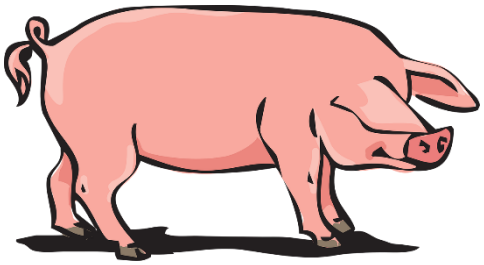


RAPPORT DE STAGE

Analyse de données métabolomiques pour résoudre un problème de bien-être animal

Master 2

Spécialité : Ingénierie de la décision et big data

Parcours : Statistiques / Actuariat

Sous la direction Jean-Marie MARION (enseignant-tuteur)

Et de Nathalie VIALANEIX (encadrante de stage)

Session : Septembre 2019

GUILMINEAU Camille

INRA

24 chemin de Borde-Rouge

31320 Auzeville-Tolosane

Faculté des Sciences

Institut de Mathématiques Appliquées

Année universitaire : 2018-2019

UCO ANGERS
UNIVERSITÉ CATHOLIQUE DE L'OUEST

3 place André Leroy | BP 10808- 49008 Angers cedex 01

UCO
ANGERS
UNIVERSITÉ CATHOLIQUE DE L'OUEST

CHARTE DE NON PLAGIAT

Protection de la propriété intellectuelle

Tout travail universitaire doit être réalisé dans le respect intégral de la propriété intellectuelle d'autrui. Pour tout travail personnel, ou collectif, pour lequel le candidat est autorisé à utiliser des documents (textes, images, musiques, films etc.), celui-ci devra très précisément signaler le crédit (référence complète du texte cité, de l'image ou de la bande-son utilisés, sources internet incluses) à la fois dans le corps du texte et dans la bibliographie. Il est précisé que l'UCO dispose d'un logiciel anti-plagiat dans lms.uco.fr, aussi est-il demandé à tout étudiant de remettre à ses enseignants un double de ses travaux lourds sur support informatique.

Cf. « Prévention des fraudes à l'attention des étudiants »

Je soussigné(e), Camille Guilmineau, étudiant(e) en Master 2 Ingénierie de la décision et big data m'engage à respecter cette charte.

Fait à Auzeville-Tolosane, le 17/04/2019

Signature :

Remerciements

Je tiens à remercier toutes les personnes ayant contribué au bon déroulement de mon stage et m'ayant aidée dans la rédaction de ce rapport.

J'adresse d'abord mes remerciements à mes encadrants de stage Nathalie Vialaneix, Gaëlle Lefort, Laurence Liaubet et Rémi Servien pour leur implication et leur confiance tout au long du stage. Leurs conseils ont été précieux dans la réussite de la mission qui m'a été confiée.

Je remercie Jean-Marie Marion, enseignant-tuteur, pour l'intérêt qu'il a porté à mon stage.

Je remercie également les autres stagiaires du bureau MIAT20, Cyril, Loukas et Typhaine pour leur bonne humeur et pour tous les bons moments que nous avons partagés.

Enfin, je souhaite remercier l'ensemble des personnes de l'unité MIAT pour leur accueil chaleureux pendant ces 6 mois.

Table des matières

Introduction **2**

1 L'INRA et l'unité MIAT **3**
1.1 L'Institut National de la Recherche Agronomique . . . 3
1.2 Le département de Mathématiques et Informatique Appliquées . . . 3
1.3 L'unité Mathématiques et Informatique Appliquées de Toulouse . . . 4

2 Présentation du contexte biologique et du sujet **5**
2.1 La métabolomique et la RMN . . . 5
2.2 Problématique du stage . . . 6

3 Description des données **8**

4 Description des méthodes de quantification **9**
4.1 Découpage en buckets . . . 9
4.2 Le package ASICS . . . 9
4.2.1 Pré-traitements du spectre d'un échantillon . . . 9
4.2.2 Pré-traitements de la librairie de référence . . . 9
4.2.3 Quantification relative des concentrations des métabolites . . . 10
4.3 Contrôle des quantifications . . . 11

5 Analyse exploratoire **13**
5.1 L'Analyse en Composantes Principales (ACP) . . . 13
5.2 Étude des quantifications ASICS . . . 13
5.2.1 Étude sur l'ensemble des individus . . . 13
5.2.2 Étude sur les individus « jugulaire24h » . . . 14
5.2.3 Étude par type d'échantillon . . . 15
5.3 Comparaison avec le découpage en buckets . . . 16
5.4 Étude des quantifications par dosages . . . 16

6 Étude des métabolites **19**
6.1 Test de Kruskal-Wallis et correction des tests multiples . . . 19
6.2 Étude des métabolites significatifs . . . 19
6.3 Modèles mixtes . . . 22
6.4 Analyse des résultats . . . 23

7 Étude des phénotypes **25**
7.1 Lien entre la lignée et la mortalité . . . 25
7.2 Poursuite de l'étude . . . 25

Conclusion **26**

Annexes **29**

Introduction

L'élevage porcin est l'une des industries agroalimentaires les plus intensives. Ces dernières années, la sélection des porcs s'est focalisée sur la prolificité, la croissance et la qualité de la viande. Cela a conduit à une augmentation du nombre de porcelets par portée, qui a été accompagnée d'une hausse de la mortalité à la naissance de l'ordre de 10 à 20%. La mortalité des porcelets est un problème social et éthique lié au bien-être de l'animal. C'est également une source de pertes économiques importantes dans la production porcine.

Le décès des porcelets a souvent lieu pendant les 72 heures suivant la naissance et est en partie dû à une moins bonne maturité. La maturité correspond à un état de développement permettant la survie à la naissance. Il est donc important de pouvoir caractériser cet état de maturité afin de comprendre et de réduire la mortalité périnatale des porcelets.

Cette problématique est étudiée à l'INRA dans le cadre du projet SubPig, notamment dans l'unité MIAT où j'ai effectué mon stage de fin d'études. Dans le cadre de ma mission, je me suis intéressée aux données métabolomiques prélevées sur les porcelets à leur naissance. Une relation a été observée entre la mortalité des porcelets et leur efficacité alimentaire. J'ai donc étudié deux sous-populations divergentes pour un critère d'efficacité alimentaire afin de comprendre les différences entre leurs métabolismes.

Ce rapport a pour but de présenter le travail réalisé au cours du stage. Je présenterai d'abord la structure qui m'accueille, puis le contexte biologique et le sujet de mon stage. Ensuite, je détaillerai les données qui sont à ma disposition. J'expliquerai les méthodes utilisées pour traiter ces données. Je décrirai ensuite l'analyse exploratoire puis l'étude que j'ai réalisée concernant les métabolites et les phénotypes.

1 L'INRA et l'unité MIAT

1.1 L'Institut National de la Recherche Agronomique

L'INRA est un organisme national de recherche scientifique publique, sous la double tutelle des ministères en charge de la Recherche et de l'Agriculture. Il a été fondé en 1946 et est désormais constitué de 13 départements scientifiques, répartis dans 17 centres (figure 1). 7900 agents titulaires composent les 250 unités de recherche et les 45 unités expérimentales.

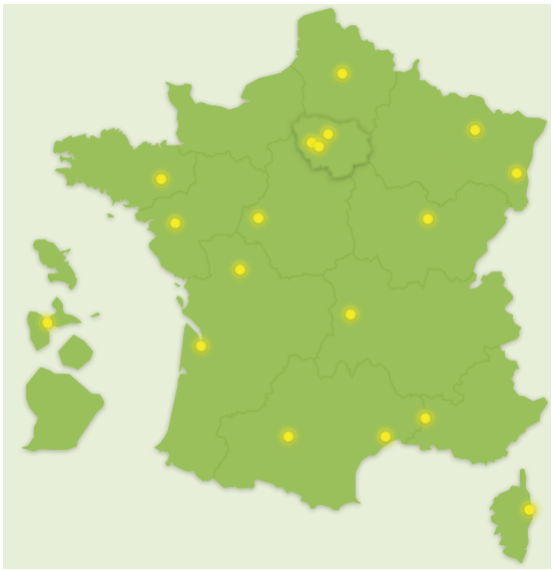


FIGURE 1 – L'implantation des centres INRA.

Les recherches de l'INRA concernent les questions liées à l'agriculture, l'alimentation et la sécurité des aliments, l'environnement et la gestion des territoires, avec une perspective de développement durable. Les missions de l'INRA sont de :

- produire et diffuser des connaissances scientifiques ;
- contribuer à l'innovation par le partenariat et le transfert ;
- former à la recherche ;
- élaborer la stratégie de recherche européenne et nationale ;
- éclairer les décisions publiques ;
- contribuer au dialogue entre sciences et société.

1.2 Le département de Mathématiques et Informatique Appliquées

Le département Mathématiques et Informatique Appliquées (MIA) participe au développement de méthodes, d'outils et de savoir-faire dans le cadre des mathématiques et de l'informatique appliquées aux domaines de l'alimentation, l'agriculture et l'environnement. Le département est organisé autour de 8 unités de recherche. Il a deux missions principales :

- Mission I : mener des recherches en maths-info et à leur interface avec d'autres disciplines, en prise étroite avec de grands enjeux de la recherche en sciences du vivant et de l'environnement.

- Mission II : accompagner le développement des maths-infos dans le contexte de la transition numérique, à l'INRA ainsi qu'auprès de ses partenaires ; en ce sens MIA offre une plus-value à l'Institut pour l'insertion de ses travaux dans la société.

1.3 L'unité Mathématiques et Informatique Appliquées de Toulouse

L'unité Mathématiques et Informatique Appliquées de Toulouse (MIAT) est une unité propre du département MIA. Elle a pour mission scientifique de mettre en œuvre des méthodes mathématiques et informatiques dans le cadre de collaborations avec les autres départements de l'INRA. L'unité est composée actuellement de deux équipes de recherche :

- MAD (Modélisation des Agro-écosystèmes et Décision) : modélisation des systèmes complexes dans les champs de l'agriculture, de l'environnement et 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 permettant de contribuer à la compréhension du vivant.

L'unité compte aussi sur l'activité de trois plateformes :

- Plateforme GENOTOUL : l'une des plateformes bioinformatiques du GIS Genotoul, son activité est centrée sur l'analyse de séquences. La plateforme administre un cluster de calcul qui sera utilisé au cours du stage pour réaliser certains calculs.
- Plateforme RECORD (Rénovation et Coordination de la modélisation des cultures pour la gestion des agro-systèmes) : elle vise à offrir un cadre et des outils informatiques communs aux modélisateurs des différentes disciplines pour la modélisation et la simulation des systèmes de culture.
- Plateforme SIGENAE (Système d'Information des Génomes des Animaux d'Elevage) : elle est issue d'un regroupement d'ingénieurs en bio-informatique qui accompagnent les biologistes des départements « animaux » (Santé Animale, Génétique Animale) de l'INRA dans le traitement de leurs données haut débit.

2 Présentation du contexte biologique et du sujet

2.1 La métabolomique et la RMN

Les métabolites sont des composés organiques issus du métabolisme ou qui y participent. Ils sont soit produits à partir d'autres métabolites grâce à des réactions chimiques, soit ils proviennent de l'extérieur en étant, par exemple, ingérés. On distingue les métabolites primaires des métabolites secondaires. Les métabolites primaires sont indispensables au fonctionnement de la cellule. Ce sont par exemple des acides aminés, essentiels en tant que constituant des protéines, ou des alcools. Les métabolites secondaires ne participent pas aux processus vitaux de la cellule mais assurent des fonctions importantes. L'ensemble des métabolites constitue le métabolome, et ils ont un effet sur le phénotype final. La métabolomique est l'étude de l'ensemble des métabolites.

Deux approches sont utilisées pour obtenir des données métaboliques : la résonance magnétique nucléaire (RMN) et la spectrométrie de masse. Ces techniques sont complémentaires et permettent de détecter des centaines de métabolites dans divers types d'échantillons. La RMN a un coût relativement abordable et permet donc de comparer le métabolome d'un grand nombre d'individus. Cette technique exploite les propriétés magnétiques de certains noyaux atomiques. En effet, lorsqu'un noyau est soumis à un rayonnement électromagnétique, il peut absorber l'énergie du rayonnement puis la relâcher. L'énergie mise en jeu lors de ce phénomène correspond à une certaine fréquence, qui est convertie en déplacement chimique.
La RMN est une technique très reproductible : si on répète l'expérience plusieurs fois, on obtient quasiment les mêmes résultats. Cependant, elle est moins sensible que la spectrométrie de masse, c'est-à-dire que les métabolites en faible concentration sont plus difficilement détectés.

Les résultats sont produits sous forme de spectres (figure 2), qui peuvent être vus comme la signature métabolique d'un individu, observé dans un état physiologique donné. Dans le spectre, l'axe des abscisses représente le déplacement chimique de la molécule avec son intensité sur l'axe des ordonnées. L'aire sous la courbe au niveau d'un pic est proportionnelle à la quantité de métabolite dans l'échantillon.

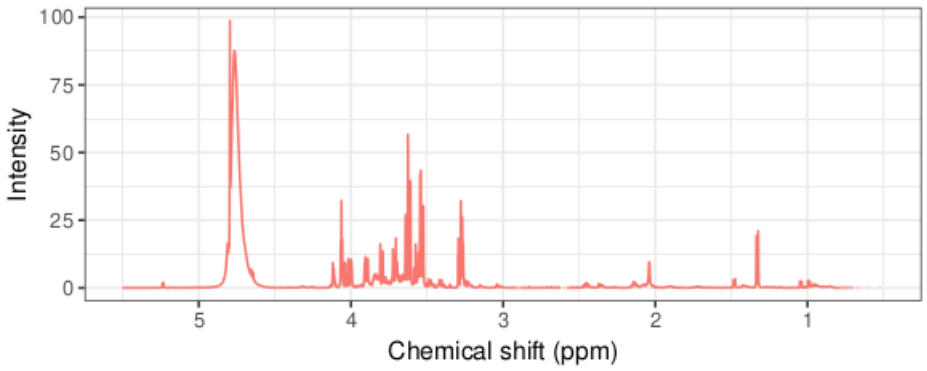


FIGURE 2 – Exemple de spectre de RMN.

La spectroscopie par RMN permet ainsi de connaître les métabolites présents dans un échantillon. Cependant, de telles données restent difficiles à analyser en raison de leur complexité et du grand nombre de signaux générés. En effet, le lien entre le spectre et les métabolites est complexe car un métabolite peut se trouver à un ou plusieurs endroits du spectre. Un métabolite n'est donc pas forcé-

ment identifié par un unique pic dans le spectre. De plus, un pic peut correspondre à un ou plusieurs métabolites (par exemple Leucine et Lysine sur la figure 3).

On ne peut donc pas connaître tous les métabolites d'un échantillon directement, c'est pourquoi il faut utiliser une méthode de quantification. Pour cela, différentes méthodes ont été développées.
Il est d'abord possible d'identifier les métabolites manuellement, avec l'aide d'un spécialiste, mais cela est peu reproductible. Des méthodes automatiques d'identification existent, comme MetaboHunter [1] ou MIDTool [2]. D'autres méthodes permettent également de quantifier automatiquement les métabolites détectés : Autofit [3], batman [4], Bayesil [5] et rDolphin [6].
Mes encadrants de stage ont développé la méthode ASICS [7], basée sur la méthode de Taridvel et al. [8]. Elle permet de quantifier les métabolites dans un spectre à partir d'une librairie de spectres purs et semble plus performante que les autres [7, 8]. Des outils de pré-traitements et post-traitements ont également été inclus.
Obtenir tous les métabolites présents dans un spectre complexe n'est donc pas une étape facile mais elle est très importante.

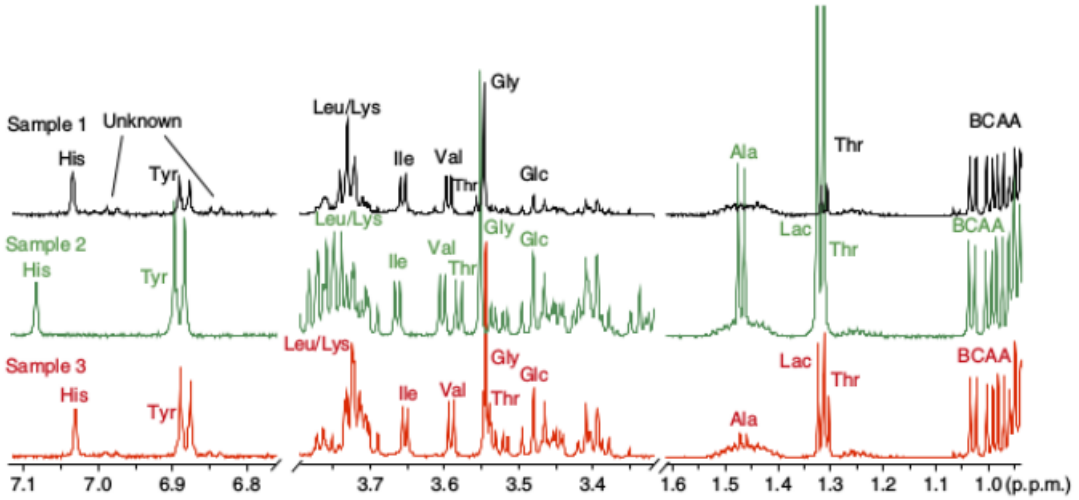


FIGURE 3 – Spectres de RMN de mélanges complexes.

2.2 Problématique du stage

Mon stage s'inscrit dans le cadre du projet SubPig, financé par un métaprogramme INRA, GISA (Gestion intégrée de la santé des animaux, 2018-2019). Ce projet doit permettre de comprendre et de réduire la surmortalité périnatale des jeunes cochons. L'amélioration du bien-être animal est un enjeu majeur du projet. Les enjeux sont également économiques car la surmortalité périnatale a un impact considérable sur la rentabilité de la production porcine.
En effet, d'après les données GTTT (Gestion Technique des Troupeaux de Truies) et IFIP (Institut du Porc) de 2015 en France [9], la mortalité entre la naissance et le sevrage est en moyenne de 13.8% pour les animaux nés vivants et 20% pour les nés incluant les mort-nés. L'espèce porcine a en moyenne 11.7 sevrés par portée pour 13.6 nés vivants. Le coût de la perte d'un porcelet avant le sevrage est évalué à 45€. Il est donc essentiel, d'un point de vue éthique comme économique, de réduire la mortalité des porcelets.
Le projet a pour objectifs scientifiques de déterminer un ensemble de gènes et/ou de métabolites permettant la prédiction de la survie ainsi que la compréhension de l'état métabolique du nouveau-né. L'étude doit aussi définir une procédure de collecte des échantillons qui soit efficace et qui n'affecte pas

le bien-être des animaux.

La problématique générale du stage consistera à analyser des données métabolomiques obtenues par RMN. Il a été observé qu'il y aurait un lien entre la mortalité des porcelets et leur alimentation. On s'intéressera donc à des échantillons de sang prélevés sur des cochons nouveaux-nés. Ces cochons sont issus de plusieurs expériences, dont l'expérience « CMJR » qui étudie deux lignées divergentes pour l'efficacité alimentaire et sur laquelle porte le sujet du stage.
De manière plus précise, l'expérience « CMJR » a consisté à créer deux sous-populations extrêmes pour un caractère d'efficacité alimentaire [10]. Celui-ci, appelé consommation moyenne journalière résiduelle (CMJR), correspond à la différence entre la consommation moyenne journalière (CMJ) observée et la consommation moyenne journalière prédite en fonction de l'estimation des besoins d'entretien et de production de l'animal. Lorsque cette quantité est positive, les animaux ont une consommation supérieure à leurs besoins théoriques et sont qualifiés d'animaux « dépensiers ». Lorsque la différence est négative, les animaux sont qualifiés « d'économes et efficaces ». Pour obtenir des populations contrastées pour ce critère, une expérience a consisté à sélectionner, pendant plusieurs générations, des individus extrêmes (faibles et fortes valeurs) pour ce critère et à les faire se reproduire entre eux : c'est ce que l'on appelle des lignées divergentes. Les animaux étudiés durant ce stage correspondent à la 10^{e} génération de l'expérience et les deux sous-populations (notées G10+ et G10-) montrent des différences marquées dans leur efficacité alimentaire.

Au cours du stage, on cherchera donc à mettre en évidence comment le métabolisme de ces deux sous-populations diffère, afin d'identifier des métabolites caractéristiques des lignées et liés à la mort des porcelets. On recherchera également la meilleure procédure de collecte des échantillons, en s'intéressant à la date et au type de prélèvement.

3 Description des données

On s'intéresse à des échantillons de sang prélevés sur des cochons nouveaux-nés. Du sang a été prélevé au niveau du cordon ombilical dès la naissance puis au niveau de la jugulaire 24 heures plus tard. Les échantillons sont de deux types : plasma ou sérum. Le sérum est le matériel restant après filtrage des globules rouges hors du sang. Le plasma est obtenu par microcentrifugation du sérum pour récupérer les globules blancs. Cependant aucun échantillon de plasma n'a été prélevé 24 heures après la naissance.

Ces échantillons ont été collectés sur deux bandes de mise-bas, c'est-à-dire dans un même élevage mais à deux temps d'élevage différents, séparés de 5 mois. Pour des raisons expérimentales, dans l'une des bandes, il n'y a pas d'échantillons prélevés 24 heures après la naissance.

Les données disponibles se décomposent en plusieurs fichiers :

- **Informations générales sur les animaux** : 120 animaux et 21 variables. Dans ce fichier on dispose, entre autres, du sexe de l'animal, de sa bande et de sa lignée.
- **Plan d'expérience de la RMN** : 290 échantillons et 14 variables. 248 échantillons proviennent d'animaux et 42 de « pools ». Un pool est un mélange de divers échantillons utilisé pour les contrôles. Ici, les pools ont été constitués en fonction du type d'échantillon, de la date de prélèvement et de la lignée de l'animal. On trouve par exemple dans ce fichier le nom de l'animal auquel correspond l'échantillon (qui permet de faire le lien avec le fichier précédent), le type d'échantillon (plasma ou sérum) et la date de prélèvement (naissance ou après 24 heures).
- **Spectres de RMN** : 248 spectres. Les noms des spectres correspondent à ceux des échantillons du fichier précédent.
- **Données de dosages** : 148 échantillons et 40 variables. Il y a 112 échantillons provenant d'animaux et 36 de pools. Ce fichier contient les quantifications directes de certains métabolites par une méthode de référence qui permet des dosages ciblés : seuls les acides aminés ont été dosés. Les échantillons utilisés sont uniquement ceux de plasma prélevés à la naissance.
- **Données phénotypiques** : 120 animaux et 12 variables. On dispose dans ce fichier de données sur le phénotype des animaux, avec des variables comme le poids à la naissance, la longueur ou la température.

4 Description des méthodes de quantification

4.1 Découpage en buckets

L'approche classique pour traiter des données issues de RMN est d'abord de découper les spectres en intervalles réguliers appelés *buckets*. Ensuite, l'aire sous la courbe est calculée pour chaque *bucket* et des analyses statistiques sont réalisées sur ces nouvelles variables. Cependant, les *buckets* ne sont pas directement liés aux métabolites puisqu'un métabolite peut avoir plusieurs pics. Il est donc nécessaire qu'un expert de la RMN identifie manuellement les *buckets* issus de l'analyse pour pouvoir interpréter biologiquement les résultats. Cette identification est longue, fastidieuse, dépend de l'expert et est difficilement reproductible. L'approche par découpage en *buckets* reste donc perfectible.

4.2 Le package ASICS

Des méthodes ont été développées pour identifier et quantifier automatiquement la concentration des métabolites dans un spectre RMN de manière reproductible. Pour une revue de ces méthodes, on peut consulter l'introduction de l'article de Tardivel et al. [8] qui définit également une nouvelle méthode appellée ASICS.
Afin que ces méthodes soient performantes, il est également nécessaire qu'elles soient combinées à des étapes de pré-traitements et de post-traitement. Le package R **ASICS** (*Automatic Statistical Identification in Complex Spectra*) a été développé dans cette optique : il associe une méthode efficace pour faire l'identification et la quantification de métabolites à un ensemble d'étapes nécessaires de pré et post-traitement.

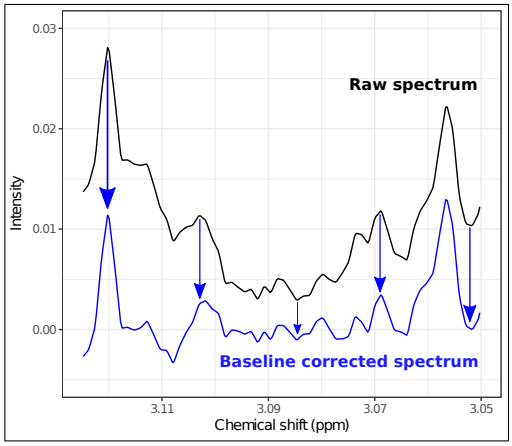
4.2.1 Pré-traitements du spectre d'un échantillon

Les sorties après RMN sont fournies sous forme de signaux appelés FID. Ils subissent des pré-traitements et sont transformés en spectres par la transformée de Fourier. J'ai réalisé ces différentes étapes avec les outils du package **PepsNMR**, inclus dans **ASICS**.
Après l'import des spectres, plusieurs pré-traitements sont recommandés pour supprimer les biais techniques. Il est possible avec **ASICS** de faire la correction de la ligne de base (figure 4(a)), l'alignement des pics des spectres entre eux (figure 4(b)) et de normaliser par l'aire sous la courbe (figure 4(c)).

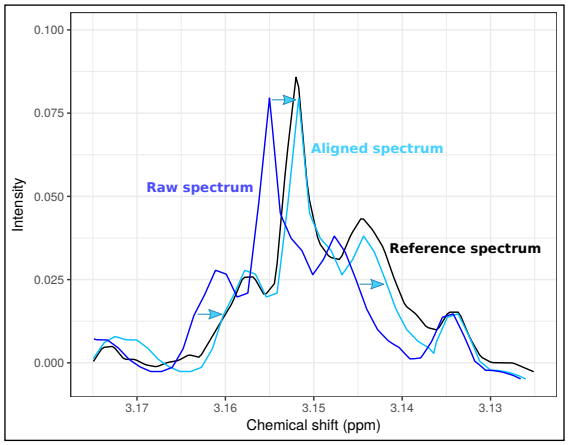
4.2.2 Pré-traitements de la librairie de référence

Une librarie de spectres de métabolites purs, c'est-à-dire de spectres composés d'un seul métabolite, est utilisée pour identifier et quantifier les métabolites dans un mélange complexe (composé de plusieurs métabolites). Cette librarie de référence est disponible dans le package. Cependant, comme pour les spectres des mélanges complexes, des pré-traitements sont nécessaires.

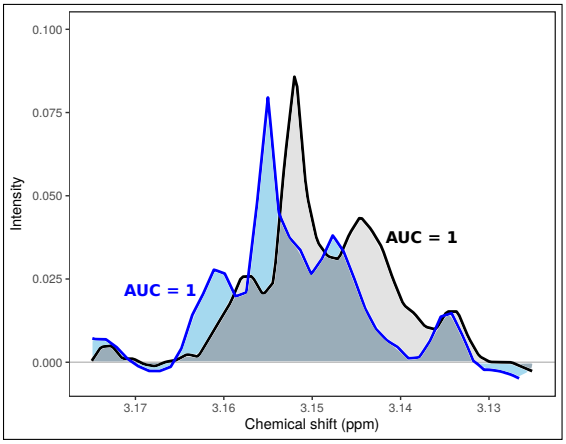
- Suppression du bruit : cela permet de déterminer plus facilement la position des pics dans les prochaines étapes.
- Première étape de sélection : un métabolite ne peut pas appartenir à un spectre complexe si tous ses pics ne sont pas présents dans le spectre. Il peut aussi y avoir des biais techniques qui font varier les déplacements chimiques des spectres. Un spectre de la librairie de référence est donc conservé uniquement si tous ses pics sont présents dans le mélange complexe et que le décalage horizontal entre deux spectres ne dépasse pas une certaine limite.



(a) Correction de la ligne de base



(b) Alignement des pics



(c) Normalisation

FIGURE 4 – Pré-traitements des spectres d'un échantillon.

- Translation et déformation : pour réaliser la quantification, il est nécessaire d'aligner les spectres de la librairie de référence avec le mélange complexe.

4.2.3 Quantification relative des concentrations des métabolites

En utilisant le mélange complexe et la librairie de référence pré-traitée, il est désormais possible de quantifier les métabolites. Le mélange complexe est défini comme une combinaison linéaire des spectres de la librairie de référence :

$$g(t) = \sum_{i=1}^{p} \beta_i f_i(\Phi_i(t)) + \epsilon(t) \tag{1}$$

avec $\beta_i \geq 0$, où g correspond au mélange complexe, $f_i \circ \Phi_i$ aux spectres de la librairie, $\beta = (\beta_1, ..., \beta_p)$ aux coefficients associés à ces spectres et ϵ au bruit. Les coefficients $(\beta_i)_i$ sont ensuite convertis en concentrations relatives des différents métabolites pour chaque échantillon. Les résultats sont fournis sous la forme d'un tableau de données avec les spectres en colonne et les métabolites en ligne.

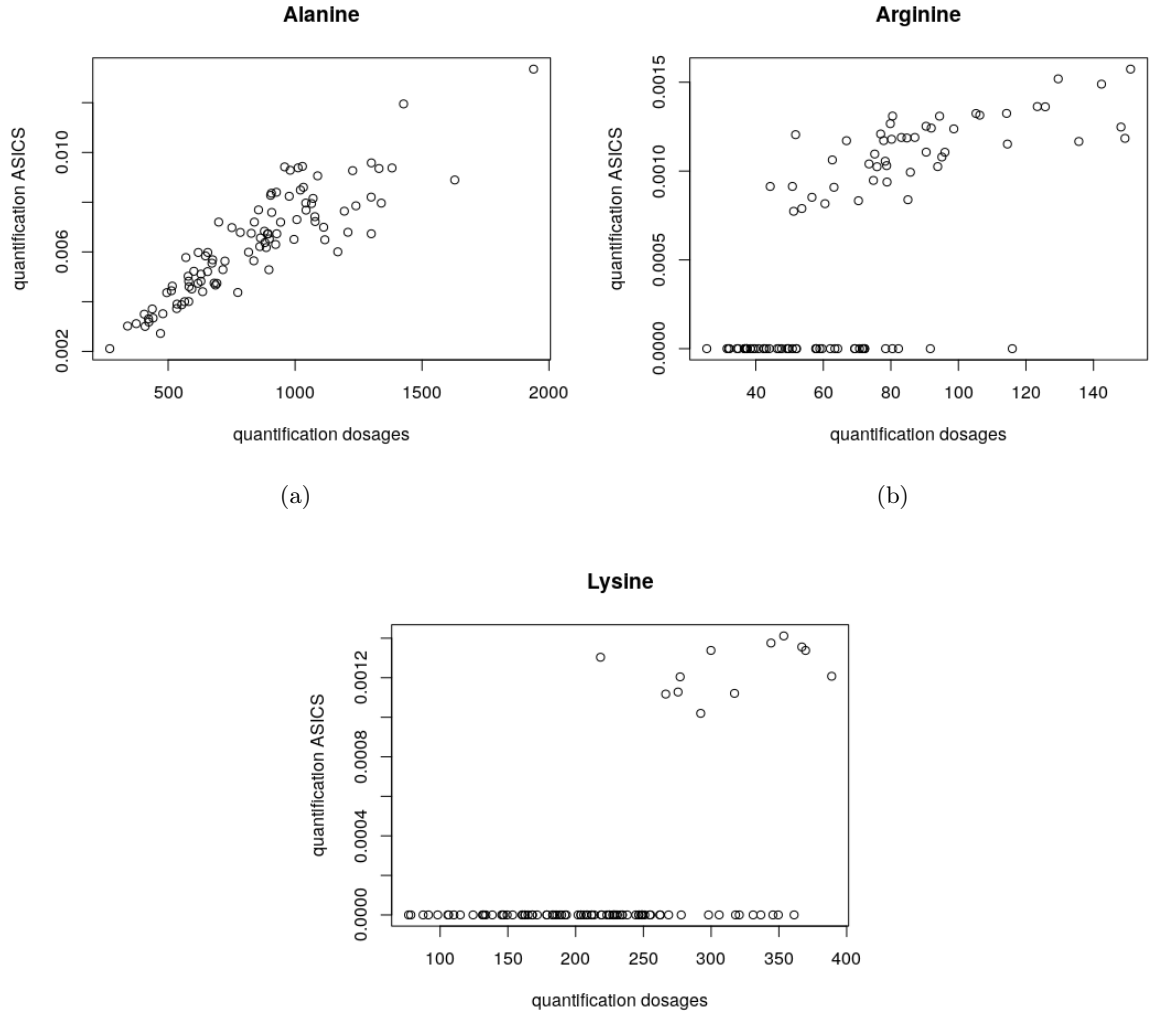
Cette méthode a été appliquée pour la quantification des spectres des animaux et des pools. 99 métabolites ont été identifiés dans les spectres des animaux et 73 pour les pools. Les quantifications des spectres des animaux et des pools ont été effectuées séparément afin de ne pas fausser les résultats. En effet, les pools doivent permettrent de contrôler les résultats des animaux. Les pools étant construits par type, date et lignée, on doit retrouver des similitudes entre les échantillons des pools et les échantillons

d'animaux de même caractéristiques.

Pour comparer les résultats avec l'approche classique, **ASICS** permet aussi le découpage des spectres en *buckets*.

4.3 Contrôle des quantifications

Afin de valider les quantifications **ASICS**, on peut les comparer avec celles effectuées par dosages. Pour chaque métabolite détecté par les deux méthodes, on peut calculer un coefficient de corrélation entre les quantifications estimées par **ASICS** et les quantifications mesurées par dosage direct. Plus la corrélation est forte, meilleure est la quantification. On peut également utiliser un nuage de points pour visualiser la corrélation.



(a) (b)

(c)

FIGURE 5 – Nuages de points des acides aminés : certains métabolites sont très bien quantifiés 5(a) mais les faibles concentrations sont mal détectées 5(b) et la lysine est difficile à quantifier 5(c).

Les métabolites quantifiés par dosages sont uniquement des acides aminés. Il y en a 19 qui sont communs aux quantifications par **ASICS** et par dosages. Pour certains acides aminés, comme l'alanine (figure 5(a)), le coefficient de corrélation est supérieur à 0.80 et le nuage de points montre bien la forte corrélation.

On remarque cependant que les faibles concentrations sont mal détectées par **ASICS**. Cela est visible sur le nuage de points de l'arginine (figure 5(b)). Les faibles concentrations dans les dosages sont nulles avec **ASICS** alors que des concentrations plus élevées sont correctement quantifiées. Cela s'explique en partie par le fait que la RMN ne détecte pas les métabolites en faible concentration, qui ne pourront donc pas être quantifiés.

Enfin, la lysine (figure 5(c)) est très peu présente dans les quantifications **ASICS**, même pour des concentrations élevées. Cela s'explique par le fait que le spectre de la lysine possède beaucoup de pics à différents endroits et qu'ils sont confondus avec d'autres métabolites. Cela rend la lysine difficile à détecter.

Le tableau 1 évalue la qualité de la quantification à partir des nuages de points et la compare au coefficient de corrélation et à la quantification moyenne par dosages. Comme on l'a vu dans les nuages de points de l'arginine et de la lysine, il y a un certain nombre de métabolites qui sont souvent quantifiés à 0, à cause de la faible sensibilité de la RMN aux petites concentrations.

Quantification des acides aminés	**Corrélation**	**Quantification moyenne par dosage**
5 très bien	> 0.70	443.46
3 très bien avec 0	$\geqslant 0.69$	96.78
1 moyen	0.58	81.74
3 moyens avec 0	$[0.47\,;\,0.53]$	107.72
7 mauvais avec 0	< 0.40	69.91

TABLE 1 – Tableau récapitulatif de la comparaison entre les quantifications ASICS et les quantifications par dosages pour les 19 acides aminés

Validation des quantifications ASICS

On a utilisé la méthode ASICS pour faire la quantification des métabolites à partir des spectres de RMN. On a comparé les résultats de cette quantification avec ceux de la quantification par dosages. Pour cela on a utilisé des nuages de points et des coefficients de corrélation.
On voit que la corrélation est liée à la quantification : plus la quantification est élevée, meilleure est la corrélation. Les faibles corrélations viennent du fait que le métabolite n'est pas détecté par la RMN donc il ne peut pas être quantifié dans ASICS. De plus, les spectres des acides aminés sont complexes, c'est pourquoi ils sont difficiles à quantifier. On peut donc valider l'utilisation de la méthode ASICS pour quantifier les métabolites, tout en étant conscient de son manque de précision pour les faibles concentrations.

5 Analyse exploratoire

L'analyse exploratoire a pour double but de vérifier que les échantillons sont bien différents pour les variables du plan d'expérience (lignée G10+/-, type de prélèvement, date du prélèvement) et de détecter d'éventuels biais expérimentaux comme des échantillons atypiques ou bien des biais techniques non souhaités qui sépareraient les échantillons selon le sexe ou la bande d'élevage par exemple. Pour cela, on réalise une Analyse en Composantes Principales (ACP) où les échantillons sont les individus et les quantifications des métabolites sont les variables. En colorant les points du graphique des individus en fonction des critères recherchés, on peut mettre en évidence l'effet des facteurs.

L'ACP a été mise en œuvre sur R grâce à la fonction **pca** du package **ASICS**.

5.1 L'Analyse en Composantes Principales (ACP)

L'ACP est une méthode d'analyse de données qui a pour but de synthétiser les informations contenues dans un jeu de données tout en minimisant la perte d'information. L'ACP peut être vue comme un changement de base ou bien comme une méthode de réduction de dimension par projection linéaire qui préserve au mieux l'inertie des données initiales. La réduction de la dimension fournit une représentation graphique facilement interprétable des données.

5.2 Étude des quantifications ASICS

5.2.1 Étude sur l'ensemble des individus

L'ACP est, dans un premier temps, réalisée pour l'ensemble des échantillons, c'est à dire sur 248 individus. On identifie rapidement trois échantillons atypiques, qui seront par la suite retirés de l'analyse. L'ACP est donc effectuée sur 245 individus et 99 variables. Les variables sont les métabolites qui ont été quantifiés par ASICS.

La coloration sur le graphique des individus de l'ACP doit permettre d'observer si des groupes se forment et s'ils correspondent à une covariable du jeu de données. Trois critères sont d'intérêt d'un point de vue biologique :

- **Le type d'échantillon** : plasma ou sérum
- **La date de prélèvement** (origine de l'échantillon) : à la naissance (cordon) ou 24 heures après la naissance (jugulaire24h)
- **La lignée** : animaux avec une meilleure efficacité alimentaire (G10-) ou animaux peu efficaces (G10+)

et trois critères correspondent à de possibles biais expérimentaux ou facteurs d'influence :

- **La bande d'élevage** : 1801 ou 1808
- **Le sexe** : mâle (M) ou femelle (F)
- **La mère**

L'ACP sur les quantifications de l'ensemble des échantillons sépare les individus en deux groupes distincts. Grâce à la coloration des points, on identifie que ces groupes correspondent à la date de

prélèvement des échantillons (figure 6). Comme attendu, la date du prélèvement de l'échantillon a un effet, qui correspond à l'évolution physiologique du porcelet à partir de la naissance.
Sur le graphique des variables, on voit qu'il y a un plus grand nombre de métabolites qui contribuent positivement à la dimension 1 (valine, proline, tyrosine...) que négativement (fructose, glycogen, glycerol).
En colorant suivant les autres critères, on constate un petit effet lié à la bande de l'individu (annexe A). Par contre, le sexe de l'individu n'a pas d'effet (annexe B).
Pour vérifier l'influence de la mère, on colore également les individus suivant ce critère. On voit clairement (annexe C) qu'il y a un effet de la mère. Il faudra en tenir compte plus tard pour ne pas introduire de biais.

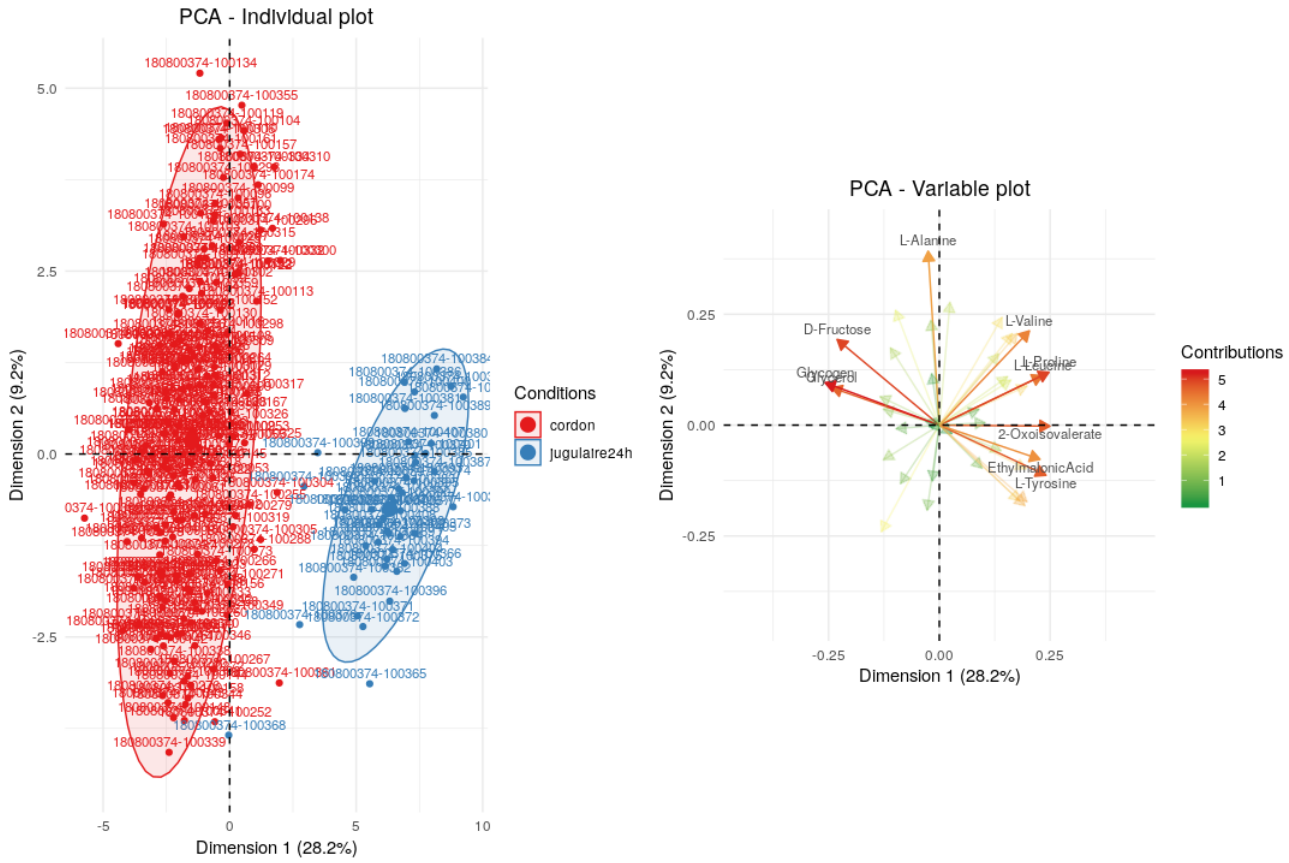


FIGURE 6 – ACP sur les quantifications de l'ensemble des individus, avec une coloration par date de prélèvement de l'échantillon.

La différence entre les groupes étant nette sur la figure 6, on séparera les données suivant la date de prélèvement de l'échantillon pour la suite de l'étude. Cela permettra de mieux visualiser les résultats sur chacun des groupes et de déterminer l'effet des autres facteurs.

5.2.2 Étude sur les individus « jugulaire24h »

Nous nous intéresserons ici aux échantillons prélevés 24 heures après la naissance des animaux, appelés « jugulaire24h ». Nous disposons donc de 47 individus et de 99 variables.

Une ACP et des colorations par critères sont effectuées, de la même manière que précédemment. On observe que l'axe 1 sépare les individus selon leur lignée (figure 7). Dans le graphique des variables, les variables corrélées positivement avec l'axe 1, qui donc correspondent à la lignée G10+ d'animaux peu

efficaces, sont des acides aminés.

On peut également noter que le métabolite myo-inositol est très liée aux individus de la lignée G10-, qui sont les animaux efficaces. C'est un métabolite qu'il sera intéressant d'étudier par la suite car il est connu chez les nouveaux-nés pour refléter la maturité et recommandé comme additif chez des prématurés [11].

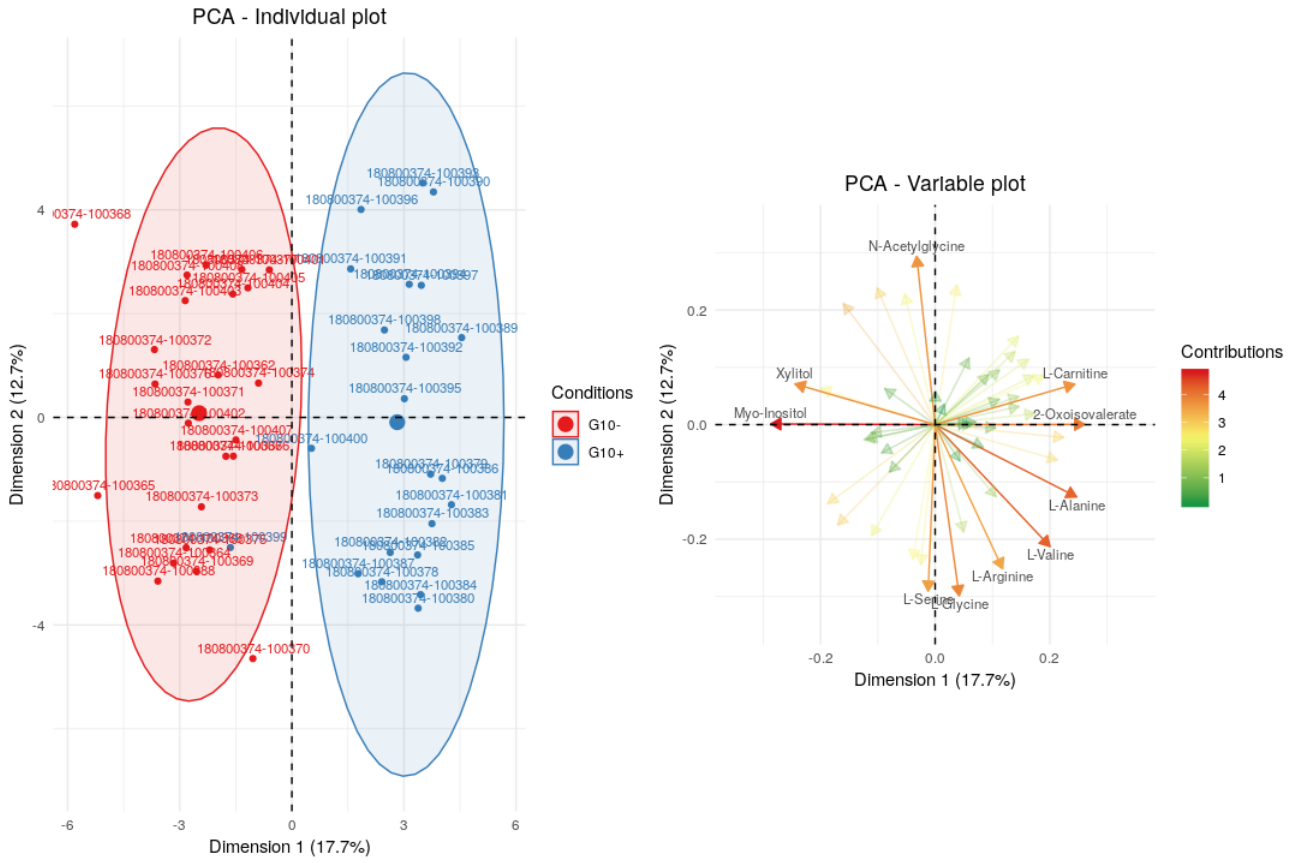
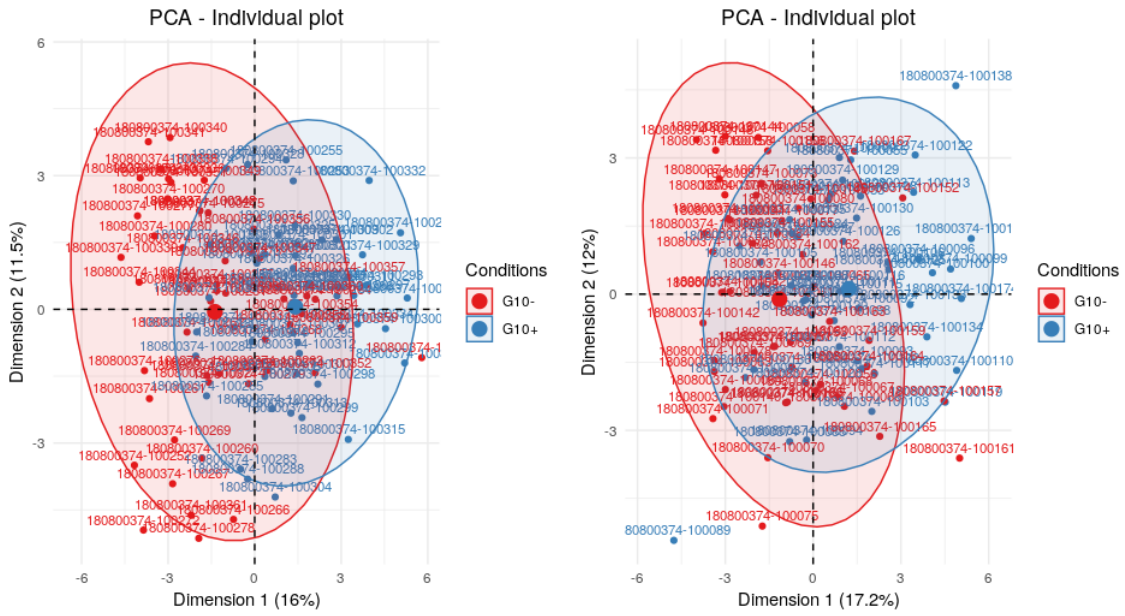


FIGURE 7 – ACP sur les quantifications des individus prélevés 24 heures après la naissance, avec une coloration par lignée.

5.2.3 Étude par type d'échantillon

Il est également intéressant de séparer les données par type d'échantillon, c'est-à-dire plasma ou sérum, avant d'étudier séparément les dates de prélèvement (naissance ou 24h) afin de savoir si l'un des types permet de mieux différencier les lignées. Les échantillons de plasma ne sont disponibles que pour les prélèvements à la naissance alors que les échantillons de sérum sont disponibles pour les deux dates de prélèvement.

L'ACP sur les échantillons de sérum prélevé après 24 heures montre que les lignées sont très bien séparées (figure 7). En revanche, pour les prélèvements à la naissance, le sérum et le plasma donnent des résultats similaires et les lignées se distinguent assez peu (figure 8). On peut en conclure que, à la naissance, aucun type ne semble être meilleur pour distinguer les lignées. Les lignées se différencient mieux dans le sérum à 24 heures.



(a) Sérum (b) Plasma

FIGURE 8 – ACP sur les quantifications des individus prélevés à la naissance de sérum 8(a) et de plasma 8(b), avec une coloration par lignée.

5.3 Comparaison avec le découpage en buckets

La même analyse a été reproduite sur les quantifications obtenues suite au découpage en *buckets* des spectres. Chaque spectre est découpé en 941 intervalles grâce à la fonction **binning** de **ASICS**.

Les ACP avec coloration sont réalisées comme précédemment, en séparant les données par date de prélèvement et par type de d'échantillon (annexe D). La séparation entre les lignées pour les échantillons prélevés 24 heures après la naissance est visible, mais moins nette que précédemment (figure 9).
La similarité de ces résultats avec les précédents montre la cohérence de la quantification effectuée par **ASICS**.

5.4 Étude des quantifications par dosages

Pour contrôler les résultats obtenus sur les données issues des quantifications par **ASICS**, on réalise une analyse exploratoire sur les données des quantifications par dosages. L'analyse est réalisée sur les 107 échantillons quantifiés par dosages et pour lesquels nous disposons d'informations sur l'animal. Ces échantillons sont les mêmes que ceux de l'analyse des quantifications **ASICS**, et concernent les prélèvements de plasma à la naissance. Les variables sont les 29 acides aminés qui ont été quantifiés par les dosages.

On note un effet lié à la bande (figure 10(a)) et une absence d'effet lié au sexe de l'animal (annexe E), comme dans les données de quantifications **ASICS**. L'effet de la lignée est peu visible (figure 10(b)). Cela est cohérent avec les résultats précédents car ce sont les échantillons de sérum prélevés 24 heures après la naissance qui permettent le mieux de distinguer les lignées. Or, la quantification par dosages a été faite sur les échantillons de plasma prélevés à la naissance. De plus, on a vu que le myo-inositol permet la séparation des lignées mais il n'est pas présent dans les dosages car ce n'est pas un acide aminé.

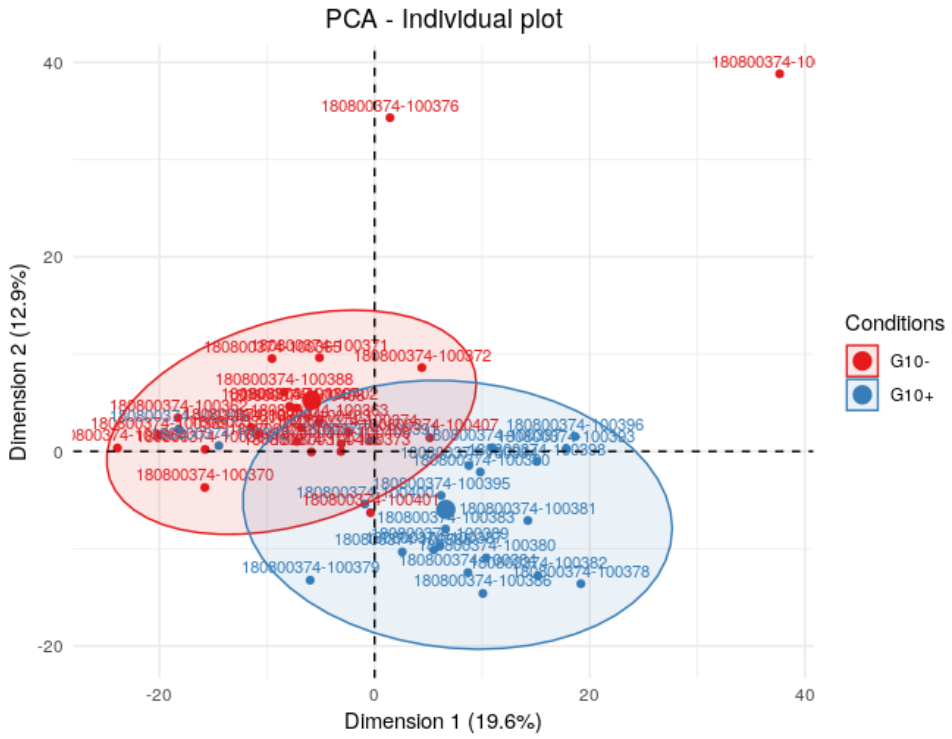
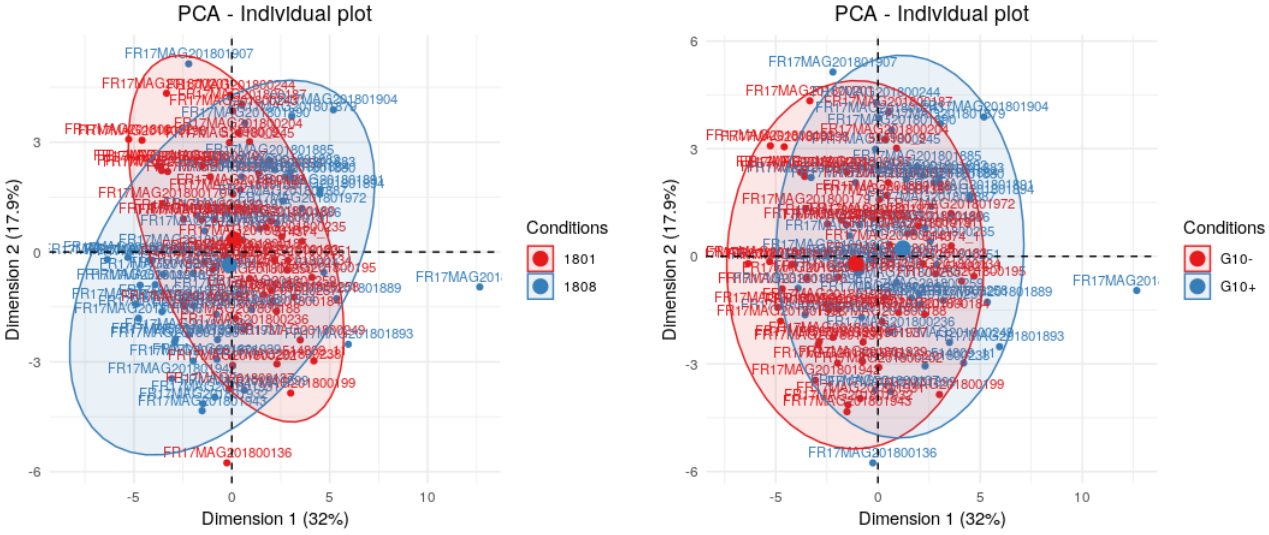


FIGURE 9 – ACP sur les *buckets* des individus prélevés 24 heures après la naissance, avec une coloration par lignée.



(a) Bande

(b) Lignée

FIGURE 10 – ACP sur les quantifications par dosages, avec une coloration par bande 10(a) et par lignée 10(b).

Un effet lignée visible à 24 heures

Des analyses en composantes principales ont été réalisées afin d'identifier des différences entre échantillons, liées à certaines covariables du jeu de données. Il a été identifié dans les données issues de la quantification **ASICS** qu'il y a des différences entre les prélèvements effectués dès la naissance des animaux et 24 heures après. La séparation entre les lignées est meilleure après 24 heures, ce qui montre une divergence physiologique des lignées après la naissance. En outre, sur les données des prélèvements effectués 24 heures après la naissance, il a été remarqué que le métabolite myo-inositol est caractéristique des individus de la lignée d'animaux efficaces.
Ces résultats ont été confirmés par l'analyse issue du découpage des spectres en *buckets* puis par l'analyse des dosages.
Afin de mieux comprendre quels sont les métabolites qui sont significativement différents pour des covariables d'intérêt, on va ensuite réaliser des tests d'hypothèses avec correction des tests multiples.

6 Étude des métabolites

6.1 Test de Kruskal-Wallis et correction des tests multiples

Pour mettre en évidence les différences entre deux conditions, on réalise des tests d'hypothèses. Le test de Shapiro-Wilk a d'abord été utilisé pour regarder la normalité des données. L'hypothèse nulle de normalité a été réjetée pour tout les tests (p-valeur ≤ 5 %) donc le test paramétrique de Student ne peut être utilisé.
On choisit ici d'appliquer un test de Kruskal-Wallis car il ne nécessite pas d'hypothèse de normalité des données. C'est un test non paramétrique et une généralisation du test de Wilcoxon. Il permet de comparer la distribution d'une variable quantitative d'intérêt entre plusieurs groupes. Ici, les groupes correspondent aux lignées et la variable d'intérêt est la concentration en métabolite. On teste l'hypothèse nulle H_0 « les lignées ne sont pas différentes pour le métabolite étudié » contre l'hypothèse alternative H_1 « les lignées sont différentes pour le métabolite étudié ». On va donc reproduire le test autant de fois qu'il y a de métabolites.

Cependant, quand un test est répété un grand nombre de fois pour tester des hypothèses indépendantes, le nombre de résultats qui sont des faux positifs (c'est-à-dire le nombre de fois où l'hypothèse nulle est rejetée à tort) augmente et la probabilité d'avoir au moins un faux positif parmi l'ensemble des tests se rapproche de manière rapide de 1. Il faut donc appliquer une correction pour les tests multiples. Pour contrôler les faux positifs, on peut se baser sur le contrôle :

- du FWER (*family-wise error rate*) : probabilité d'avoir au moins un faux-positif
- du FDR (*false discovery error*) : proportion de faux-positifs

Le principe consiste à calculer une p-valeur ajustée, c'est-à-dire une p-valeur corrigée par un facteur d'ajustement. La méthode généralement utilisée pour cela en biologie est celle de Benjamini et Hochberg [12]. Celle-ci permet de contrôler le FDR.

Les tests de Kruskal-Wallis avec correction des tests multiples par la méthode de Benjamini et Hochberg ont été mis en œuvre sur R avec la fonction **kruskalWallis** du package **ASICS** sur les données issues des quantifications et des dosages.

6.2 Étude des métabolites significatifs

Les tests permettent d'identifier les métabolites qui sont significativement différents en fonction de la lignée. On compare ensuite ces métabolites suivant les types d'échantillons et les dates de prélèvement et avec les dosages. On trace des boxplots de concentration par lignée pour chaque métabolite significatif dans les tests. On pourra ensuite comparer ces boxplots.

On étudie donc les métabolites dans les 4 cas suivants :

- Quantifications ASICS : prélèvements de sérum à la naissance
- Quantifications ASICS : prélèvements de plasma à la naissance
- Quantifications ASICS : prélèvements de sérum après 24 heures
- Dosages : prélèvements de plasma à la naissance

Le diagramme de Venn (figure 11) montre le nombre de métabolites significatifs en commun entre les différents cas suite aux tests de Kruskal-Wallis. On remarque qu'il y a 8 métabolites communs au plasma et sérum à la naissance, ce qui montre que les deux types se ressemblent.

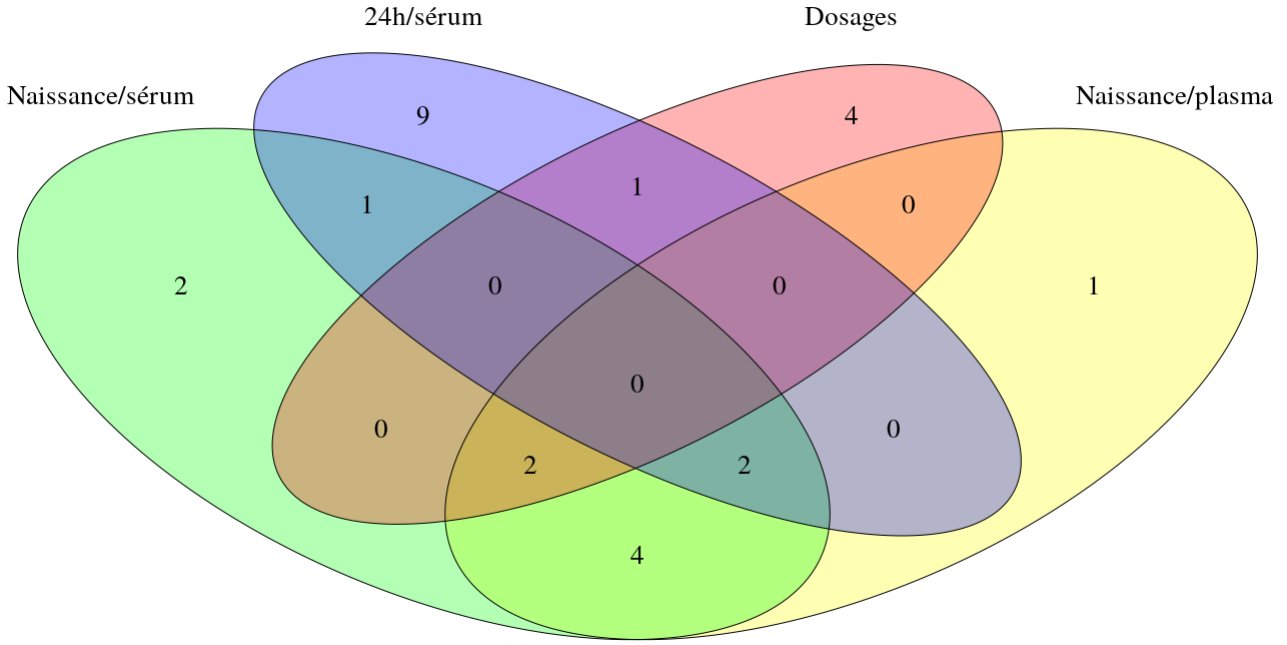


FIGURE 11 – Diagramme de Venn du nombre de métabolites significatifs dans les tests de Kruskal-Wallis.

Ce sont les prélèvements de sérum à 24h qui ont le plus de métabolites significatifs. La majorité de ces métabolites n'est pas significative dans les autres cas et le myo-inositol en fait partie.

On voit dans les boxplots (figure 12) du sérum à la naissance, où le myo-inositol n'est pas significatif, que la quantité du métabolite dans les deux lignées est proche. En revanche, dans le sérum à 24 heures, il est significatif et on observe une diminution de la quantité chez les G10+ alors qu'elle reste stable chez les G10-. Ainsi les boxplots sont différenciées entre les lignées.

La glutamine aussi est uniquement significative dans le sérum à 24 heures suite au test de Kruskal-Wallis. Quand on compare les boxplots du sérum à 24 heures avec celles à la naissance (figure 13), on remarque que la concentration de glutamine augmente chez les G10- à 24 heures alors qu'elle reste stable chez les G10+. La glutamine est un nutriment très utilisé par les cellules du tube digestif et elle aurait un effet bénéfique sur le système immunitaire du tube digestif [13]. Or la maturation du tube digestif et de son système immunitaire est un facteur majeur pour la santé et la croissace des jeunes porcelets, notamment dans les premières 24 heures après la naissance. Les concentrations plus élevées dans la lignée des animaux les plus efficaces (G10-) pourraient donc constituer un avantage pour eux.

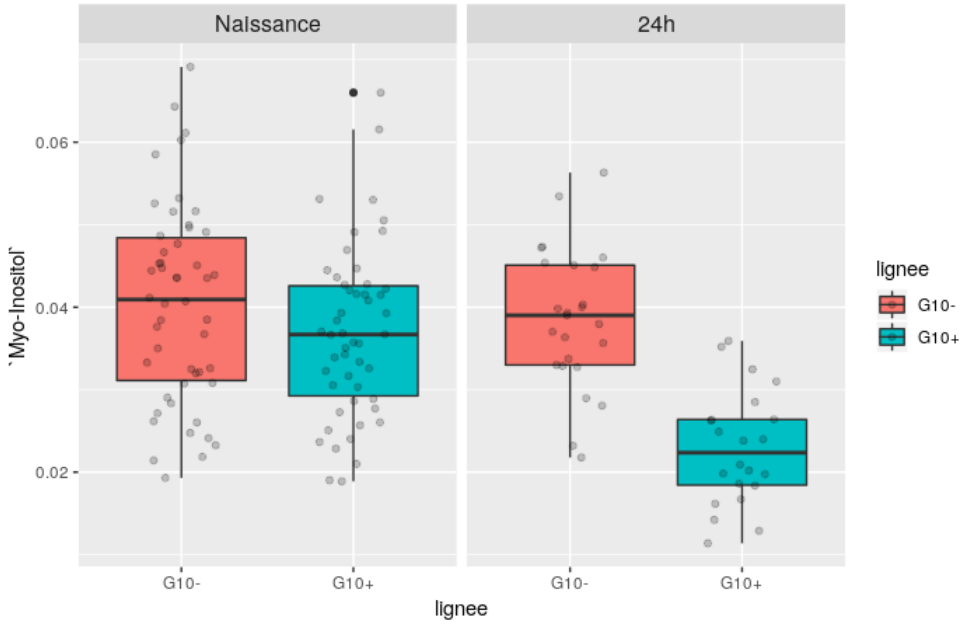


FIGURE 12 – Quantité de myo-inositol par lignée dans les prélèvements de sérum à la naissance (non significatif) et à 24h (significatif).

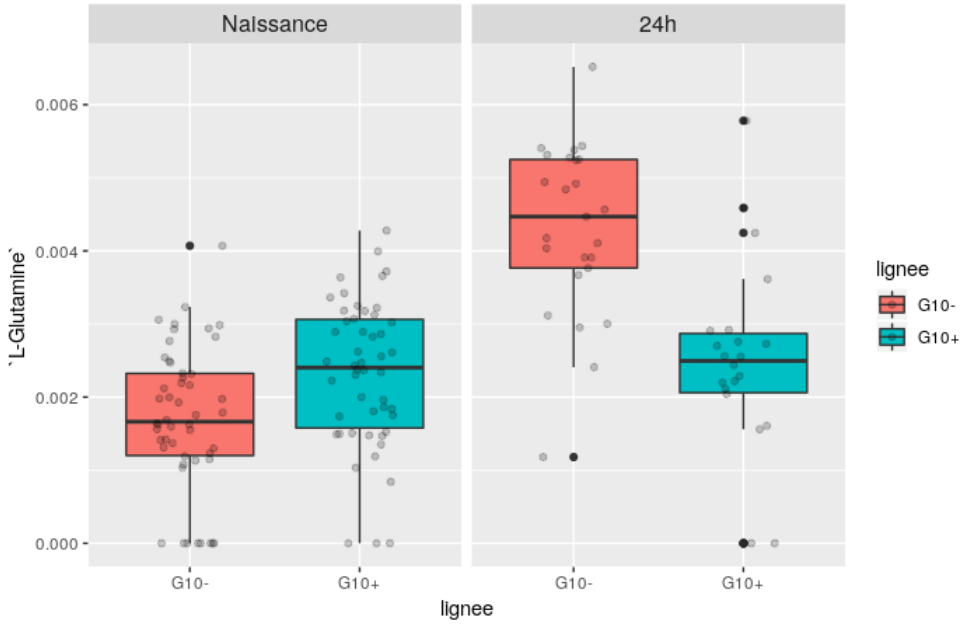


FIGURE 13 – Quantité de glutamine par lignée dans les prélèvements de sérum à la naissance (non significatif) et à 24h (significatif).

L'alanine est statistiquement significative dans les prélèvements de sérum à la naissance et à 24 heures mais pas dans les prélèvements de plasma à la naissance. Dans le boxplot du sérum à 24 heures, les lignées sont très bien différenciées et la distance inter-quartile est faible. L'alanine n'est pas statistiquement significative dans le plasma à la naissance et les boxplots des lignées sont moins différenciées mais on observe que la tendance est toujours visible (figure 14). L'alanine est un acide aminé qui a un rôle d'échange entre les tissus, comme transporteur de carbone et d'azote. Il peut aussi servir à la synthèse du glucose s'il est transporté dans le foie.

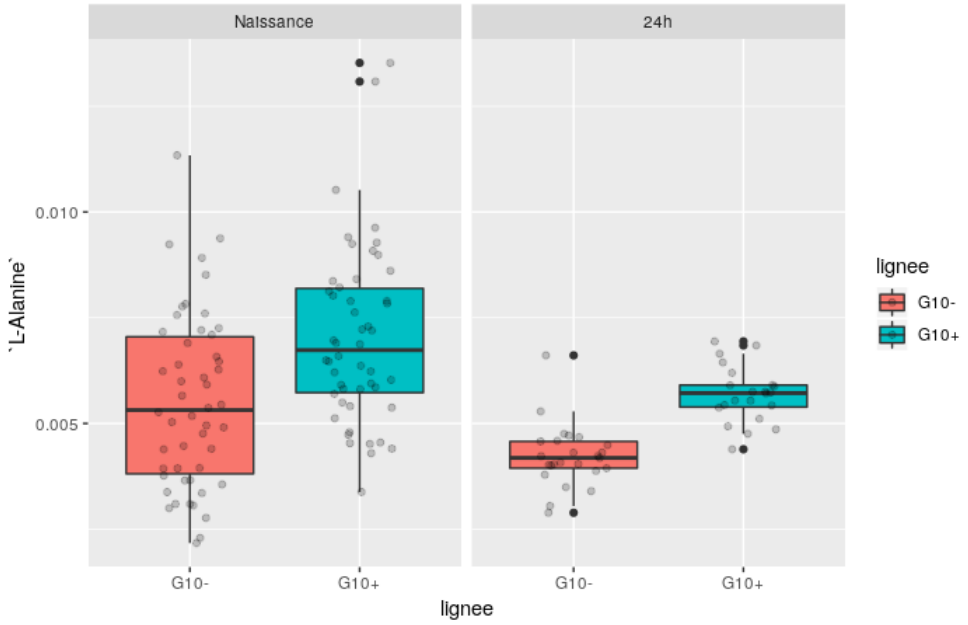


FIGURE 14 – Quantité d'alanine par lignée dans les prélèvements de sérum à la naissance (non significatif) et à 24h (significatif).

6.3 Modèles mixtes

On construit également des modèles mixtes afin de mieux comprendre les effets sur les métabolites des phénomènes comme la lignée, la date de prélèvement, de type de prélèvement et l'aléa lié à la mère, ainsi que leurs interactions. Ce sont des modèles qui comportent à la fois des facteurs fixes et des facteurs aléatoires [14]. On construit d'abord un modèle avec tous les phénomènes, que l'on compare avec un modèle réduit composé uniquement de l'aléatoire. Pour tous les métabolites pour lesquels le modèle complet est significativement différent du modèle nul, on construit des sous-modèles et on en sélectionne le meilleur par utilisation du critère BIC [15]. Cette procédure a déjà été utilisée et est décrite dans l'article de Voillet et al. [16].
Le *Bayesian Information Criterion* est définit par :

$$BIC = -2\ln(L) + k\ln(N) \tag{2}$$

avec L la vraisemblance du modèle, k le nombre de paramètres et N le nombre d'observations. Le modèle sélectionné est celui qui minimise le critère BIC.

On va répéter la procédure de construction et sélection de modèles d'abord avec les effets date, lignée et mère puis avec les effets type, lignée et mère. L'objectif est d'étudier l'impact de la date et du type d'échantillon sur les métabolites, pour déterminer s'ils sont importants à prendre en compte dans le processus de collecte des échantillons.

Le modèle complet et les sous-modèles sont :

- $y_{complet} = \alpha_{lignee} + \beta_{date} + \gamma_{lignee/date} + \delta_{mere} + \varepsilon$
- $y_{lignee,date} = \alpha_{lignee} + \beta_{date} + \delta_{mere} + \varepsilon$
- $y_{lignee} = \alpha_{lignee} + \delta_{mere} + \varepsilon$
- $y_{date} = \beta_{date} + \delta_{mere} + \varepsilon$

avec ε l'erreur du modèle, qui suit une loi normale.

Puis on construit aussi ces modèles en remplaçant la date (naissance ou 24h) par le type (sérum ou plasma) :

- $y_{complet} = \alpha_{lignee} + \beta_{type} + \gamma_{lignee/type} + \delta_{mere} + \varepsilon$
- $y_{lignee,type} = \alpha_{lignee} + \beta_{type} + \delta_{mere} + \varepsilon$
- $y_{lignee} = \alpha_{lignee} + \delta_{mere} + \varepsilon$
- $y_{type} = \beta_{type} + \delta_{mere} + \varepsilon$

avec ε l'erreur du modèle, qui suit une loi normale.

Ce travail est réalisé sur R avec le package **nlme**.

6.4 Analyse des résultats

La construction des modèles mixtes a permis d'identifier, pour chaque métabolite significatif dans les modèles, le meilleur des sous-modèles. Les tableaux suivants (table 2) présentent pour chaque modèle le nombre de métabolites pour lesquels le modèle est le meilleur. Les modèles $y_{complet}$ et y_{date} ou y_{type} sont les meilleurs pour la plupart des métabolites. Aucun métabolite n'a pour meilleur modèle celui de la lignée seule.

	$y_{complet}$	$y_{lignee,date}$	y_{lignee}	y_{date}
Nombre de métabolites	26	3	0	30

	$y_{complet}$	$y_{lignee,type}$	y_{lignee}	y_{type}
Nombre de métabolites	2	2	0	17

TABLE 2 – Nombre de métabolites significatifs par modèle

On compare l'information donnée par les modèles à celle des boxplots pour mieux comprendre le rôle des métabolites.
Le meilleur modèle pour le myo-inositol est y_{date}. On voit en effet sur les boxplots (figure 12) qu'il y a un effet lié à la date. En fonction de la date, les lignées sont plus ou moins différenciées.
Dans les boxplots de la glutamine (figure 13), on avait remarqué un effet lignée et date. Cela est cohérent avec le modèle sélectionné qui est $y_{complet}$.
Le modèle y_{date} est le meilleur pour l'alanine. L'effet de la date était en effet visible dans les boxplots (figure 14).

On met en évidence avec ces résultats l'importance de la date de prélèvement. Cela permet également d'obtenir des conclusion intéressantes d'un point de vue biologique, notamment sur la présence du myo-inositol chez les G10- (animaux efficaces) qui est cohérente avec les informations bibliographiques sur l'intérêt du myo-inositol chez les prématurés.

Des métabolites significativement différents entre les lignées

La réalisation de tests de Kruskal-Wallis avec correction des tests multiples a permis d'identifier des métabolites significativement différents pour les lignées. On a comparé ces métabolites entre les différents types et dates de prélèvement et on a construit des boxplots avec les quantités par lignée.

On a ensuite construit des modèles mixtes pour modéliser les effets de la lignée, de la date, du type de prélèvement et de l'aléa de la mère. Pour chaque métabolite, on a sélectionné le meilleur modèle et on a comparé la cohérence avec les boxplots.

Avec ces analyses, on peut confirmer que les échantillons prélevés 24 heures après la naissance sont meilleurs pour distinguer les lignées. Cela permet d'identifier des métabolites qui sont caractéristiques de ce type d'échantillon et significativement différents pour les lignées.

Cependant, les boxplots ne sont pas plus différenciées entre les lignées pour les prélèvements de plasma que de sérum. De plus, le diagramme de Venn confirme que les métabolites significatifs sont en majorité les mêmes dans les deux types. Cela signifie donc qu'il est possible de faire des prélèvements de sérum au lieu de plasma sans impacter les résultats. L'intérêt pratique est qu'il est plus simple d'obtenir du sérum que du plasma.

7 Étude des phénotypes

Le phénotype est l'ensemble des caractéristiques observables d'un individu, dus aux facteurs héréditaires ou aux modifications apportées par l'environnement. Il est utile de s'y intéresser dans notre étude car des traits phénotypiques peuvent être liés aux lignées ou a des métabolites.

7.1 Lien entre la lignée et la mortalité

On s'intéresse ici à la mortalité des porcelets. On veut savoir s'il y a un lien entre la lignée d'un porcelet et sa survie. Pour cela, on fait un test de Fisher. Il permet de tester l'indépendance entre deux variables qualitatives à partir d'une table de contingence, qui donne le nombre d'individus dans chacune des modalités.
La table 3 va permettre de tester les critères G10-/G10+ et vivant/mort. On teste l'hypothèse nulle H_0 « la lignée et la survie sont indépendantes » contre l'hypothèse alternative H_1 « la lignée et la survie ne sont pas indépendantes ».

	Vivant	**Mort**
G10–	111	15
G10+	105	17

TABLE 3 – Table de contingence entre les lignées et la survie

La p-valeur du test est de 0.7. On en conclut que la survie des porcelets ne dépend pas de leur lignée.

7.2 Poursuite de l'étude

Il serait intéressant de poursuivre l'étude des phénotypes, en s'intéressant aux autres variables phénotypiques disponibles : poids, longueur, largeur, circonférence, température.
On pourra réaliser une régression PLS afin de mettre en lien les métabolites et les phénotypes et essayer de prédire les phénotypes à partir des métabolites. Par manque de temps, je n'ai pas pu terminer cette analyse.

Conclusion

En résumé, les échantillons de sang prélevés sur des porcelets à la naissance ont été analysés par résonance magnétique nucléaire. Les métabolites dans les spectres ont ensuite été identifiés et quantifiés par la méthode ASICS. La comparaison entre les quantifications ASICS et les quantifications par dosages a permis de comprendre que plus la concentration d'un métabolite est élevée, meilleure est sa quantification ASICS. On remarque aussi que certains métabolites ne sont pas quantifiés par ASICS car ils ne sont pas détectés lors de la RMN.
Ensuite, on a réalisé une analyse exploratoire sur les données des quantifications. Cela a permis de valider le fait que le sexe et la bande d'élevage ne créent pas de biais. On a également remarqué la différence entre les prélèvements dès la naissance et 24 heures après, qui correspond à l'évolution physiologique du porcelet.
On a étudié les métabolites avec des tests de Kruskal-Wallis. La similarité des métabolites significatifs dans les tests pour les prélèvements de plasma et de sérum a montré que les deux types d'échantillons sont semblables. On peut donc prélever l'un ou l'autre sans affecter les résultats.
Grâce aux tests et à des modèles mixtes, on a identifié des métabolites d'intérêt, comme le myo-inositol et la glutamine. La quantité de ces métabolites change entre la naissance et 24 heures pour l'une des lignées et cela pourrait avoir un lien avec la survie du porcelet. Le nombre de métabolites significatifs pour le modèle mixte incluant la date de prélèvement montre qu'il y a des différences importantes entre les deux dates.
Enfin, on a débuté l'étude des phénotypes et mis en évidence le fait que la survie ne dépend pas de la lignée de l'animal.

Pour la poursuite de ce travail, on pourra approfondir l'étude des phénotypes, en étudiant d'autres variables et en les mettant en lien avec les métabolites. On pourra également travailler sur les voies métaboliques [1] car ici on ne s'est intéressé qu'à la quantité des métabolites.

D'un point de vue personnel, ce stage a été une réelle opportunité pour moi de découvrir les biostatistiques. J'ai trouvé très intéressant d'échanger avec les biologistes car cela permet de comprendre leurs problématiques. Les enjeux de ce projet étaient très concrets, et les analyses ont permis de tirer des conclusions pratiques pour les biologistes. J'ai ainsi pu mesurer l'apport que peuvent avoir les statistiques dans d'autres disciplines. Ce stage m'a également permis d'acquérir quelques notions de biologie et de me familiariser avec le domaine de la recherche, que j'ai beaucoup apprécié. Je souhaite continuer dans ce sens car les perspectives en biologie sont riches et variées.

J'ai également eu l'opportunité au cours de mon stage de participer, en tant que volontaire, à la conférence UseR ! 2019, organisée à Toulouse. Cela a été l'occasion pour moi de voir les multiples possibilités qu'offre le logiciel R dans des domaines très divers, et de découvrir des outils spécifiques à la biostatistique et la bioinformatique. Cela a été une expérience enrichissante que j'ai beaucoup appréciée.

1. Succession de réactions chimiques au cours desquelles un substrat initial est transformé et donne naissance à des produits finis, via une série de métabolites intermédiaires. (source : https://www.aquaportail.com/definition-6064-voie-metabolique.html)

Références

[1] Dan Tulpan, Serge Léger, Luc Belliveau, Adrian Culf, and Miroslava Čuperlović-Culf. MetaboHunter : an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. *BMC bioinformatics*, 12(1) :400, 2011.

[2] Arianna Filntisi, Charalambos Fotakis, Pantelis Asvestas, George K Matsopoulos, Panagiotis Zoumpoulakis, and Dionisis Cavouras. Automated metabolite identification from biological fluid 1H NMR spectra. *Metabolomics*, 13(12) :146, 2017.

[3] Aalim M Weljie, Jack Newton, Pascal Mercier, Erin Carlson, and Carolyn M Slupsky. Targeted profiling : quantitative analysis of 1H NMR metabolomics data. *Analytical chemistry*, 78(13) :4430–4442, 2006.

[4] Jie Hao, William Astle, Maria De Iorio, and Timothy MD Ebbels. BATMAN - an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics*, 28(15) :2088–2090, 2012.

[5] Siamak Ravanbakhsh, Philip Liu, Trent C Bjordahl, Rupasri Mandal, Jason R Grant, Michael Wilson, Roman Eisner, Igor Sinelnikov, Xiaoyu Hu, Claudio Luchinat, et al. Accurate, fully-automated NMR spectral profiling for metabolomics. *PloS one*, 10(5) :e0124219, 2015.

[6] Daniel Cañueto, Josep Gómez, Reza M Salek, Xavier Correig, and Nicolau Cañellas. Rdolphin : A GUI R package for proficient automatic profiling of 1D 1H-NMR spectra of study datasets. *Metabolomics*, 14(3) :24, 2018.

[7] Gaelle Lefort, Laurence Liaubet, Cecile Canlet, Patrick Tardivel, Marie-Christine Pere, Helene Quesnel, Alain Paris, Nathalie Iannuccelli, Nathalie Vialaneix, and Remi Servien. ASICS : an R package for a whole analysis workflow of 1D 1H NMR spectra. *bioRxiv*, page 407924, 2018.

[8] Patrick JC Tardivel, Cécile Canlet, Gaëlle Lefort, Marie Tremblay-Franco, Laurent Debrauwer, Didier Concordet, and Rémi Servien. ASICS : An automatic method for identification and quantification of metabolites in complex 1D 1H NMR spectra. *Metabolomics*, 13(10) :109, 2017.

[9] Institut du Porc. https://www.ifip.asso.fr.

[10] Hélène Gilbert, Yvon Billon, Ludovic Brossard, Justine Faure, Philippe Gatellier, Florence Gondret, Etienne Labussière, Bénédicte Lebret, Louis Lefaucheur, N Le Floch, et al. Divergent selection for residual feed intake in the growing pig. *animal*, 11(9) :1427–1439, 2017.

[11] Mikko Hallman, Kristina Bry, Kalle Hoppu, Marjatta Lappi, and Maija Pohjavuori. Inositol supplementation in premature infants with respiratory distress syndrome. *New England Journal of Medicine*, 326(19) :1233–1239, 1992.

[12] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate : a pratical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1) :289–300, 1995.

[13] Guoyao Wu, Fuller W Bazer, Gregory A Johnson, Darrell A Knabe, Robert C Burghardt, Thomas E Spencer, XL Li, and JJ Wang. Triennial Growth Symposium : important roles for L-glutamine in swine nutrition and production. *Journal of Animal Science*, 89(7) :2017–2030, 2011.

[14] Mary J Lindstrom and Douglas M Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687, 1990.

[15] Émilie Lebarbier and Tristan Mary-Huard. Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la Société française de statistique*, 147(1) :39–57, 2006.

[16] Valentin Voillet, Magali SanCristobal, Yannick Lippi, Pascal GP Martin, Nathalie Iannuccelli, Christine Lascor, Florence Vignoles, Yvon Billon, Laurianne Canario, and Laurence Liaubet. Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity. *BMC genomics*, 15(1) :797, 2014.

Annexes

Annexe A : ACP sur les quantifications, avec une coloration par bande d'élevage

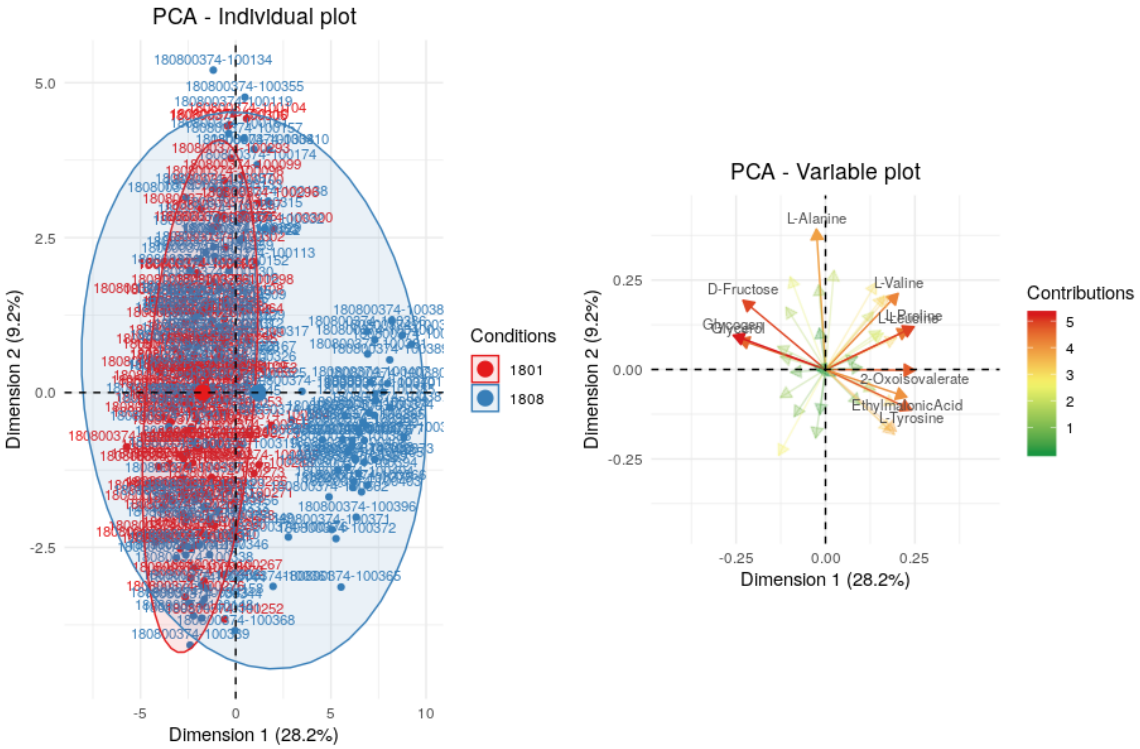


FIGURE 15 – ACP sur les quantifications de l'ensemble des individus, avec une coloration par bande d'élevage

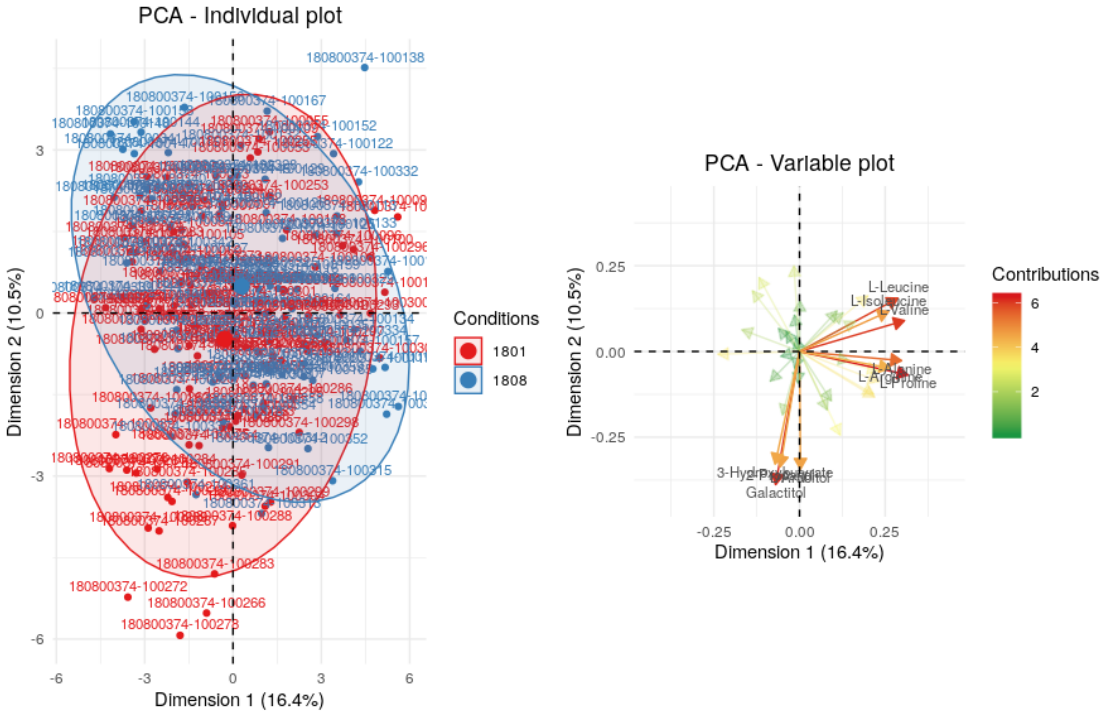
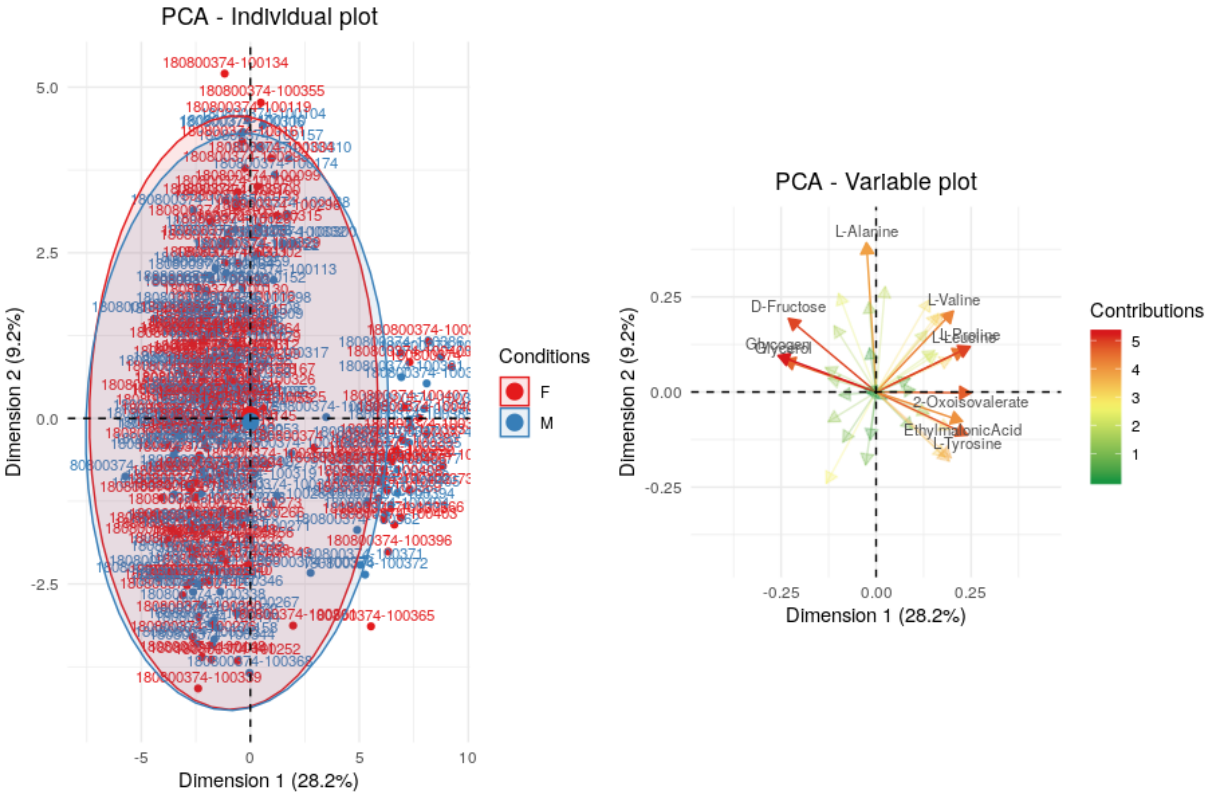
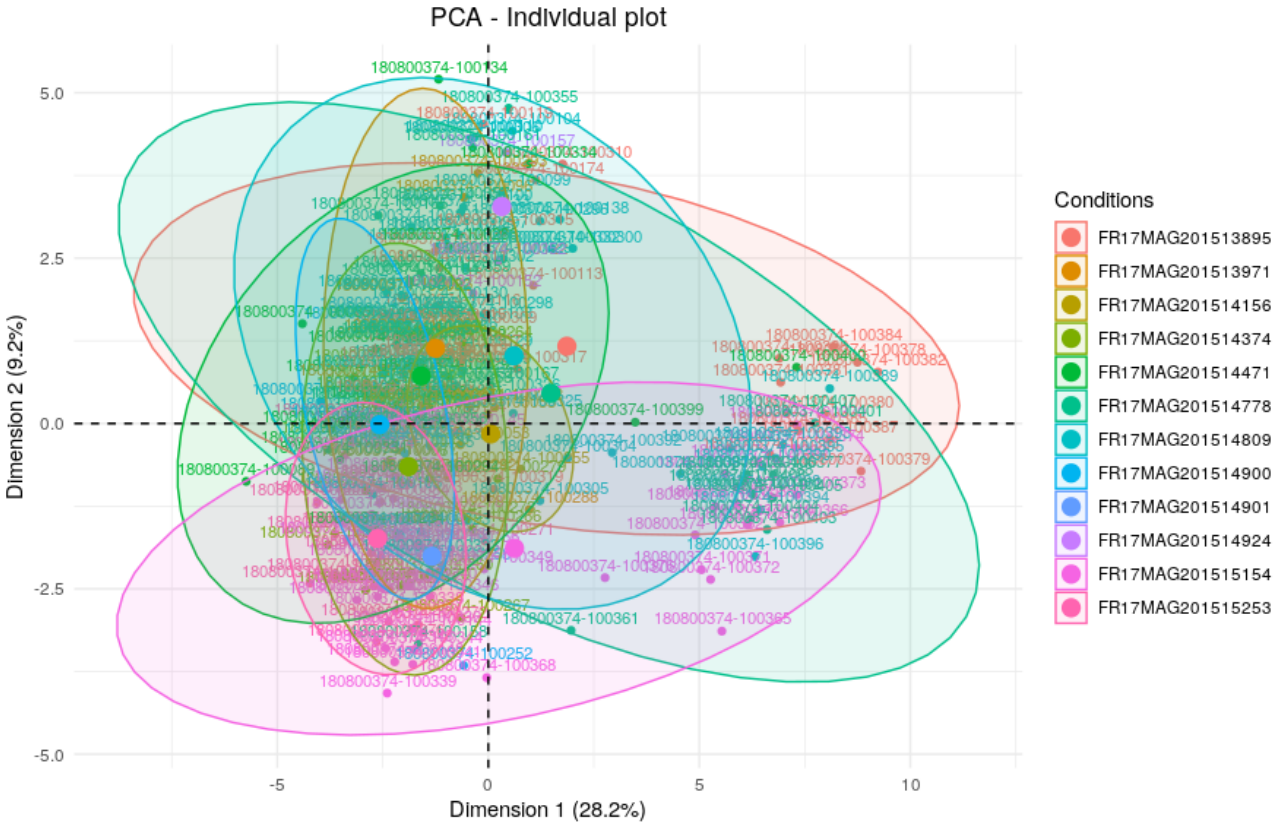


FIGURE 16 – ACP sur les quantifications des individus prélevés à la naissance, avec une coloration par bande d'élevage

Annexe B : ACP sur les quantifications de l'ensemble des individus, avec une coloration par sexe



Annexe C : ACP sur les quantifications de l'ensemble des individus, avec une coloration par mère



Annexe D : ACP sur les *buckets*

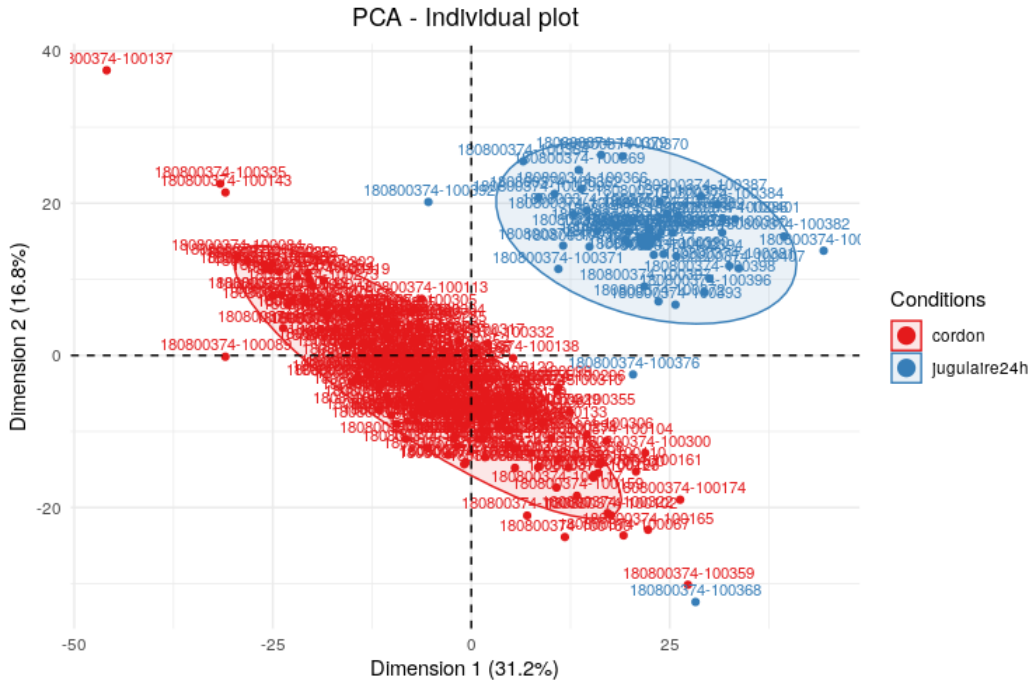


FIGURE 17 – ACP sur les *buckets* pour l'ensemble des individus, avec une coloration par type d'échantillon

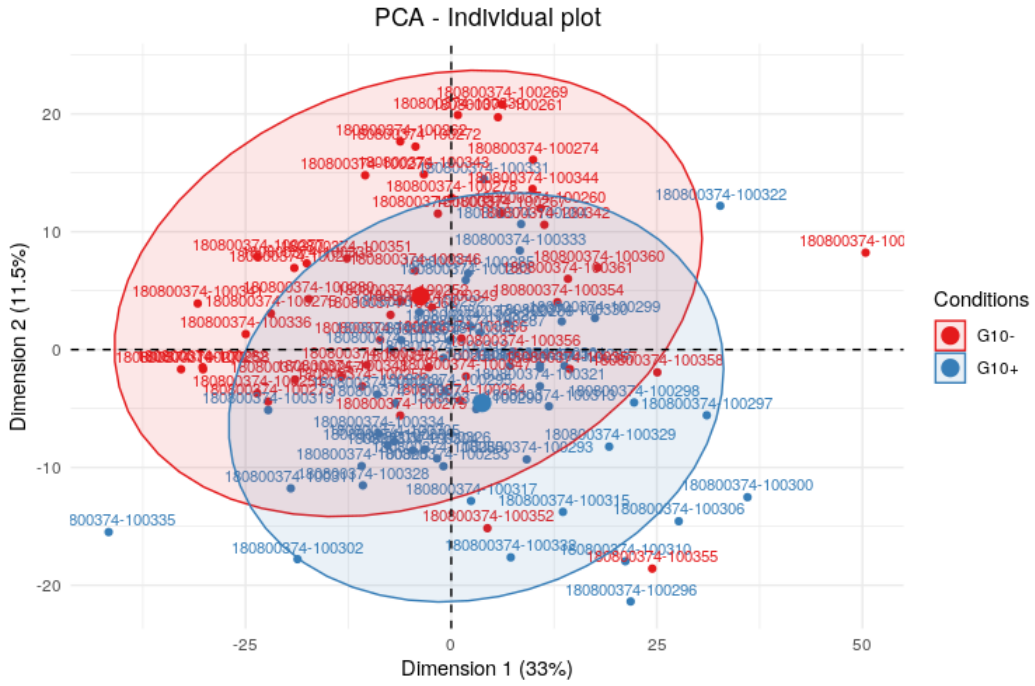


FIGURE 18 – ACP sur les *buckets* des individus prélevés de sérum à la naissance, avec une coloration par lignée

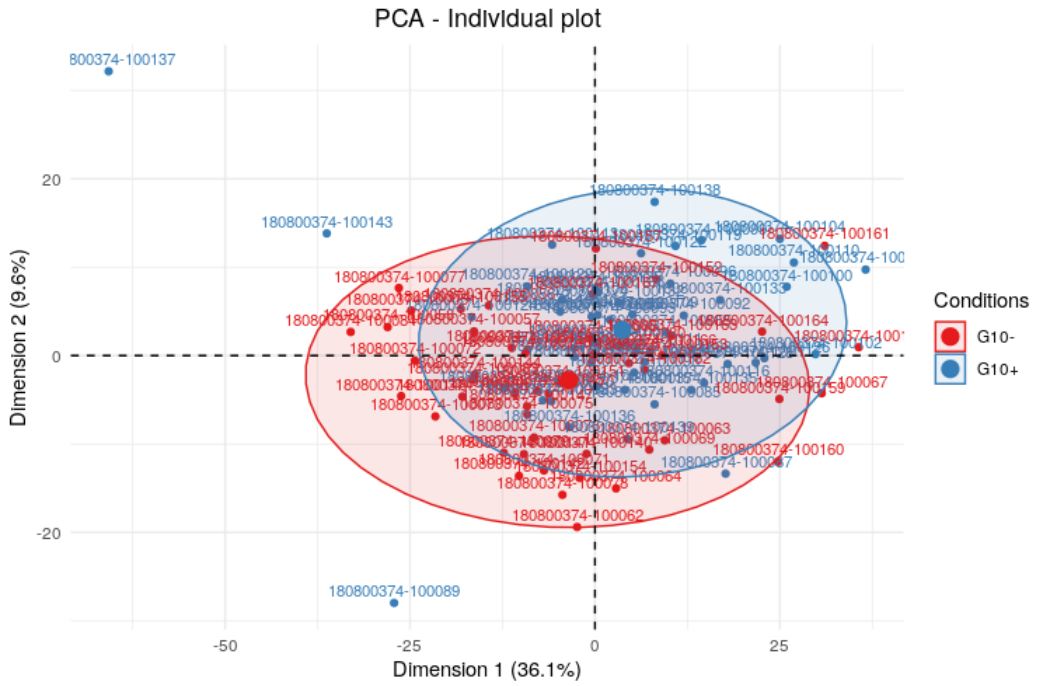
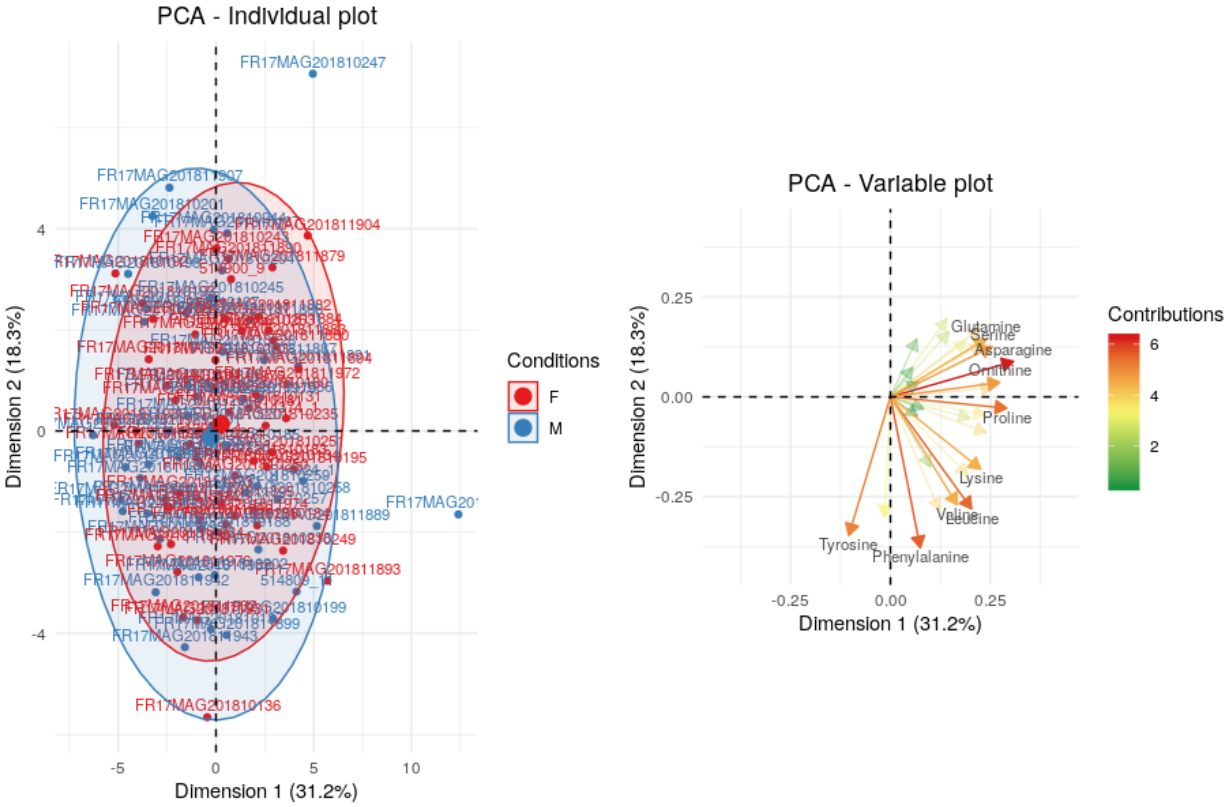


FIGURE 19 – ACP sur les *buckets* des individus prélevés de plasma à la naissance, avec une coloration par lignée

Annexe E : ACP sur les quantifications par dosages, avec une coloration par sexe



Résumé

Ce rapport présente le travail que j'ai réalisé au cours de mon stage de fin d'études à l'INRA, au sein de l'unité Mathématiques et Informatique Appliqués de Toulouse. L'objectif est d'analyser les données métabolomiques du projet SubPig.
La surmortalité périnatale des porcelets est un problème économique et éthique, lié au bien être de l'animal. On va donc étudier le profil métabolomique des porcelets à la naissance. On s'intéressera particulièrement aux différences entre deux lignées divergentes pour un critère d'efficacité alimentaire, qui jouerait un rôle dans la survie de l'animal.
Les données métabolomiques sont obtenues par Résonance Magnétique Nucléaire, sous forme de spectres. On utilise la méthode ASICS pour identifier et quantifier les métabolites présents dans les spectres.
On réalise ensuite une analyse exploratoire à partir des données de quantification. L'Analyse en Composantes Principales permet notamment de mettre en évidence l'importance de la date de prélèvement, qui correspond à l'évolution physiologique du porcelet.
Ensuite, des tests d'hypothèses et la construction de modèles mixtes permettent d'étudier les métabolites. On identifie des métabolites d'intérêt, comme le myo-inositol ou la glutamine.
Enfin, une première étude des phénotypes montre qu'il n'y a pas de relation entre la lignée et la survie des porcelets. On pourrait par la suite étudier de manière plus approfondie les données phénotypiques, afin de les relier avec les métabolites.

Mots clés : Métabolomique, ASICS, ACP, tests d'hypothèses, modèles mixtes

Abstract

This report presents the work I did during my final internship at INRA, in the Mathematics and Applied Informatics unit of Toulouse. The objective is to analyze the metabolomic data of the SubPig project.
Perinatal over-mortality of piglets is an economic and ethical problem, linked to the animal's well-being. We will therefore study the metabolomic profile of piglets at birth. Particular attention will be paid to the differences between two divergent lines for a feed efficiency criterion, which would play a role in the animal's survival.
Metabolomic data are obtained by Nuclear Magnetic Resonance, in the form of spectra. The ASICS method is used to identify and quantify the metabolites present in the spectra.
An exploratory analysis is then carried out based on the quantification data. In particular, the Principal Composant Analysis makes it possible to highlight the importance of the sampling date, which corresponds to the physiological evolution of the piglet.
Then, hypothesis tests and the construction of mixed models allow the metabolites to be studied. Metabolites of interest, such as myo-inositol or glutamine, are identified.
Finally, a first study of phenotypes shows that there is no relationship between the lineage and piglet survival. Phenotypic data could then be further investigated to link them with metabolites.

Keywords : Metabolomics, ASICS, PCA, hypothesis testing, mixed-effects models