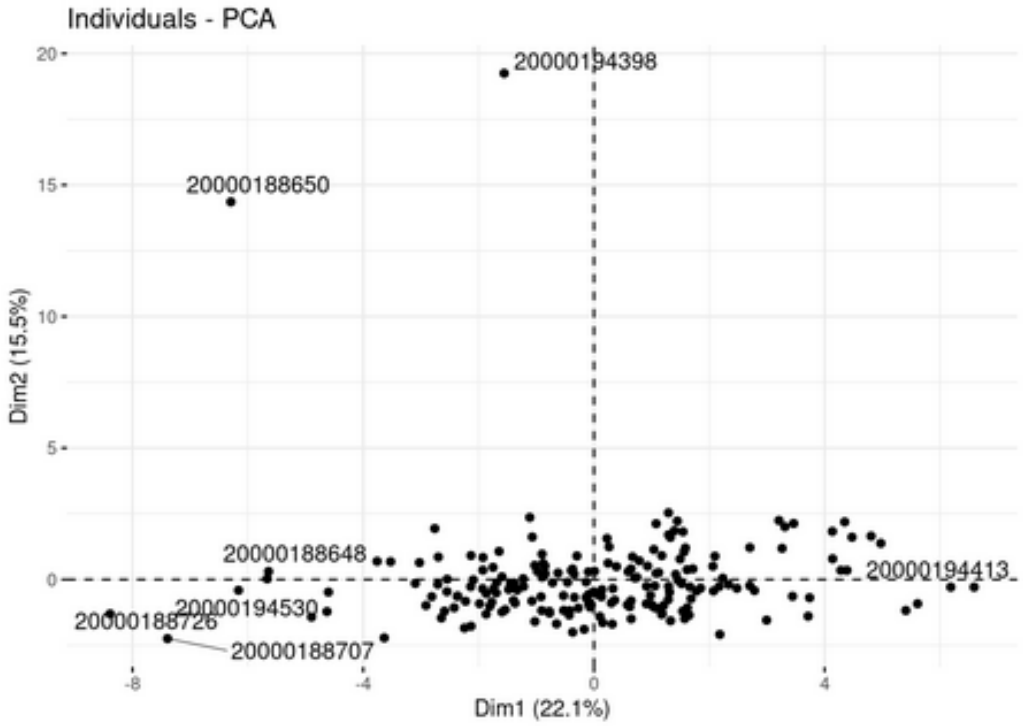
# Multivariate analysis

## PCA

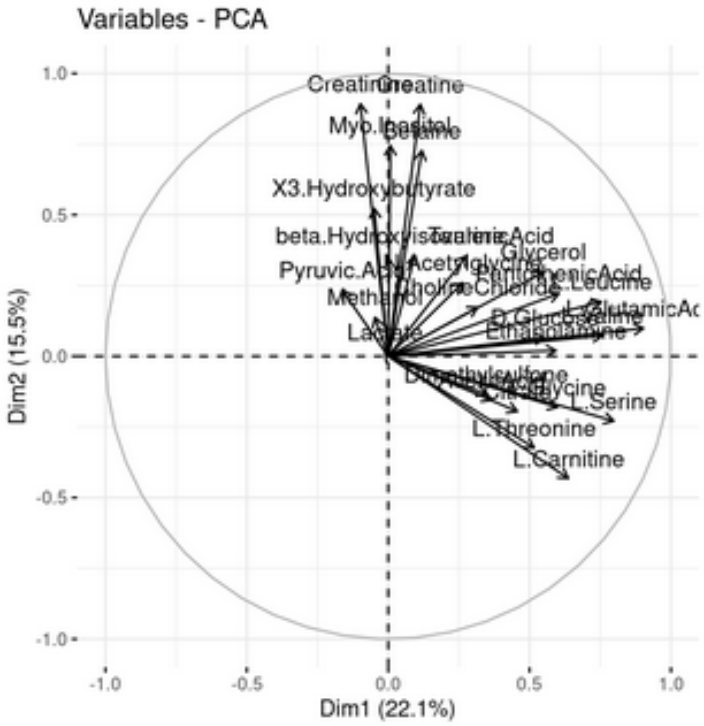
```
tab_plasmadac <- merge(PlasmaRMN_dac, design, by = "row.names")
rownames(tab_plasmadac) <- tab_plasmadac$Row.names
tab_plasmadac <- tab_plasmadac[, -1]
tab_plasmadac$batch <- as.factor(tab_plasmadac$batch)
tab_plasmadac$year <- as.factor(tab_plasmadac$year)

pca_plasmadac = PCA(tab_plasmadac[, c(1:26)], scale.unit = TRUE, graph = FALSE)

fviz_pca_ind (pca_plasmadac, repel = TRUE)
```

```
## Warning: ggrepel: 189 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

FIGURE 6.2 – Rapport sur l'analyse du métabolome du plasma en DAC.



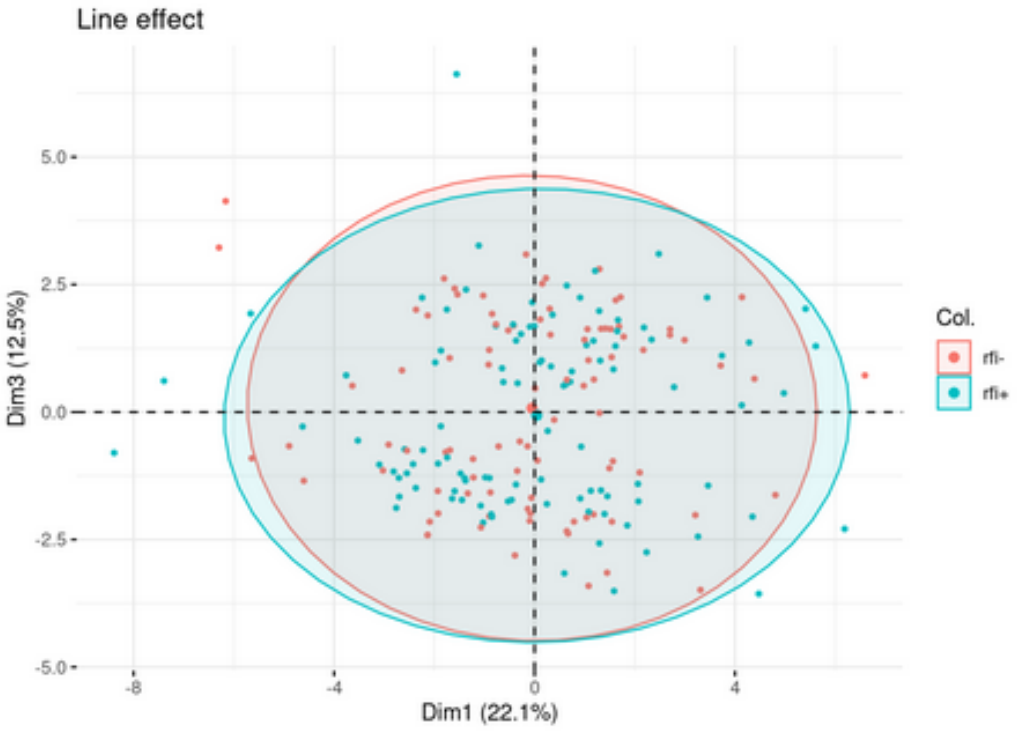

```
fviz_pca_var(pca_plasmadac, col.var = "black")
```




Colors by line, batch and year are then added:

```
fviz_pca_ind(pca_plasmadac,
             axes = c(1, 3),
             geom.ind = "point",
             pointshape = 20,
             col.ind = tab_plasmadac$line,
             addEllipses = TRUE) +
ggtitle("Line effect")
```

FIGURE 6.3 – Rapport sur l'analyse du métabolome du plasma en DAC.




```
fviz_pca_ind(pca_plasmadac,
             axes = c(1, 3),
             geom.ind = "point",
             pointshape = 20,
             col.ind = tab_plasmadac$batch,
             palette = palette_batches,
             addEllipses = FALSE) +
ggtitle("Batch effect")
```

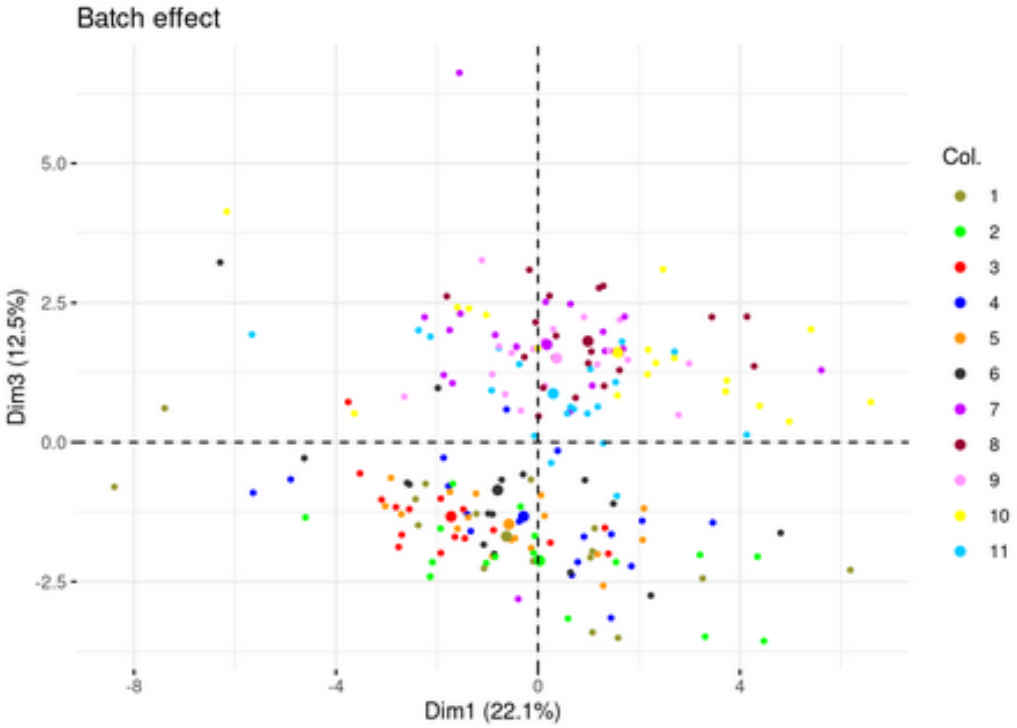
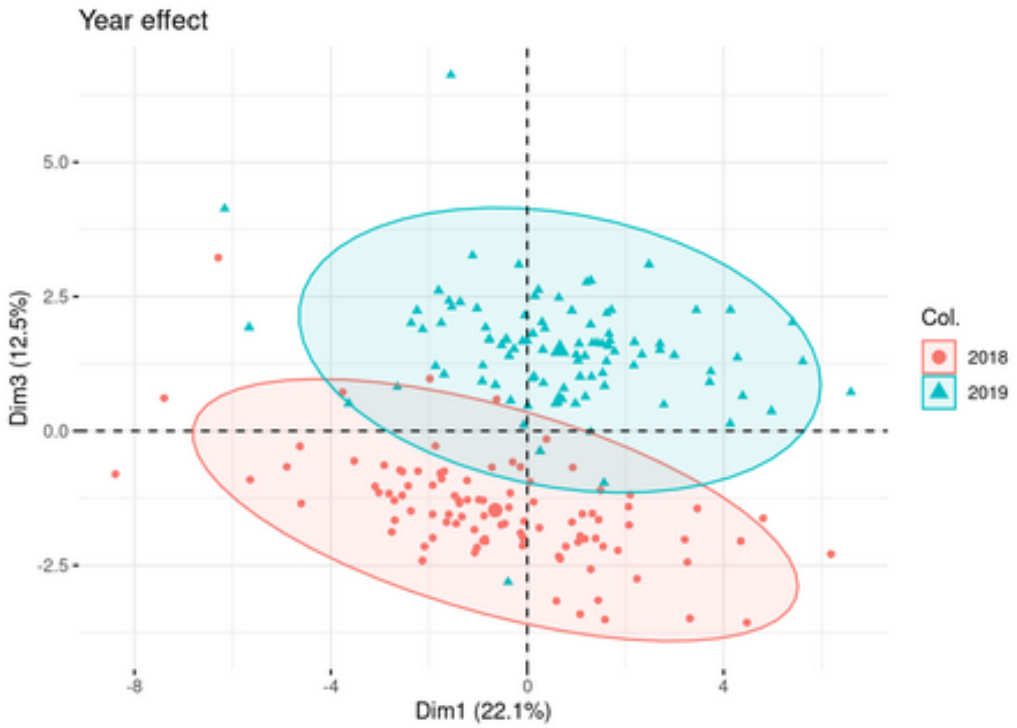



FIGURE 6.4 – Rapport sur l'analyse du métabolome du plasma en DAC.

```
fviz_pca_ind(pca_plasmadac,
             axes = c(1, 3),
             geom.ind = "point",
             col.ind = tab_plasmadac$year,
             addEllipses = TRUE) +
ggtitle("Year effect")
```




There is not a visible line effect. However we can observe a batch effect on both axis and a strong year effect on the second axis.

## Correction of batch effect

We use `ComBat` to correct batch effect. As the batch is directly related to the performances we add `model.matrix(~line, data = design)` as an argument of `ComBat`, in order to maintain the line effect:

```
plasmadac_cor <- t(ComBat(dat = t(PlasmaRMN_dac), batch = design$batch, mod = model.matrix(~lin
e, data = design)))
```

```
## Found11batches
```

```
## Adjusting forlcovariate(s) or covariate level(s)
```

```
## Standardizing Data across genes
```

```
## Fitting L/S model and finding priors
```

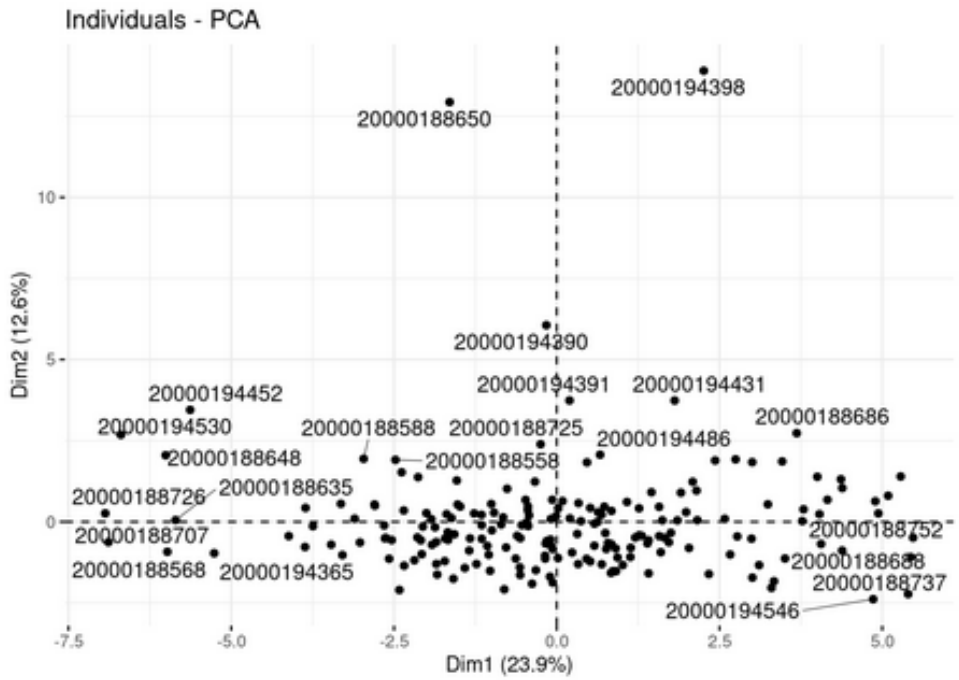
```
## Finding parametric adjustments
```

```
## Adjusting the Data
```

FIGURE 6.5 – Rapport sur l'analyse du métabolome du plasma en DAC.

```
tab_plasmadac2 <- merge(plasmadac_cor, design, by = "row.names")
rownames(tab_plasmadac2) <- tab_plasmadac2$Row.names
tab_plasmadac2 <- tab_plasmadac2[, -1]
tab_plasmadac2$batch <- as.factor(tab_plasmadac2$batch)
tab_plasmadac2$year <- as.factor(tab_plasmadac2$year)

pca_plasmadac_cor = PCA(tab_plasmadac2[, c(1:26)], scale.unit = TRUE, graph = FALSE)

fviz_pca_ind (pca_plasmadac_cor, repel = TRUE)
```

```
## Warning: ggrepel: 174 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
fviz_pca_var(pca_plasmadac_cor, col.var = "black")
```

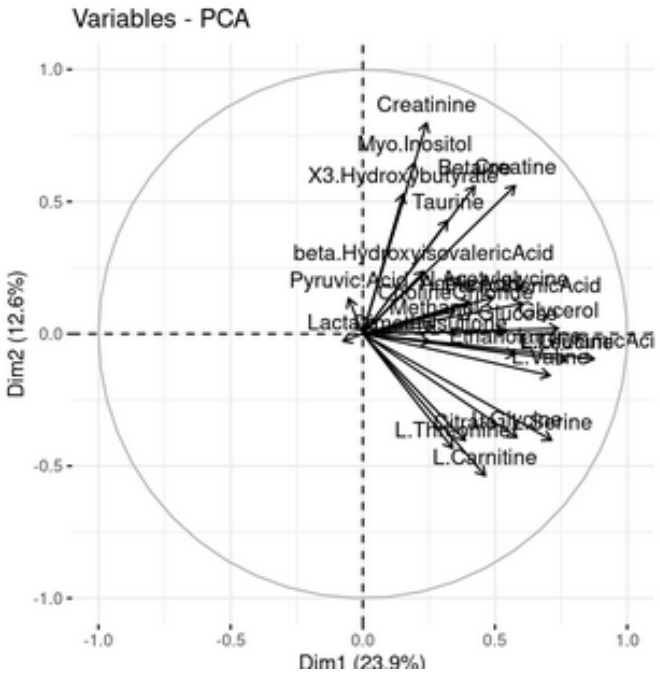
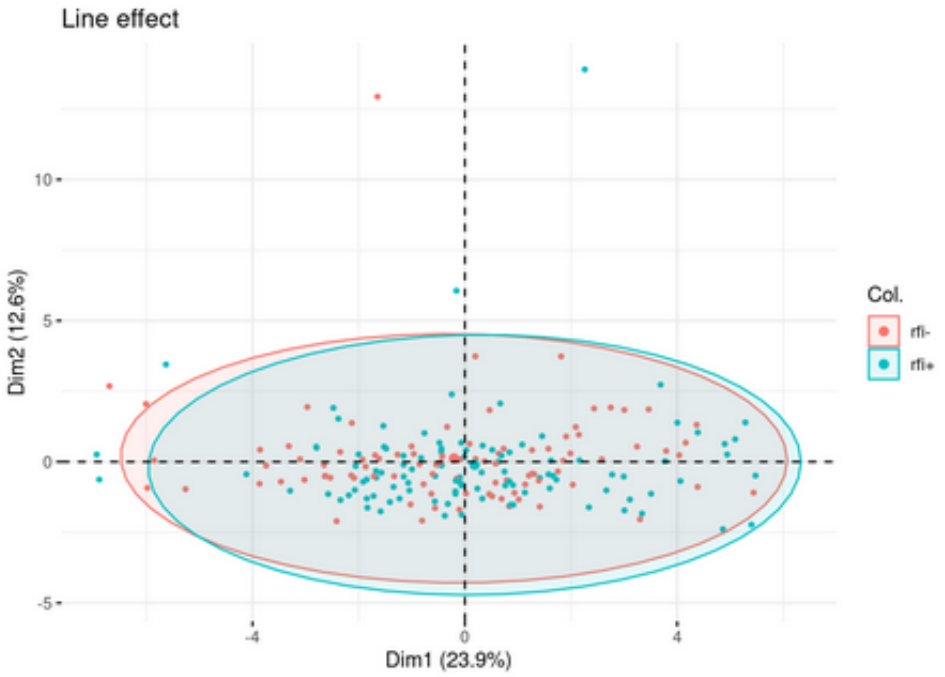



FIGURE 6.6 – Rapport sur l'analyse du métabolome du plasma en DAC.

```
fviz_pca_ind(pca_plasmadac_cor,
             axes = c(1, 2),
             geom.ind = "point",
             pointshape = 20,
             col.ind = tab_plasmadac2$line,
             addEllipses = TRUE) +
ggtitle("Line effect")
```




```
fviz_pca_ind(pca_plasmadac_cor,
             axes = c(1, 2),
             geom.ind = "point",
             pointshape = 20,
             col.ind = tab_plasmadac2$batch,
             palette = palette_batches,
             addEllipses = FALSE) +
ggtitle("Batch effect")
```

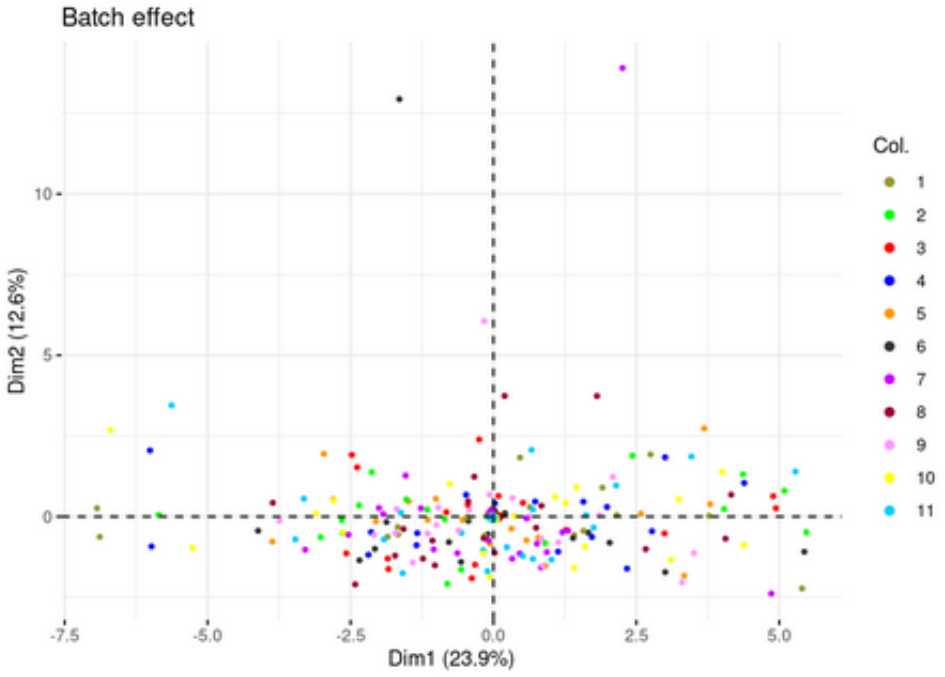
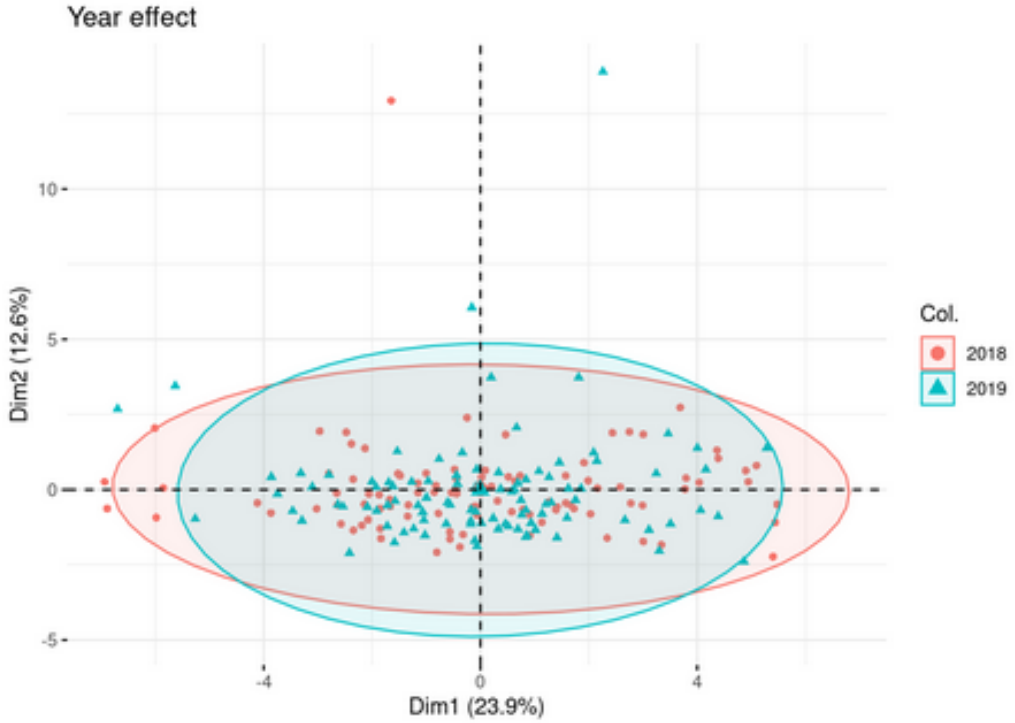



FIGURE 6.7 – Rapport sur l'analyse du métabolome du plasma en DAC.

```
fviz_pca_ind(pca_plasmadac_cor,
             axes = c(1, 2),
             geom.ind = "point",
             col.ind = tab_plasmadac2$year,
             addEllipses = TRUE) +
ggtitle("Year effect")
```




The line effect is still not visible. However there is not a batch effect and a year effect anymore.

# Tests

## Student's t-test

```
student <- function(variable) {
  df <- tab_plasmadac2[, c("line", variable)]
  res <- t.test(df[, 2] ~ df[, 1], var.equal = TRUE)
  return(res$p.value)
}
pval_student <- sapply(names(tab_plasmadac2)[1:26], student)
adj_student <- p.adjust(pval_student, method = "BH")
adj_student
```

```
##                 D.Glucose                   Betaine
##                 0.9929643                 0.7885902
##                AceticAcid                   Lactate
##                 0.9929643                 0.9929643
##                 L.Glycine             L.GlutamicAcid
##                 0.5132657                 0.5132657
##                  Creatine               L.Threonine
##                 0.7885902                 0.1717501
##                  L.Valine                  Glycerol
##                 0.9981928                 0.4041979
##                  L.Serine                  L.Leucine
##                 0.2821154                 0.7885902
```

Figure 6.8 – Rapport sur l'analyse du métabolome du plasma en DAC.

```
##            N.Acetylglycine               Ethanolamine
##                  0.9929643                  0.9929643
##                    Taurine            Dimethylsulfone
##                  0.9201451                  0.9929643
##                    Citrate         X3.Hydroxybutyrate
##                  0.1717501                  0.9272401
##                 L.Carnitine               Myo.Inositol
##                  0.5132657                  0.9929643
## beta.HydroxyisovalericAcid                 Creatinine
##                  0.1717501                  0.9929643
##                    Methanol               Pyruvic.Acid
##                  0.9929643                  0.8870569
##            PantothenicAcid            CholineChloride
##                  0.9929643                  0.9929643
```

```
names(which(adj_student <= 0.05))
```

```
## character(0)
```

## Wilcoxon test

```
wilcoxon <- function(variable) {
  df <- tab_plasmadac2[, c("line", variable)]
  res <- wilcox.test(df[, 2] ~ df[, 1])
  return(res$p.value)
}
pval_wilcoxon <- sapply(names(tab_plasmadac2)[1:26], wilcoxon)
adj_wilcoxon <- p.adjust(pval_wilcoxon, method = "BH")
adj_wilcoxon
```

```
##                  D.Glucose                    Betaine
##                 0.96880335                 0.59697123
##                 AceticAcid                    Lactate
##                 0.96880335                 0.96880335
##                  L.Glycine              L.GlutamicAcid
##                 0.44334544                 0.59697123
##                   Creatine                 L.Threonine
##                 0.72102102                 0.25112990
##                   L.Valine                    Glycerol
##                 0.96880335                 0.44334544
##                   L.Serine                   L.Leucine
##                 0.25286631                 0.96880335
##            N.Acetylglycine               Ethanolamine
##                 0.96880335                 0.96880335
##                    Taurine            Dimethylsulfone
##                 0.79113662                 0.96880335
##                    Citrate         X3.Hydroxybutyrate
##                 0.03789934                 0.77774557
##                 L.Carnitine               Myo.Inositol
##                 0.59697123                 0.92245033
## beta.HydroxyisovalericAcid                 Creatinine
##                 0.25112990                 0.77774557
```

FIGURE 6.9 – Rapport sur l'analyse du métabolome du plasma en DAC.

```
##                   Taurine            Dimethylsulfone
##                0.79113662                 0.96880335
##                   Citrate          X3.Hydroxybutyrate
##                0.03789934                 0.77774557
##               L.Carnitine                Myo.Inositol
##                0.59697123                 0.92245033
## beta.HydroxyisovalericAcid                 Creatinine
##                0.25112990                 0.77774557
##                  Methanol                Pyruvic.Acid
##                0.65926633                 0.92245033
##           PantothenicAcid             CholineChloride
##                0.96880335                 0.96880335
```

```
names(which(adj_wilcoxon <= 0.05))
```

```
## [1] "Citrate"
```

The Wilcoxon test shows that only citrate present significant differences between the two lines.

## Significant variable

```
p <- ggplot(tab_plasmadac2, aes(x = line, y = Citrate)) +
     geom_boxplot(outlier.colour = "red", outlier.size = 2) +
     xlab("Line") + ylab("Citrate Concentration") +
     theme_bw()
p
```

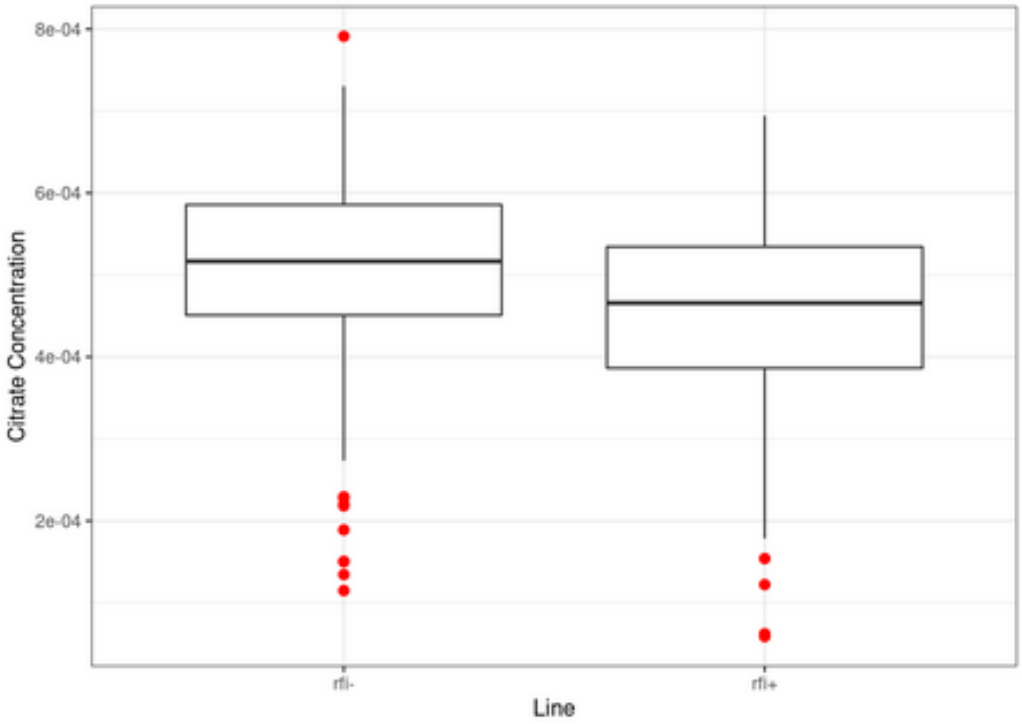
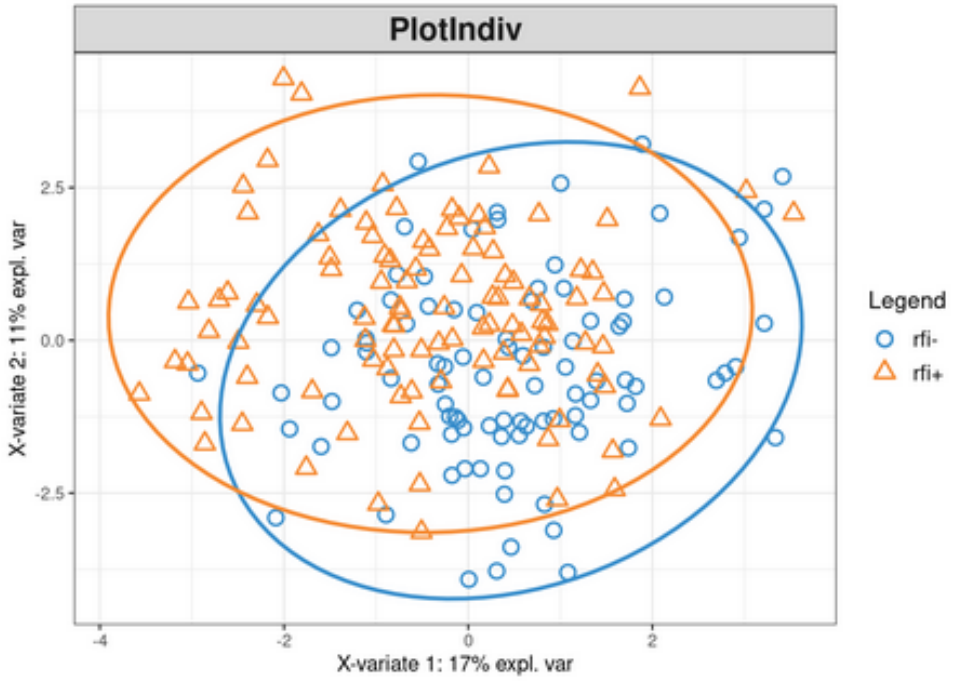



FIGURE 6.10 – Rapport sur l'analyse du métabolome du plasma en DAC.

# PLS-DA

```
y = design$line
x = plasmadac_cor

plsda_plasmadac <- plsda(X = x, Y = y, ncomp = 10)

plotIndiv(plsda_plasmadac , comp = 1:2,
          ind.names = FALSE, ellipse = TRUE,
          legend = TRUE)
```

```
## Warning: It is deprecated to specify `guide = FALSE` to remove a guide. Please
## use `guide = "none"` instead.
```




```
plotVar(plsda_plasmadac, var.names = TRUE, cutoff = 0.6)
```

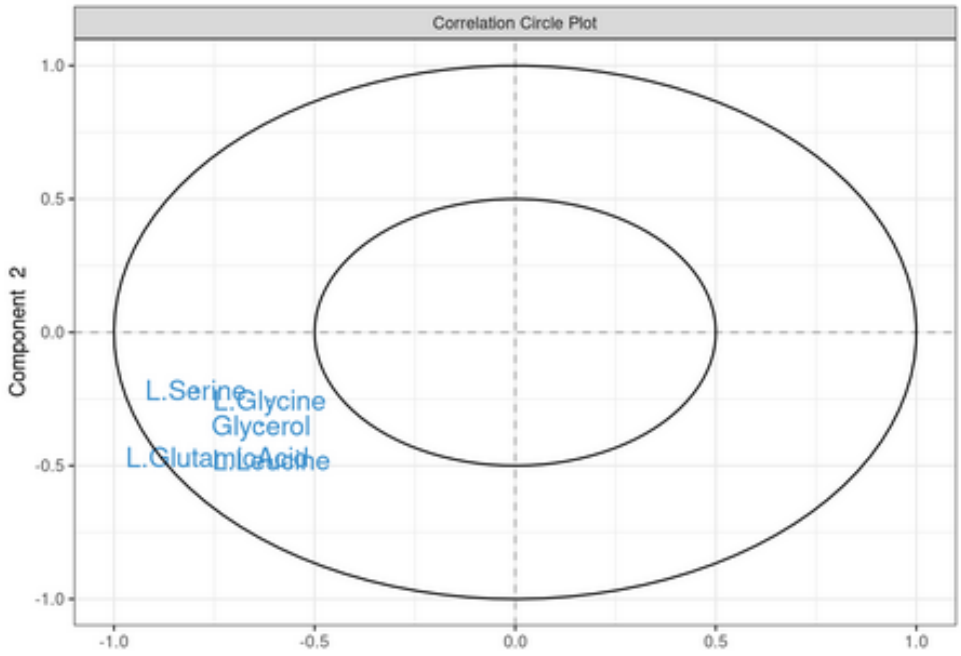



FIGURE 6.11 – Rapport sur l'analyse du métabolome du plasma en DAC.