Quantification automatique de métabolites dans un spectre RMN et application à la description de la maturité périnatale chez le porc

Gaëlle Lefort

To cite this version:

Gaëlle Lefort. Quantification automatique de métabolites dans un spectre RMN et application à la description de la maturité périnatale chez le porc. Statistiques [math.ST]. Université Toulouse Capitole, 2021. Français. NNT : . tel-03647188

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 1 Capitole

Présentée et soutenue par

Gaëlle LEFORT

Le 2 juillet 2021

Quantification automatique de métabolites dans un spectre RMN et application à la description de la maturité périnatale chez le porc

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Mathématiques et Applications**

Unité de recherche :
MIAT-INRA : Unité de Mathématiques et Informatique Appliquées Toulouse

Thèse dirigée par
Nathalie VIALANEIX et Rémi SERVIEN

Jury

Mme Kim-Anh Lê Cao, Rapporteure
M. Patrick Giraudeau, Rapporteur
Mme Anne Ruiz-Gazen, Examinatrice
M. Jaap van Milgen, Examinateur
M. Etienne Thévenot, Examinateur
Mme Nathalie Vialaneix, Directrice de thèse
M. Rémi Servien, Co-directeur de thèse
Mme Laurence Liaubet, Co-directrice de thèse

Résumé

Parmi les nombreuses données *omiques* qui décrivent le fonctionnement biologique d'un organisme, le métabolome suscite un intérêt croissant car il est plus proche des phénotypes d'intérêt et qu'il a donc avoir un potentiel important pour la recherche de *biomarqueurs*. La spectrométrie par résonance magnétique nucléaire (RMN) est une technologie haut-débit qui produit des spectres caractéristiques du mélange complexe de métabolites présents dans un échantillon d'intérêt. Cependant, leur interprétation biologique est difficile car ceux-ci ne donnent pas une mesure explicite des différentes quantités de métabolites présents dans l'échantillon.

Une approche prometteuse pour l'analyse de ces données consiste à identifier et quantifier les métabolites présents dans le mélange complexe à partir de son spectre et à réaliser l'analyse statistique sur les résultats de cette quantification. Une première partie de cette thèse a consisté en l'amélioration d'une méthode de quantification existante, ASICS, ainsi qu'à son implémentation dans un package R/Bioconductor. Une nouvelle méthode, prenant en compte l'ensemble des spectres d'une expérience lors de la quantification, a aussi été proposée dans le but d'améliorer la fiabilité des résultats.

Un second volet de cette thèse concerne l'application de cette méthode au problème de mortalité néonatale des porcelets et plus précisément à la description des mécanismes impliqués dans la mise en place de la maturité. L'analyse des spectres RMN de plasma, d'urine et de liquide amniotique de fœtus en fin de gestation a permis d'identifier des voies métaboliques impliquant de nombreux acides aminés et sucres (croissance et apport d'énergie) ainsi que le métabolisme du glutathion (stress oxydatif).

Abstract

Among all the *omics* data which characterize the biological functioning of an organism, metabolomics is a promising approach in systems biology for phenotype characterization or biomarker discovery. ^{1}H Nuclear Magnetic Resonance (NMR) is a high-throughput technology that allows to obtain characteristic spectra of a complex mixture of metabolites present in a sample of interest. However, the analysis of ^{1}H NMR spectra remains difficult, mainly due to the fact that the measure of different amounts of metabolites present in the sample is indirect with this technique.

A promising approach for the analysis of such data is to identify and quantify metabolites present in the complex mixture from its spectrum and to perform the statistical analysis on the results of this quantification. A first part of this thesis consists in improving an existing method of metabolite quantification, named ASICS, and in its implementation in an R/Bioconductor package. A new method that integrates information obtained from several complex spectra of the same experiment during the quantification process is also proposed in order to improve the reliability of the results.

A second part of this thesis concerns the application of this method to a problem of neonatal mortality in piglets and more precisely the description of the mechanisms involved in the establishment of maturity. Analysis of NMR spectra of plasma, urine and amniotic fluid from late gestation fetuses allow to identify metabolic pathways involving numerous amino acids and sugars (growth and energy supply) as well as glutathione metabolism (oxidative stress).

Remerciements

Je tiens tout d'abord à remercier mes directrices de thèses : Nathalie Vialaneix, Rémi Servien, Laurence Liaubet et Hélène Quesnel. Nathalie, je te remercie de la confiance que tu m'as apportée et de m'avoir supportée durant toutes ces années, travailler avec toi est un plaisir et j'ai énormément appris à tes côtés. Rémi, merci de ton soutien, tes conseils, ta patience et ta disponibilité pour répondre à mes questions ou venir régulièrement à Castanet (même en vélo par 35°C). Laurence, tu as toujours été là, pendant et avant la thèse, pour discuter de biologie (j'espère que je dis moins de bêtises aujourd'hui) ou quand j'en avais besoin et je t'en suis extrêmement reconnaissante. Hélène, même de plus loin, merci d'avoir été là lorsque j'ai eu besoin de tes connaissances et tes conseils.

Je remercie également l'institut de convergence #DigitAg et l'INRAE plus spécialement les départements MIA, GA et SA pour le financement de cette thèse.

Je remercie par ailleurs Kim-Anh Lê Cao et Patrick Giraudeau pour avoir accepté de rapporter cette thèse et pour vos remarques pertinentes. Un grand merci également à Anne Ruiz-Gazen, Étienne Thévenot et Jaap van Milgen pour avoir accepté de faire partie de mon jury. Je souhaite aussi remercier Cécile Canlet, Marie-José Mercat et Alain Paris, membres de mon comité de thèse, pour m'avoir conseillé durant ces trois années.

Merci aussi à tous les membres des unités MIAT et GenPhySE de m'avoir accueilli durant ces trois années (et les précédentes). Tout particulièrement, merci à notre serviteur Sylvain, de m'avoir toujours soutenue pour la réalisation de cette thèse. Merci à nos formidables gestionnaires Fabienne, Nathalie, Alain, Papa et Benjamine, ça a été un bonheur de vous avoir pour tout gérer si efficacement.

Merci à toute l'équipe GenoRobust, Agnès, Mireille, Cécile pour votre soutien et vos remarques toujours pertinentes, Laure et Lisa pour toutes nos discussions et ce que vous m'avez appris sur les manips. Un grand merci à Nathalie pour être toujours là quand j'ai des questions sur la métabo et pour ces bons moments au Japon (entre autres).

Merci aussi à toute la team cantine, Marie-Stéphane, Céline, Jérôme, Sarah, Patrice, Cédric, Philippe, nos discussions autour du repas m'ont manqué durant cette dernière année mais j'espère qu'on pourra se retrouver très bientôt. Merci aussi d'avoir toujours goûté toutes sortes de gâteau que j'ai pu faire. Marie-Stéphane, je te remercie pour tout ce que tu m'as appris que ce soit lié au travail ou non, nos nombreux échanges vont me manquer (surtout au petit-dej quand les couloirs sont déserts). Je n'oublie pas Margot, Camille et Joanna, ma thèse n'aurait pas été la même sans nos (trop ?) longues discussions au café ou ailleurs.

Merci à tous les autres membres de la plateforme BioInfo, Christine, je te remercie pour cette année pré-thèse. Merci Claire pour tes précieux conseils et Didier de m'avoir accueilli dans ton bureau.

Merci à ma collègue, voisine mais surtout amie, Claire, pour tes super dessins de cochons qui ont embelli cette thèse, ton soutien et d'avoir nourri la panthère au péril de ta vie lorsque j'étais en déplacement.

Merci Nathanaël de m'avoir supportée dans le bureau durant ces trois années, de m'avoir écoutée quand ça n'allait pas et surtout pour tous ces bons moments partagés.

Un énorme merci collectif à tous les doctorants et post-doc que j'ai croisés, Alyssa, Audrey, Emilie, Fulya, Harmony, Lise, Louise, Manon, Maxime, Océane, Quentin, Tiphaine, Vanille (et tous ceux que j'ai pu oublier...) pour ces moments au bureau et à l'extérieur.

Et enfin, je ne peux finir sans remercier ma famille, Maman, Papa, merci de votre soutien durant toutes ces années, d'avoir toujours été là quand j'en avais besoin et de m'avoir encouragée. C'est grâce à vous si j'en suis là aujourd'hui. Merci à mon petit frère (seulement par l'âge pas pour tout le reste), ça a été un bonheur que tu viennes à Toulouse.

Contributions

Articles

- G. Lefort *et al.* (2019). ASICS : an R package for a whole analysis workflow of 1D 1H NMR spectra. *Bioinformatics*, 35(21) : 43564363
- G. Lefort *et al.* (2020). The maturity in fetal pigs using a multi-fluid metabolomic approach. *Scientific Reports*, **10**, 19912
- G. Lefort *et al.* (2021). Joint automatic metabolite identification and quantification of a set of ^{1}H NMR spectra. *Analytical Chemistry*, 93(5) : 28612870

Communications orales et poster dans des conférences

- G. Lefort *et al.* (2018). ASICS : un package R pour l'identification et la quantification de métabolites dans un spectre RMN 1H. *7es Rencontres R*. Présentation orale.
- G. Lefort *et al.* (2018). ASICS : a new R package for identification and quantification of metabolites in complex 1D 1H NMR spectra. *European Conference on Computational Biology (ECCB 2018)*. Poster.
- G. Lefort *et al.* (2018). ASICS : identifier et quantifier des métabolites à partir d'un spectre RMN ^{1}H. *51es Journées de Statistique de la SFdS (JdS2019)*. Présentation orale.
- G. Lefort *et al.* (2019). ASICS : identification and quantification of metabolites in complex 1D 1H NMR spectra. *15th Annual Conference of the Metabolomics Society (Metabolomics 2019)*. Poster.
- G. Lefort *et al.* (2019). Study of fetal pig maturity in relation with neonatal survival using a multi-fluids metabolomic approach. *15th Annual Conference of the Metabolomics Society (Metabolomics 2019)*. Poster.
- G. Lefort *et al.* (2019). ASICS : a new R package for identification and quantification of metabolites in complex 1H NMR spectra. *UseR ! 2019*. Poster.

- G. Lefort (2019). ASICS : identifier et quantifier des métabolites à partir d'un spectre RMN ^{1}H. *StatMathAppli 2019*. Présentation orale.
- G. Lefort *et al.* (2020). ASICS : identification and quantification of metabolites in complex ^{1}H NMR spectra. *European RFMF Metabomeeting 2020*. Présentation orale.
- G. Lefort *et al.* (2020). Étude de la maturité des porcelets en fin de gestation par une approche métabolomique multifluide. *52es journées de la recherche porcine*. Poster.
- G. Lefort *et al.* (2020). Joint automatic metabolite identification and quantification of a set of ^{1}H NMR spectra. *Metabolomics 2020*. Poster.

Développement

- G. Lefort *et al.* (2018). **ASICS** : Automatic Statistical Identification in Complex Spectra. R package version 2.0.1. http://bioconductor.org/packages/ASICS/

Table des matières

Chapitre 1

Introduction

1.1 Métabolomique par Résonance Magnétique Nucléaire (RMN)

La métabolomique en quelques mots

Les données dites *omiques* permettent de décrire le fonctionnement biologique d'un organisme à divers niveaux de l'échelle du vivant (Figure 1.1) : génome (séquence d'ADN), transcriptome (ensemble des ARN présents dans une cellule ou un tissu), protéome (ensemble des protéines), métabolome (ensemble des métabolites) (Wishart, 2007). Leur analyse et leur intégration est un enjeu important pour la biostatistique moderne qui doit permettre de mieux comprendre les mécanismes moléculaire sous-jacents (Rohart et al., 2017 ; Mariette and Villa-Vialaneix, 2018). Contrairement au transcriptome et au protéome, dont les fonctions dépendent respectivement de régulations épigénétiques et post-transcriptionnelles, le métabolome représente une image directe de l'activité biochimique d'un organisme (Patti et al., 2012). Par conséquent, les métabolites, qui sont le produit final du processus de régulation cellulaire, sont plus directement corrélés au phénotype d'intérêt observé situé à l'échelle de l'individu (Fiehn, 2002). Même si la métabolomique reste sous-utilisée par rapport aux autres *omiques*, son potentiel important en terme de biomarqueurs suscite un intérêt croissant et en fait un champ d'étude employé dans divers domaines (Giraudeau, 2017 ; Sévin et al., 2015 ; Patti et al., 2012). Par exemple, dans le domaine médical, l'analyse du métabolome est utilisée pour trouver de nouvelles caractéristiques biologiques (des biomarqueurs) facilement mesurables permettant, entre autres, de détecter une maladie (dépistage médical) ou d'adapter le traitement à des caractéristiques propres à l'individu (médecine personnalisée).

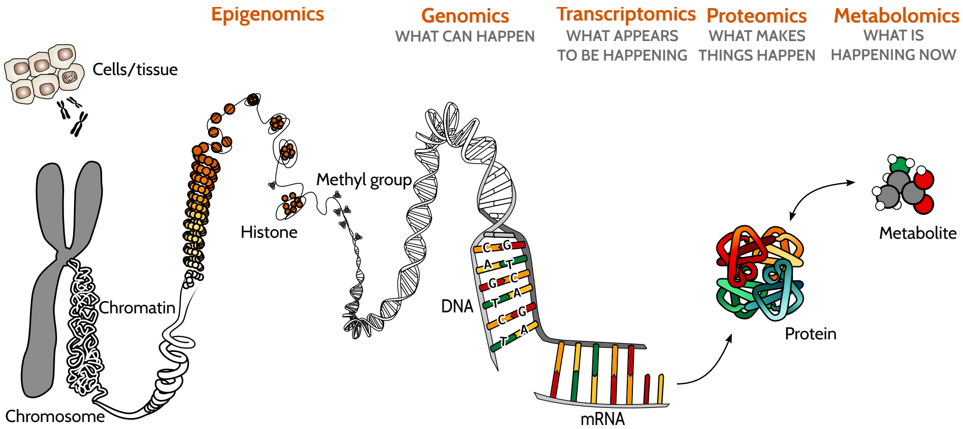


FIGURE 1.1 – Représentation schématique de différents niveaux de caractérisation des systèmes biologiques inspirée en partie de https://fr.wikipedia.org/wiki/Fichier:Chromosome_fr.svg.

Principe de la Résonance Magnétique Nucléaire (RMN)

Diverses approches permettent de caractériser le métabolome en mesurant simultanément un grand nombre de métabolites dans les fluides biologiques et les tissus. Parmi elles, on trouve en particulier la spectrométrie de masse (*Mass spectrometry*, MS) ou la résonance magnétique nucléaire (RMN). Ces deux techniques sont complémentaires (Emwas, 2015 ; Emwas et al., 2019 ; Dunn et al., 2011). Historiquement, le profilage du métabolome par RMN a été proposé par l'équipe de Nicholson (Nicholson et al., 1999). Cette technique a l'avantage d'être très reproductible, non destructive, moins sensible aux biais techniques et de couvrir un grand nombre de métabolites présents dans les différents réseaux métaboliques. Au contraire, la spectrométrie de masse est plus sensible mais demande plus de temps pour préparer l'échantillon et plus de temps d'acquisition des spectres. En outre, elle doit être couplée à de la chromatographie liquide (*liquid chromatography*, LC) ou gazeuse (*gaz chromatography*, GC) pour permettre la simplification du mélange. Dans le cadre d'analyses non ciblées, de recherches de biomarqueurs, et d'analyses conjointes avec d'autres types de *omiques*, la RMN reste la technique la plus utilisée.

Sur le plan technique, la RMN se base sur la possibilité de certains noyaux atomiques d'interagir avec un champ magnétique (Hore, 1995 ; Canet et al., 2002). Sans entrer dans les détails, la RMN consiste à observer les transitions entre deux niveaux d'énergie très proches d'un noyau, quel qu'il soit, soumis à un champ magnétique. Cette transition d'énergie induit un courant qui est mesuré en fonction du temps. La courbe qui en résulte, appelée FID (*Free Induction Decay*), correspond a une somme de sinusoïdes amorties de caractéristiques différentes dans le cas où le nombre de noyaux est important (Figure 1.2). Le noyau le plus couramment utilisé est celui de l'hydrogène : on parle alors de ^{1}H RMN ou RMN du proton (le noyau de l'hydrogène n'étant constitué que d'un seul proton).

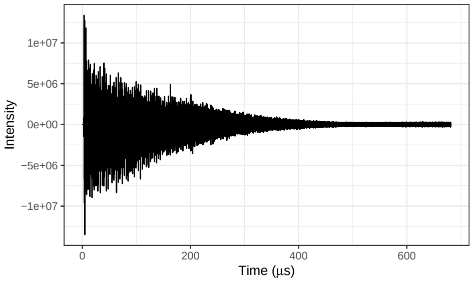


FIGURE 1.2 – FID obtenu à partir d'un mélange synthétique de 21 métabolites.

Des prétraitements sont alors nécessaires pour obtenir un spectre interprétable (Martin et al., 2018). Parmi ces prétraitements, la transformée de Fourier est utilisée pour passer du domaine temporel dans lequel se trouve le spectre FID au domaine fréquentiel (Figure 1.3). Cela induit en particulier une conversion de l'échelle de mesure en Hertz en déplacements chimiques mesurés en *parts-per-million* (ppm) ce qui permet de s'affranchir du champ magnétique et du spectromètre utilisés.

Le spectre ainsi obtenu permet de distinguer plus simplement les différents métabolites composant le mélange. En effet, la structure chimique de chaque métabolite et, plus précisement, de chaque groupe de protons qui le compose, va induire une position spécifique des pics (Figure 1.4). Les couplages entre les groupes voisins (*i.e.* les liaisons entre les groupes) vont jouer sur la multiplicité d'un groupe de pics (*i.e.*, le nombre de pics du groupe). Par exemple pour l'éthanol (Figure 1.4), il y a trois groupes de protons (de couleurs différentes), chacun ayant une multiplicité différente : le groupe OH est représenté par un seul pic (singulet) alors que le groupe H_3C est repré-

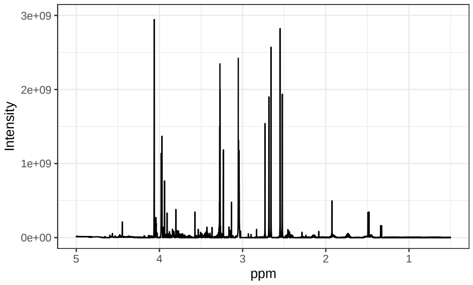


FIGURE 1.3 – Spectre obtenu après transformée de Fourier du spectre FID de la Figure 1.2.

senté par trois pics (triplet). La présence d'un métabolite dans un mélange complexe est alors déduite de la présence de tous ses pics aux positions théoriques avec les bonnes multiplicités dans le spectre obtenu à partir de ce mélange complexe.

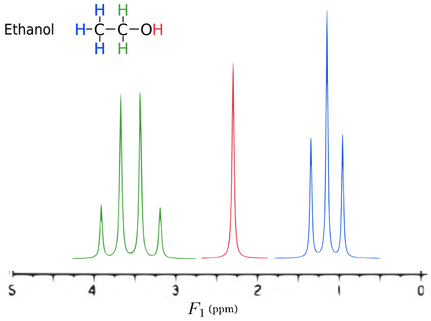


FIGURE 1.4 – Spectre de l'éthanol sur lequel les couleurs permettent d'identifier les pics correspondants à des groupes de protons donnés.

La concentration de chaque métabolite dans le mélange est, elle, proportionnelle à l'aire sous chacun des pics du métabolite (Tofts and Wray, 1988). De ce fait, un métabolite avec peu de pics mais avec une intensité élevée peut être moins concentré qu'un métabolite avec beaucoup de pics d'intensités moindres.

Difficultés d'utilisation des spectres RMN

Le nombre important de métabolites dans un échantillon biologique induit un nombre important de pics dans le spectre RMN (Figure 1.3). Les pics des différents métabolites vont se chevaucher et rendre complexe l'identification des métabolites présents. Pour identifier les métabolites, il faut regarder l'ensemble du spectre pour voir si tous les pics d'un métabolite sont présents. Ceci est long, fastidieux, nécessite des connaissances approfondies en RMN et est donc dépendant de la personne réalisant l'identification (Tredwell et al., 2011). Pour améliorer la résolution du spectre et limiter le chevauchement, il est possible d'augmenter le champ magnétique. Néanmoins, même dans ce cas, beaucoup de pics continuent à se chevaucher.

Une difficulté supplémentaire de l'identification est que les déplacements chimiques peuvent être décalés d'un spectre à l'autre à cause des variations expérimentales lors de l'acquisition des spectres, et donc parfois chevaucher les pics d'un autre métabolite dans un spectre différent (voir l'exemple de spectres de l'alanine dans la Figure 1.5, obtenus dans une même expérience). En effet, les déplacements chimiques sont sensibles à certaines conditions expérimentales comme la température ou le pH de la solution (Weljie et al., 2006).

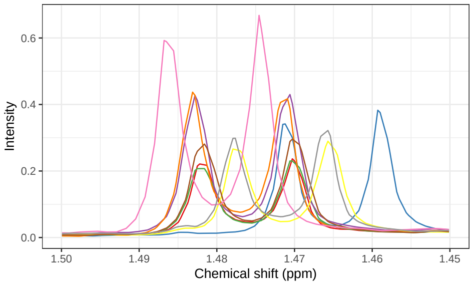


FIGURE 1.5 – Doublet de l'alanine pour des spectres venant de l'expérience PORCINET décrites dans la Section 1.3.

Objectifs de la thèse

Malgré leur utilisation fréquente, l'interprétation biologique des spectres RMN en une dimension reste compliquée. L'objectif de cette thèse est le développement d'une méthode permettant une identification et une quantification automatiques des métabolites dans ce type de spectre. Il s'agira, dans un premier temps, d'améliorer une méthode existante, **ASICS**, en la calibrant par l'utilisation de données réelles (Section 1.2). Ensuite, pour améliorer la précision des quantifications, une méthode jointe prenant en compte l'ensemble des spectres d'une même expérience est présentée (Section 1.2.4). Ces différentes méthodes seront appliquées à la description de la maturité périnatale chez le porc (Section 1.3).

1.2 Traitement de spectres RMN en une dimension

Avant de pouvoir réaliser les analyses statistiques qui nous intéressent, des traitements des spectres sont nécessaires. En plus des prétraitements utiles pour transformer le spectre FID, l'alignement des spectres complexes et leur normalisation permettent de s'affranchir des conditions expérimentales. Dans la Section 1.2.1, les différentes méthodes d'alignement seront détaillés car aucune n'est encore acceptée comme méthode de référence au contraire des prétraitements antérieurs qui sont maintenant réalisés en routine. Ensuite, le spectre est découpé en *buckets* ou les métabolites sont quantifiés pour obtenir de nouvelles variables utilisables dans les analyses futures (Section 1.2.2 et 1.2.3).

Quelques notations utiles

Deux types de spectres seront utilisés dans la suite :

- les spectres dits *complexes* représentent les spectres obtenus à partir d'un échantillon biologique. Ces spectres comportent plusieurs pics appartenant à plusieurs métabolites (Figure 1.3). Ils sont notés $g_j(t)$ où $j \in \{1, \ldots, n\}$ correspond au numéro de l'expérience et $t \in \{1, \ldots, q\}$ au point de la grille des déplacements chimiques (ppm) où g_j est observé. Lorsqu'un seul spectre complexe est analysé, il est noté g ;

- les spectres dits *purs* correspondent aux spectres d'un métabolite seul comme par exemple le spectre de l'éthanol (Figure 1.4). Ces spectres purs sont notés $f_i(t)$ où $i \in \{1, \ldots, p\}$ représente le métabolite. Lorsque l'on considérera la totalité des p spectres purs, on parlera de librairie des spectres purs.

Enfin, la contribution du métabolite i au spectre g_j sera notée β_{ji} (ou β_i si il n'y a qu'un spectre) et la quantification associée sera $b_{ji} = \beta_{ji}/u_i$ où u_i est le nombre de protons du métabolite i.

1.2.1 Prétraitement des spectres

L'utilisation de la transformée de Fourier n'est pas suffisante pour analyser des spectres RNM. Différentes autres étapes de prétraitements sont nécessaires (Martin et al., 2018 ; Alonso et al., 2015 ; Sumner et al., 2007). La plupart de ces pré-traitements sont usuels et réalisés en routine par des logiciels comme TopSpin (V2.1, Bruker, Biospin, Munich, Germany) ou le package R **PepsNMR** (Martin et al., 2018). Avant de réaliser la transformée de Fourier, une amélioration de la résolution du spectre est effectuée en ajoutant des zéros au signal FID (*zero-filling*). Puis, un filtrage (apodisation) est utilisé pour améliorer le rapport signal sur bruit. Le spectre obtenu par transformée de Fourier est alors soumis à un rephasage, permettant de rendre les pics symétriques, à un calibrage à l'aide d'une référence interne ajoutée avant l'acquisition et à une correction de la ligne de base. Ensuite, une étape d'alignement des pics est nécessaire pour s'affranchir des différences de conditions expérimentales. Mes travaux y ayant plus particulièrement trait les méthodes existantes sont détaillées dans le paragraphe suivant. Enfin, il est nécessaire de réaliser une normalisation des spectres pour rendre minimiser les variations dues aux différences de dilution des échantillons. L'une des méthodes les plus utilisées est la normalisation par l'aire sous la courbe, ou *constant sum normalization*, qui produit des spectres avec une intensité spectrale totale équivalente (Craig et al., 2006). D'autres types de normalisations comme la normalisation PQN (*Probabilistic Quotient Normalisation*, Dieterle et al. (2006)) ou par la médiane ou le premier quartile existent (Smolinska et al., 2012).

Alignement des pics

Comme évoqué précédemment, les différences de conditions expérimentales lors de lacquisition des spectres (variations de température ou de pH par exemple) peuvent entraîner des décalages plus ou moins importants de l'emplacement des pics sur le spectre (Figure 1.5). Avant toute analyse, il est nécessaire d'aligner chacun des pics de tous les spectres complexes pour les rendre comparables.

Il existe deux types de méthodes d'alignement de spectres RNM (Alonso et al., 2015) : les méthodes basées sur la déformation (*warping*) et celles basées sur la segmentation (*segmenting*). Les méthodes de déformation consistent en la transformation non linéaire de l'axe des déplacements chimiques pour maximiser la corrélation entre les spectres (Tomasi et al., 2004). Cela permet d'étirer ou de contracter localement le spectre. Deux algorithmes appelés *Correlation Optimized Cwarping* (COW) et *Dynamic Time Warping* (DTW) sont particulièrement utilisés. Ce type de méthode a l'avantage de bien fonctionner sur des zones densément peuplées en pics cependant le temps de calcul nécessaire à l'alignement est très long (Giskeødegård et al., 2010 ; Savorani et al., 2010). Les méthodes de segmentation consistent en l'application d'un décalage constant à tout le spectre ou à des intervalles. Parmi ces méthodes, icoshift (Savorani et al., 2010) et speaq (Vu et al., 2011 ; Beirnaert et al., 2018) sont les plus utilisées et se distinguent par leur algorithme de définition des bornes de ces intervalles. Une technique automatique de définition est disponible dans icoshift mais, pour obtenir de meilleurs résultats, il est recommandé de définir manuellement les bornes des intervalles. Dans la méthode du package **speaq**, les auteurs ont choisi d'implémenter une classification hiérarchique des pics qui est aussi très utilisé pour l'alignement de spectres en spectrométrie de masse (De Souza et al., 2006 ; Kazmi et al., 2006). Une telle classification autorise des décalages de pics en sens opposés ce qui en fait une meilleure méthode que celles précédemment citées (Vu et al., 2011). Plus précisément, la partition des intervalles, réalisée de manière itérative, permet d'aligner les spectres de plus en plus finement (Figure 1.6). Lors de la première étape, l'alignement est réalisé sur l'ensemble du spectre. Ensuite le spectre est divisé en deux grâce à la classification hiérarchique et l'alignement est réalisé sur chacun des sous-spectres. Ces deux étapes (classification et alignement) sont ensuite réalisées jusqu'à ce que le domaine traité ne comporte plus qu'un seul pic.

Pour améliorer le temps de calcul des méthodes de *warping*, la corrélation croisée de la transformée de Fourier (*Fast Fourier Transform cross-correlation*) est utilisée (Wong et al., 2005) pour aligner les spectres dans icoshift (Savorani et al., 2010) et speaq (Vu et al., 2011 ; Beirnaert et al., 2018). Pour utiliser cette corrélation, un spectre de référence g_1 est choisi sur lequel sont alignés les autres spectres. Plusieurs décalages s sont ensuite testés dans l'intervalle $[-m_1, m_1]$ où m_1 est le décalage maximal autorisé fixé par l'utilisateur en fonction des conditions expérimentales. Pour chacun de

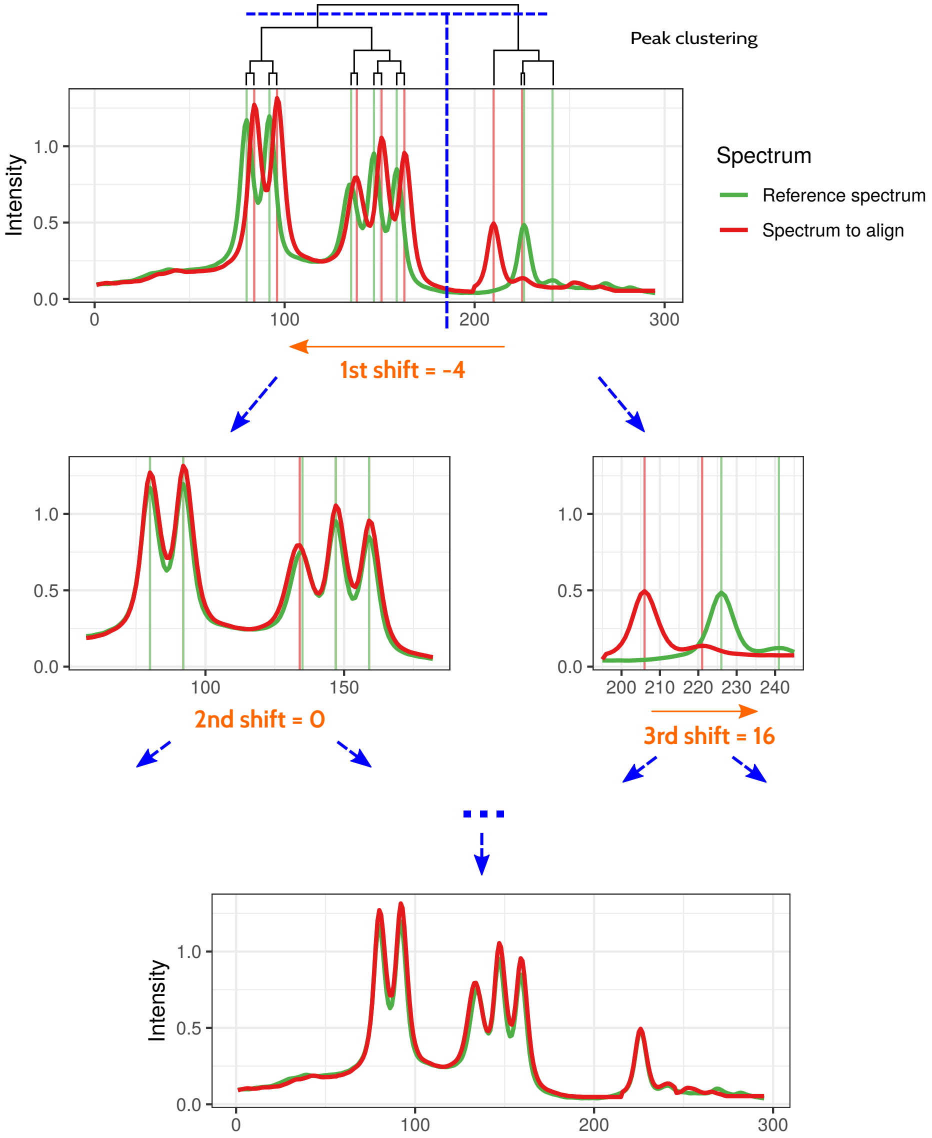


FIGURE 1.6 – Algorithme incrémental d'alignement implémenté dans le package **speaq**.

ces décalages, les FFT des spectres g_1 et $\tilde{g}_{2,s} : t \to g_2(t+s)$ sont calculées :

$$\forall\, l = 0,\dots,L-1, \qquad g_{jl}^{\text{FFT}} = \sum_{l'=0}^{L-1} g_j(t_{l'})e^{-\frac{2j\pi}{L}\times ll'} \tag{1.1}$$

avec $t_1,\dots,t_L$ les points (ppm) auxquels g_j est observé et g_j est soit g_1 soit $\tilde{g}_{2,s}$. Le meilleur décalage est alors choisi comme celui pour lequel la corrélation croisée est maximisée :

$$s^* = \underset{s\in\{s_1,\dots,s_K\}\subset[-M,M]}{\arg\max} \operatorname{Cor}(g_1^{\text{FFT}}, \tilde{g}_{2,s}^{\text{FFT}}) \tag{1.2}$$

Cette corrélation est utilisée dans icoshift et speaq pour calculer le décalage optimal pour chacun des intervalles.

1.2.2 Analyse des spectres : *buckets vs.* déconvolution

Deux approches sont possibles pour utiliser des spectres RMN (Alonso et al., 2015) dans des analyses statistiques. La plus simple et couramment utilisée est le *bucketing* aussi appelé *binning*. L'idée du *bucketing* est de découper le spectre en petits intervalles, appelé *buckets*, puis de calculer l'aire sous le spectre de chacun de ces *buckets* pour obtenir de nouvelles variables (Figure 1.7). Généralement la taille des *buckets* est fixée et égale pour tout le spectre. Il est aussi possible de choisir manuellement les bornes de chacun des *buckets*. Dans ce cas, un *bucket* correspond idéalement à un pic du spectre ou à un groupe de pics qu'on sait appartenir au même métabolite. Avec cette approche, la grande reproductibilité de la RMN permet d'obtenir des résultats de bonne qualité en terme de prédiction ou de différence entre conditions même si les différences entre les spectres sont faibles. Cependant, une étape d'identification des pics ou *buckets* pour leur associer un métabolite est ensuite nécessaire pour pouvoir interpréter ces résultats.

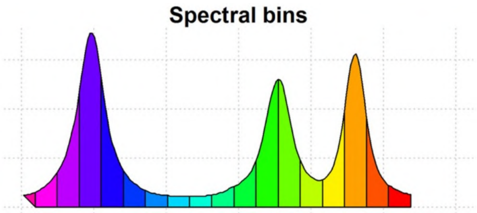


FIGURE 1.7 – Découpage d'un spectre en *buckets*.

Cette identification est généralement réalisée en comparant la position des pics dans les spectres de métabolites purs à celle dans les spectres des mélanges complexes. Plusieurs bases de données contenant des spectres purs sont disponibles : HMDB (*Human Metabolome Database*, Wishart et al. (2007)) ou BMRB (BioMagResBank, Ulrich et al. (2008)), par exemple. Il est aussi possible d'identifier automatiquement les métabolites avec des méthodes comme MetaboHunter (Tulpan et al., 2011) ou celle présentée dans Jacob et al. (2013). Pour la première, la liste des pics ou celle des *buckets* dintérêt est comparée automatiquement à la liste des pics des métabolites contenue dans MetaboHunter pour obtenir un score de présence par métabolite. Comme seul la position des pics est prise en compte et non les liens possible entre eux, cela entraîne un fort taux de faux positifs. Pour la seconde méthode, une classification est réalisé grâce aux corrélations entre les pics ou *buckets* puis l'identification est réalisé sur chacun des clusters ce qui permet d'améliorer l'identification car il est supposé que les pics d'un même cluster appartiennent au même métabolite (Alonso et al., 2015).

Alternativement à l'approche utilisant les *buckets*, il est possible de quantifier automatiquement les métabolites avant toute analyse. Plusieurs méthodes existent, toutes basées sur la déconvolution des mélanges complexes grâce à des spectres de métabolites purs (Figure 1.8) :

$$g(t) \simeq \sum_{i=1}^{p} \beta_i f_i(t) \tag{1.3}$$

où $g(t)$ correspond au mélange complexe, $f_i(t)$ au spectre pur du métabolite i ($i \in \{1 \dots p\}$), et β_i à la contribution dans le mélange du métabolite i. Une librairie de spectres purs est donc nécessaire et généralement alignée sur le spectre du mélange complexe g pour corriger les décalages existant entre les pics (comme vu dans la Section 1.2.1). Cette partie est essentielle pour une bonne quantification car les spectres de la librairie ne sont généralement pas acquis en même temps que les mélanges complexes, voire ils ont été obtenus avec un spectromètre différent. Les décalages spectres purs *vs.* mélanges complexes sont donc potentiellement très différents (d'amplitude plus importante) des décalages entre mélanges complexes.

Je décris, ci-dessous, quelques-unes des méthodes les plus couramment utilisées pour la quantification automatique. Chacune d'elles utilise un alignement ainsi qu'un méthode de quantification qui lui sont propres.

AutoFit (Mercier et al., 2011 ; Weljie et al., 2006) AutoFit est l'une des premières méthodes de quantification automatique qui a été développée. Elle est incluse dans le logiciel commercial Chenomx (Chenomx

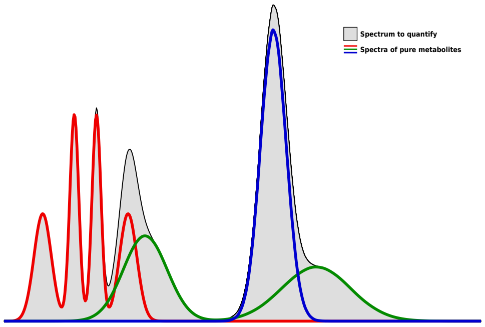


Figure 1.8 – Déconvolution d'un spectre à l'aide de spectres de métabolites purs.

Inc., Edmonton, AB, Canada) et bénéficie d'une librairie de métabolites purs très complète. Chacun des spectres purs a été acquis à partir de différents niveau de pH dans la solution (entre 4.0 et 9.0 avec des intervalles de 0.5 unité de pH). À l'aide de ces différents spectres, la librairie est calibrée pour estimer au mieux les concentrations. Pour estimer les concentrations, chaque spectre pur est découpé en intervalles correspondant aux groupes de pics, $k \in \{1, ..., z_i\}$ et le groupe de pics est résumé par une fonction (fonction lorentzienne) dépendant du spectre pur, $\tilde{f}_{ik}$. On ajuste alors le modèle :

$$g(t) = \sum_{i=1}^{p} \beta_i \sum_{k=1}^{z_i} \tilde{f}_{ik}(t + s_{ik}) + \epsilon \tag{1.4}$$

avec $s_{i,k}$ le déplacement chimique du groupe de pics k du métabolite i.

Les contributions des métabolites, β_i, ainsi que les décalages optimaux, s_{ik}, sont estimés en minimisant la fonction objectif $\|\epsilon\|^2$ grâce à des algorithmes de recuit simulé et des algorithmes génétiques.

BATMAN (Astle et al., 2012 ; Hao et al., 2012, 2014) BATMAN n'est pas fourni avec une librairie de référence et pour créer celle-ci, il faut donc utiliser des spectres purs obtenus par ailleurs ou ceux de bases de données telles que HMDB. Pour chaque métabolite, un pré-traitement des spectres purs est alors nécessaire qui consiste à extraire des déplacements chimiques, couplages et ratios d'intensité de chaque groupe de pics. Ces étapes demandent un travail supplémentaire par rapport à la simple utilisation de spectres. Cependant, une telle librairie d'environ 750 métabolites est disponible dans le package. De plus, l'alignement des spectres est réalisé de manière semi-automatique à l'aide d'un module Matlab indépendant grâce à une approche à base de fonctions splines. La quantification des métabolites est effectuée dans un cadre Bayésien grâce au modèle suivant :

$$g(t) = \sum_{i=1}^{p} \beta_i f_i(t) + \sum_{i>p} \beta_i f_i(t) + \epsilon \tag{1.5}$$

où $(f_i)_{i>p}$ correspond aux métabolites n'appartenant pas à la librairie.

Comme pour AutoFit, les spectres purs de la librairie ne sont pas utilisés directement mais au travers du même modèle que décrit dans l'équation (1.4). La partie inconnue, $\sum_{i>p} \beta_i f_i(t)$, est représentée sur une base d'ondelettes. Le modèle est, enfin, estimé grâce à un algorithme MCMC (*Markov chain Monte Carlo*) implémenté dans le package R **batman**. Ce type d'algorithme est très long donc le calcul des quantifications pour un spectre peut prendre plusieurs heures, voire des jours pour cette approche.

BAYESIL (Ravanbakhsh et al., 2015) Contrairement à BATMAN, BAYESIL inclut tous les prétraitement décrits dans la Section 1.2.1 depuis les spectres FID. Comme pour les méthodes précédentes, le spectre complexe est divisé en intervalles. Cependant les bornes des intervalles sont définies à partir du spectre complexe au lieu des spectres purs. Un modèle graphique probabiliste est ensuite utilisé pour attribuer chaque intervalle à un métabolite. Un modèle de mélange similaire au modèle d'AutoFit est enfin utilisé, dans un cadre bayésien, pour estimer les concentrations de chaque métabolites ainsi que les décalages grâce à des méthodes d'inférence de Monte Carlo. BAYESIL est disponible facilement à travers une interface Web. Cependant, les prétraitements des spectres étant réalisés avec cette méthode, il est nécessaire des respecter les conditions expérimentales spécifiées pour l'utiliser.

Dolphin (Gómez et al., 2014 ; Cañueto et al., 2018) Tout comme pour BATMAN, la librairie utilisée par la méthode Dolphin est créée à partie des déplacements chimiques, couplages et ratios d'intensité de chacun des pics des métabolites. Ensuite, toujours comme dans BATMAN, les spectres purs sont divisés en intervalles (ici appelé *region of interest*, RIO) en utilisant les groupes de pics comme dans l'équation (1.4). Une première étape d'identification des métabolites est réalisée en comparant les spectres des mélanges complexes et les spectres purs pour chaque intervalle. Les quantifications sont estimées région par région dans le but de maximiser un critère de qualité de l'ajustement. Pour améliorer les quantifications réalisées par intervalle, les ratios d'intensités disponibles dans la librairie sont utilisés dans l'estimation. Cette méthode est implémentée dans un package R nommé **rDolphin**. En plus de la quantification, des indicateurs de qualité et de fiabilité sont inclus. Cependant, comme les estimations sont calculées par régions, Dolphin produit plusieurs quantifications pour un même métabolite, ce qui complique l'utilisation des quantifications et leur interprétation.

ASICS (Tardivel et al., 2017) La librairie de spectres purs utilisée par cette méthode contient la totalité du spectre et non un résumé comme pour BATMAN et Dolphin. Contrairement aux méthodes précédemment détaillées, les pics ne sont pas modélisés grâce à des fonctions lorentziennes. ASICS permet de quantifier les métabolites dans un spectre en quatre étapes (Figure 1.9). Dans un premier temps, les spectres purs des métabolites pour lesquels tous les pics ne sont pas présents dans le spectres complexe sont écartés car il ne peuvent alors être présents dans ce mélange. Ensuite, les spectres purs restants sont alignés sur le spectre complexe en utilisant une fonction ϕ de déformation des déplacements chimiques. Les quantifications sont alors estimées par minimisation d'un critère de moindres carrés pour le modèle :

$$g(t) = \sum_{i=1}^{p} \beta_i \, f_i(\phi_i(t)) + \epsilon(t) \tag{1.6}$$

où ϕ_i correspond à la déformation appliquée à ce spectre et ϵ au bruit. Enfin, par rapport aux autres méthodes, une étape de sélection des métabolites présents, basée sur le contrôle du FWER (*FamilyWise Error Rate*), est mise en place. Cette méthode est disponible sur R avec une librairie de près de 200 spectres purs.

Tardivel et al. (2017) comparent ces différentes méthodes (sauf Dolphin qui n'était pas encore publiée) et montrent la supériorité d'ASICS en termes de quantifications et d'identification sur un mélange synthétique de 21 métabolites et sur des données réelles d'un échantillon de plasma. La première contribution de ma thèse a consisté à améliorer cette approche et à la rendre disponible et facilement utilisable dans un package R, comme détaillé dans la sous-section suivante.

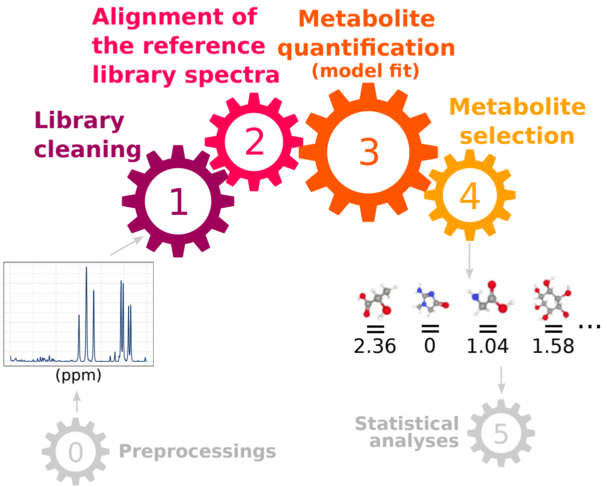


FIGURE 1.9 – Étapes de la quantification des métabolites à l'aide de la méthode ASICS.

1.2.3 Contribution du chapitre 2

Production scientifique

Cette sous-section résume la publication "G. Lefort, L. Liaubet, C. Canlet, P. Tardivel, M.-C. Père, H. Quesnel, A. Paris, N. Iannuccelli, N. Vialaneix, R. Servien (2019). ASICS : an R package for a whole analysis workflow of 1D 1H NMR spectra. *Bioinformatics*, 35(21) : 43564363". La méthode est implémentée dans le package R **ASICS** disponible sur Bioconductor.

Amélioration de l'alignement des spectres de la librairie

Dans la version d'ASICS de Tardivel et al. (2017), l'alignement de chacun des pics des métabolites de la librairie est réalisé grâce à la fonction de déformation $\phi(t) = at(1-t)+t$ où t correspond aux déplacements chimiques d'origine, $\phi(t)$ aux déplacements chimiques déformés et a au coefficient de déformation. Le coefficient a est optimisé de sorte que la déformation soit la meilleure possible tout en respectant un décalage maximum m_1 entre le pic d'origine et le pic déformé.

L'inconvénient de cette méthode est que la déformation est réalisée indépendamment pour chacun des pics d'un même métabolite. Les déformations ne sont donc pas toujours adaptées à tous les pics. Pour améliorer cela, j'ai proposé de réaliser l'alignement en deux étapes (Figure 2.2 page 30). La première étape aligne le spectre pur dans sa globalité grâce à la corrélation croisée de la transformée de Fourier (équation 1.2 page 10). Dans un deuxième temps, la déformation ϕ est utilisée mais sur un intervalle plus petit correspondant à un seul pic. Au lieu d'autoriser une déformation d'amplitude comprise dans $[-m_1, m_1]$, cette deuxième déformation a une amplitude comprise dans $[-\frac{m_1}{5}, \frac{m_1}{5}]$. Pour cela, le paramètre a est contraint par $a \in \left[-\frac{m_2}{\frac{\tau_2-\tau_1}{0.5^2}}, \frac{m_2}{\frac{\tau_2-\tau_1}{0.5^2}}\right] \cap [-1, 1]$ avec $m_2 = \frac{m_1}{5}$ et (τ_1, τ_2) les bornes inférieures et supérieures des déplacements chimiques initiaux.

Amélioration et validation des quantifications

Dans la version initiale de l'approche, l'optimisation des moindres carrés était réalisée sans contrainte sur la valeur de β_i, induisant des quantifications estimées potentiellement négatives qui étaient alors corrigées par $\tilde{\beta_i} = \max(\beta_i, 0)$. J'ai ajouté une contrainte de positivité, $\beta_i \geq 0$, au modèle 1.6. Cette contrainte permet d'améliorer les quantifications mais d'un point de vue théorique lors de la sélection des métabolite le FWER n'est plus contrôlé. Cependant, les différences étant minimes, il nous a paru raisonnable de garder cette procédure d'identification qui permet d'obtenir de très bon résultats. Les quantifications alors obtenues ont été validées plus précisément grâce à deux jeux de données comportant plus de 100 spectres. Ces jeux de données ont permis de mieux calibrer le choix par défaut des paramètres (décalage maximum autorisé m_1, redéfinition de l'amplitude pour le paramètre des déformations a, contrainte de positivité...) mais aussi de comparer les différentes méthodes existantes. La méthode Dolphin qui n'avait pas été comparer à ASICS lors de la publication de Tardivel et al. (2017) a, ici, pu l'être. De plus, une comparaison entre l'approche standard par *buckets* et l'approche par quantification avec ASICS a été réalisée et montre des résultats semblables. Les quantifications sont donc fiables et utilisables pour faciliter l'interprétation biologique des résultats.

Implémentation de la méthode dans le package R ASICS

Pour faciliter l'utilisation de cette méthode, un package R **ASICS** a été implémenté. Celui-ci permet une quantification automatique d'un ensemble de mélanges complexes. Celle-ci est rapide (environ 1 à 2 min par spectre) et est réalisable sur un ordinateur de bureau. De plus, des fonctions permettant le prétraitement des spectres depuis les spectres bruts FID ont été ajoutée tout comme des fonctions d'analyse des quantifications (ACP, OPLS-DA ou tests de Kruskal-Wallis). Ces fonctions permettent à l'utilisateur de réaliser toute son analyse sans avoir à se préoccuper de mettre en forme ses données à chaque étape. Enfin, une librairie importante de près de 200 spectres purs est disponible et il est facile pour l'utilisateur d'y ajouter des nouveaux spectres purs d'intérêt.

1.2.4 Contribution du chapitre 3

Production scientifique

Cette sous-section résume la publication "G. Lefort, L. Liaubet, N. Marty-Gasset, C. Canlet, N. Vialaneix, R. Servien (2021). Joint automatic metabolite identification and quantification of a set of ^{1}H NMR spectra. *Analytical Chemistry*, 93(5) : 28612870". La méthode est implémentée dans la version 2.0 du package R **ASICS** disponible sur Bioconductor.

Dans la majorité des expériences réalisées, on dispose de plusieurs spectres complexes. Ces spectres sont similaires (*i.e.*, à peu près les mêmes métabolites sont présents dans tous les échantillons) car les individus d'une même expérience sont relativement proches. Dans les méthodes décrites précédemment, l'alignement des spectres de la librairie et la quantification sont réalisés indépendamment pour chacun des spectres complexes. Lorsque l'on analyse un ensemble de spectres provenant d'une même expérience, les spectres purs sont donc alignés sur chaque spectre complexe et les métabolites sont quantifiés pour chacun des spectres complexes un par un. Ces spectres complexes étant similaires, on pourrait souhaiter améliorer ces deux étapes en utilisant tous les spectres en même temps, afin de profiter de l'information jointe qu'ils amènent.

À notre connaissance, aucune méthode d'alignement de spectres ne prend en compte l'ensemble des spectres d'une même expérience. L'alignement est toujours réalisé pour chaque spectre contre un spectre choisi comme référence comme pour les méthodes décrites dans la section 1.2.1. La quantification aussi est toujours réalisée indépendamment pour chacun des spectres, cependant des modèles statistiques à réponses multiples existent et permettraient d'estimer simultanément les quantifications d'un ensemble de spectres. Nous avons donc développé une méthode d'alignement joint ainsi qu'une méthode de quantification simultanée d'un ensemble de spectres complexes qui utilise une approche group-Lasso.

Alignement joint des spectres de la librairie pour un ensemble de spectres complexes

Comme décrit dans le chapitre 2, l'alignement des spectres purs sur chacun des mélanges complexes est réalisé en deux étapes : d'abord le même décalage est réalisé globalement sur l'ensemble du spectre pur et ensuite chacun des pics du spectre pur est déformé localement. L'alignement local permet de prendre en compte les différences de décalage le long du spectre mais aussi les faibles variations qui peuvent exister entre les différents mélanges complexes. Il est donc important que l'alignement local soit réalisé pour chacun des spectres complexes indépendamment. Lors de l'alignement global, les FFT sont utilisées pour trouver le meilleur décalage. Cet indicateur, utilisé dans beaucoup de méthodes d'alignement (comme icoshift ou **speaq**), permet d'obtenir de très bons résultats lorsque l'on aligne deux spectres complexes. Cependant, dans le cas des spectres purs, l'alignement n'est pas toujours adéquat. Un moyen d'améliorer l'alignement est alors de contraindre le décalage maximum autorisé m_1 ou mieux de l'adapter à chacun des spectres purs comme ceux-ci peuvent avoir été obtenus dans des conditions expérimentales très différentes. Pour calibrer ce décalage maximal m_1, nous avons choisi d'utiliser l'information de l'ensemble des spectres complexes de l'expérience. En utilisant cette méthode jointe, les résultats sont meilleurs qu'avec l'approche indépendante ou en utilisant les autres méthodes existantes sur des données simulées ainsi que sur des données réelles de plasma de porcelets à la naissance.

Quantification et identification jointes d'un ensemble de spectres complexes

En se basant sur la même idée, nous avons voulu quantifier les métabolites de l'ensemble des spectres simultanément. Lorsque les quantifications sont réalisées indépendamment, il arrive très souvent qu'un métabolite ne soit identifié que dans très peu de spectres ce qui correspond généralement à une mauvaise identification (*i.e.* un métabolite absent tout le temps mais détecté peu souvent). Nous souhaiterions pouvoir nous servir de l'information jointe des différents spectres afin d'éviter ces mauvaises identifications, en utilisant par exemple une pénalité sur les métabolites. Le modèle *group-Lasso* à réponses multiples, décrit dans Simon et al. (2013), est particulièrement approprié dans ce cas :

$$\underset{\beta \in \mathbb{R}^{p \times n}}{\arg\min} \left\{ \frac{1}{2} \|G - \Gamma F \beta^{\top}\|_F^2 + \lambda \sum_{i=1}^{p} \|\beta_i\|_2 \right\}, \quad \text{st } \beta_{ij} \geq 0 \tag{1.7}$$

avec Γ la matrice diagonale de covariance des résidus et $\|.\|_F$ la norme de Frobenius. Malgré les bons résultats sur les jeux de données réelles et simulées en terme de quantification, la contrainte de parcimonie du modèle 1.7 ne donne pas d'aussi bons résultats en terme d'identification que la sélection de métabolites basée sur le FWER. Pour des résultats optimaux, une combinaison des deux est donc possible : le modèle 1.7 est estimé seulement sur un sous ensemble de métabolites identifiés dans plus de r_c spectres complexes, ce seuil étant fixé par l'utilisateur.

1.3 Métabolome du porcelet en fin de gestation

Pour cette thèse, j'ai utilisé des données provenant de deux projets portant sur le métabolome du porcelet en fin de gestation et à la naissance : PORCINET (ANR-09-GENM-005) et SuBPig (INRAE GISA 2018-2019). C'est la nécessité d'une méthode de quantification et d'identification pour faciliter l'interprétation des données obtenues dans le premier projet qui a motivé cette thèse. En plus de fournir des données réelles ayant permis l'amélioration de la méthode ASICS, une partie de cette thèse a été consacrée à l'analyse des données du premier projet cité.

1.3.1 Mortalité périnatale chez le porc

Durant les dernières décennies, la sélection pour la prolificité chez le porc (*Sus scrofa*) a été associée à une augmentation importante de la mortalité néonatale (Canario et al., 2006, 2007). On observe aujourd'hui un taux de mortalité néonatale à la naissance compris entre 10 et 20 % (Edwards and Baxter, 2015), soit environ 15 000 morts par jour en France. Réduire la mortalité des porcelets en production porcine et améliorer leur vitalité est un enjeu important pour concilier des impératifs de productivité économique et une meilleure acceptabilité de la filière en termes de bien-être animal et d'impact sur l'environnement (par exemple, pour réduire l'usage des produits médicamenteux en élevage).

Les premières 24-48 heures après la naissance représentent la période la plus critique pour la survie du porcelet. Plusieurs facteurs influençant la survie du porcelet à la naissance, mais aussi après celle-ci, ont été identifiés (Edwards and Baxter, 2015). Ces facteurs sont liés à des effets maternels (durée de mise bas, santé. . .), au porcelet (type génétique, vitalité. . .), à des caractéristiques du porcelet qui sont partiellement sous contrôle maternel (poids du porcelet à la naissance) ou à une combinaison de ces facteurs (Baxter et al., 2008 ; Edwards and Baxter, 2015). La maturité du porcelet, qui est définie comme le complet développement permettant la survie à la naissance, est aussi un facteur important de sa survie et de sa croissance (Leenhouwers et al., 2002 ; Basso and Wilcox, 2010). Chez le porc, le processus de maturation est considéré comme se déroulant lors du dernier mois de gestation (entre 90 et 114 jours environ). La maturité à la naissance, avec les conditions environnementales, ont donc des conséquences majeures sur la mortalité néonatale du porcelet.

1.3.2 Projet PORCINET

Le projet PORCINET (ANR-09-GENM-005) a pour objectif de mieux décrire les mécanismes se produisant durant le dernier tiers de la gestation et qui influence la mortalité périnatale. Dans ce but, deux races de porcs ont été comparées (Large White et Meishan) à deux stades de la fin de gestation (90 jours et 110 jours de gestation ; Foxcroft et al. (2006)). Ces deux races ont été choisies car elles diffèrent en terme de survie néonatale et peuvent donc permettre d'identifier des différences en lien avec une augmentation de mortalité néonatale. La race Large White est une race européenne qui a été sélectionnée pour une viande plus maigre et une prolificité plus élevée. Le haut taux de mortalité périnatale observé chez ces porcs est partiellement dû à une plus faible maturité à la naissance (Canario et al., 2007). Au contraire, la race Meishan présente un faible taux de mortalité à la naissance et les porcelets sont considérés plus matures malgré un petit poids de naissance (Herpin et al., 1993 ; Canario et al., 2006). De plus, les truies de ces deux races ont été inséminées avec une semence mixte de Large White et Meishan ce qui a permis a des fœtus au génotype croisé de grandir dans le même environnement utérin que des foetus de génotype pur.

Des prélèvements (fluides et tissus) ont été effectués sur près de 600 fœtus pour l'un ou l'autre des stades de gestation sur plusieurs tissus (Figure 1.10).

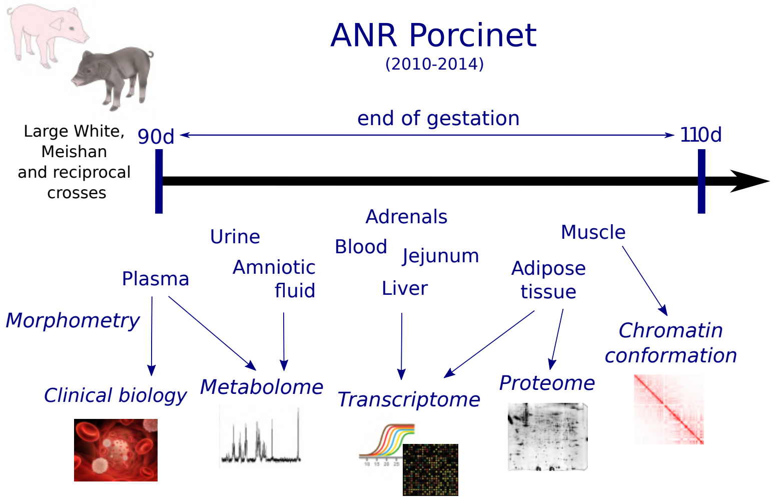


FIGURE 1.10 – Plan expérimental et données disponibles du projet PORCINET.

1.3.3 Contribution du chapitre 4

Production scientifique

Cette sous-section résume la publication “G. Lefort, R. Servien, H. Quesnel, Y. Billon, L. Canario, N. Iannuccelli, C. Canlet, A. Paris, N. Vialaneix, L. Liaubet (2020). The maturity in fetal pigs using a multi-fluid metabolomic approach. *Scientific Reports*, **10**, 19912”.

Dans cette étude, nous nous sommes concentrés sur l’analyse de l’un des types de données du projet PORCINET, le métabolome, le but étant de fournir une description du métabolomique des porcelet en fin de gestation. Ces résultats viennent compléter ceux déjà publiés sur les données transcriptomiques et protéomiques du muscle, de l’intestin et du tissu adipeux (Voillet et al., 2014, 2018 ; Gondret et al., 2018 ; Yao et al., 2017). Les profils métaboliques des fœtus ont été obtenus sur trois fluides : le plasma, l’urine et le liquide amniotique. Les spectres de RMN ont été quantifiés avec la méthode ASICS.

Pour chacun des fluides, des modèles mixtes ont permis d’identifier des métabolites pour lesquels la concentration varie entre les deux stades de gestation et/ou entre les génotypes. Certains métabolites, comme le myo-inositol et la proline, avaient déjà été identifiés chez le porc et d’autres mammifères pour être des biomarqueurs de survie néonatales et/ou de retards de croissance intra-utérins (Nissen et al., 2011 ; Liu et al., 2019). Ensuite, des analyses denrichissement de voies métaboliques ont été réalisées pour chacun des fluides. Les voies métaboliques enrichies en métabolites différentiels impliquent de nombreux acides aminés et sucres (croissance et apport d'énergie) ainsi que le métabolisme du glutathion (stress oxydatif) ce qui permet une meilleure description des mécanismes biologiques les plus pertinents pour décrire la fin de gestation.

Chapter 2

ASICS: an R package for a whole analysis workflow of 1D ^{1}H NMR spectra

Scientific production

The content of this chapter corresponds to the article “G. Lefort, L. Liaubet, C. Canlet, P. Tardivel, M.-C. Père, H. Quesnel, A. Paris, N. Iannuccelli, N. Vialaneix, R. Servien (2019). ASICS : an R package for a whole analysis workflow of 1D 1H NMR spectra. *Bioinformatics*, 35(21) :43564363”. The method is implemented in the R package **ASICS** available on Bioconductor.

Abstract

Motivation: In metabolomics, the detection of new biomarkers from NMR spectra is a promising approach. However, this analysis remains difficult due to the lack of a whole workflow that handles spectra pre-processing, automatic identification and quantification of metabolites and statistical analyses, in a reproducible way.
Results: We present **ASICS**, an R package that contains a complete workflow to analyse spectra from NMR experiments. It contains an automatic approach to identify and quantify metabolites in a complex mixture spectrum and uses the results of the quantification in untargeted and targeted statistical analyses. **ASICS** was shown to improve the precision of quantification in comparison to existing methods on two independent datasets. In addition, **ASICS** successfully recovered most metabolites that were found important to explain a two level condition describing

the samples by a manual and expert analysis based on bucketting. It also found new relevant metabolites involved in metabolic pathways related to risk factors associated with the condition.

Availability: ASICS is distributed as an R package, available on Bioconductor.

2.1 Introduction

Metabolomics is the comprehensive characterization of the small molecules involved in metabolic chemical reactions. It is a promising approach in systems biology for phenotype characterization or biomarker discovery, and it has been applied to many different fields such as agriculture, biotechnology, microbiology, environment, nutrition or health. Complementary analytical approaches, such as Nuclear Magnetic Resonance (NMR) or High-Resolution Mass Spectrometry, can be used to obtain metabolic profiles. These technologies allow routine detection of hundreds of metabolites in different biological samples (cell cultures, organs, biofluids...). But, due to their high complexity and to the large amount of generated signals, the analysis of such data remains a major challenge for high-throughput metabolomics.

This article focuses on NMR data, that is a promising tool to detect interesting biomarkers. The most common approach to deal with ^{1}H NMR spectra is to first divide them into intervals called buckets. The areas under the curve are computed for every bucket and every spectrum and these data are given as inputs to statistical methods to provide a list of buckets of interest (for instance buckets that are significantly different between two conditions). Since buckets are not directly connected to metabolites, this approach requires that ^{1}H NMR experts identify the metabolites from the extracted buckets. Not only is this identification step tedious, time consuming, expert dependent and not reproducible but it also leads to a serious loss of information since the identification of metabolites is restricted to the ones that correspond to extracted buckets (Considine et al., 2018).

Some methods have thus been developed to automatically identify metabolites from ^{1}H NMR spectra (MetaboHunter (Tulpan et al., 2011), MIDTool (Filntisi et al., 2017)) and others to automatically quantify the concentration of detected metabolites (Autofit (Weljie et al., 2006), **batman** (Hao et al., 2012), Bayesil (Ravanbakhsh et al., 2015) and **rDolphin** (Cañueto et al., 2018)); see Bingol (2018) for a complete review. Recently, Tardivel et al. (2017) defined a new statistical method to automatically identify and quantify metabolites that outperforms the other approaches. However, the approach mainly focuses on the quantification step and needed to be embedded in a complete pre-processing and post-processing analysis workflow, available through a simple tool. To our knowledge, such analysis workflows already existed (see a review in Misra (2018)) but they were usually restricted to some steps of the global analysis (post-processing, bucketing or statistical analysis). The only exception seems to be the W4M e-infrastructure (Guitton et al. (2017), available through the Galaxy platform[1]), whose automatic identification and quantification step is based on an earlier version of **ASICS** but the environment only allows one-by-one spectrum analysis. Furthermore, none of the existing workflow is as flexible, easily installed and embedded with other tools than an R package can be.

The R package **ASICS** (Automatic Statistical Identification in Complex Spectra) was designed to fill this gap. The identification and quantification method is partially based on Tardivel et al. (2017) but has been strongly revisited and improved to provide a fine tuning of all the parameters. Changes on the identification step (library distortion) and on the quantification step (model fitting) have also been implemented to improve the results and to reduce the computational cost. In addition, the method, that was only available under the form of separate and undocumented scripts, is now properly packaged and documented and the preprocessing of the spectra and post quantification statistical analyses have been implemented and are now part of the pipeline.

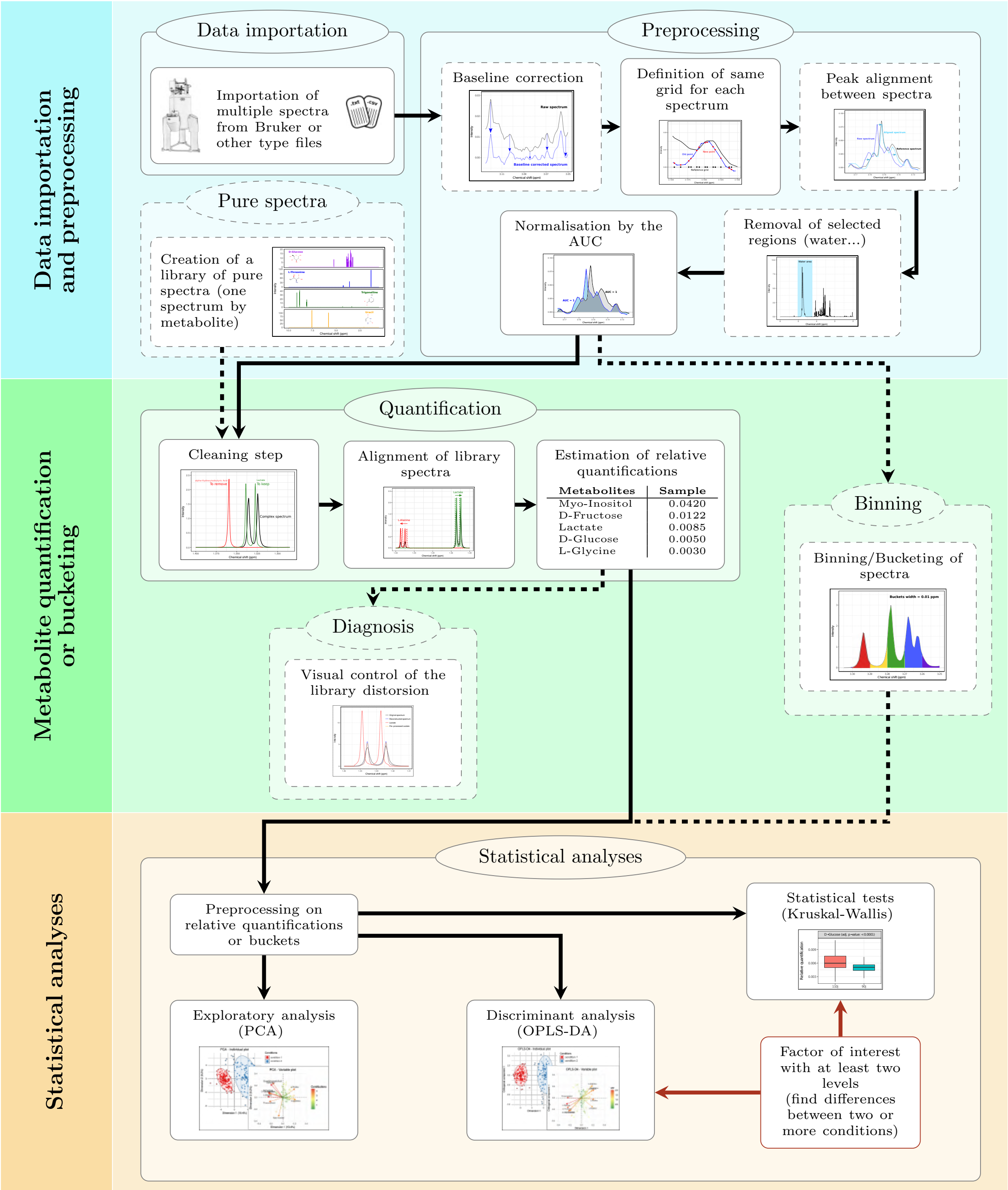
2.2 Material and methods

ASICS is an R package available on Bioconductor (Gentleman et al. (2004), http://bioconductor.org/packages/ASICS/) that combines all the steps of the analysis of ^{1}H NMR spectra (library of pure spectra management, preprocessing, quantification, post-quantification statistical analyses). The package also includes functions to directly perform statistical analyses on buckets and diagnosis tools to assess the quality of the quantification. All functionalities of the **ASICS** package are summarized in Figure 2.1 and described in the next sections.

2.2.1 Preprocessing the complex mixture spectrum

After the data are imported from raw 1D Bruker spectral data files or other types of files, several preprocessing steps are recommended in order to re-

[1] https://usegalaxy.org/


Data importation and preprocessing
Data importation
Importation of multiple spectra from Bruker or other type files
Preprocessing
Baseline correction
Raw spectrum
Baseline corrected spectrum
Definition of same grid for each spectrum
Peak alignment between spectra
Reference spectrum
Pure spectra
Creation of a library of pure spectra (one spectrum by metabolite)
Normalisation by the AUC
Removal of selected regions (water...)
Metabolite quantification or bucketing
Quantification
Cleaning step
To remove
To keep
Complex spectrum
Alignment of library spectra
Estimation of relative quantifications
Binning
Binning/Bucketing of spectra
Buckets width = 0.01 ppm
Diagnosis
Visual control of the library distorsion
Statistical analyses
Statistical analyses
Preprocessing on relative quantifications or buckets
Statistical tests (Kruskal-Wallis)
Exploratory analysis (PCA)
Discriminant analysis (OPLS-DA)
Factor of interest with at least two levels (find differences between two or more conditions)

Metabolites	Sample
Myo-Inositol	0.0420
D-Fructose	0.0122
Lactate	0.0085
D-Glucose	0.0050
L-Glycine	0.0030

Figure 2.1 – Schematic representation of **ASICS** workflow. Bottom box (with brown background): supplementary data (factor corresponding to experimental conditions for the different spectra) are required for this part of the analysis.

move technical biases.

Baseline correction

Most of ^{1}H NMR spectra have baseline distortions coming from various sources like instrument instability. These distortions can induce an increase or a decrease in peak intensities and skew the results of quantification. Wang et al. (2013a) developed a method to estimate the baseline for a spectrum by classifying each point as a signal or a noise point and by using a linear interpolation between noise points to construct the baseline. Then, the baseline is subtracted from its spectrum.

Peak alignment

Due to pH or temperature variations between the acquisition of multiple spectra, peak positions of the same metabolite can change between spectra. It is better to align all peaks before analyses, especially if a binning algorithm is used. Vu et al. (2011) developed an algorithm, implemented in the R package **speaq**, to carry out this alignment. It is based on continuous wavelet transform to detect peaks and hierarchical clustering to align all spectra on a reference one.

Removal of unwanted regions

It is also frequent to exclude a part of the spectra from the analysis. For instance, the part corresponding to water (4.5-5.1 ppm) is of no interest for most biological analyses and thus frequently removed prior to statistical analyses. Urea region (5.5-6.5 ppm) is also frequently excluded in case of urine samples.

Normalisation

A normalisation is mandatory before any analysis to make samples comparable. It allows to minimise systematic variations due to differences in sample dilutions. One of the most used methods is the normalisation to a constant sum (Craig et al. (2006)). As a result, the total spectral intensity is the same for each spectrum.

In **ASICS**, all preprocessing steps are available and the normalisation is the only mandatory one (it is systematically performed when the data are loaded). In the two following steps of the quantification method (preprocessing of the reference library, described in Section 2.2.2, and quantification itself, described in Section 2.2.3), all complex mixture spectra are processed individually and independently from each other. The method is thus described for only one complex mixture spectrum (and repeated similarly for all the others).

2.2.2 Preprocessing the reference library

A library of pure metabolite spectra is used as a reference to identify and quantify metabolite concentrations in the (complex mixture) spectra of interest. This library is a set of spectra of pure compounds, that have been acquired independently from samples. Such a reference library is available in **ASICS**. This library is composed of 190 spectra for which the noise has already been removed. The spectra acquisition procedure is detailed in Tardivel et al. (2017). In addition, **ASICS** provides functions to add or remove some spectra from the available reference library or to use another (user provided) reference library.

In addition to removing noise of each library spectrum, preprocessing steps are needed to clean and adapt the library to each spectrum of interest.

Noise thresholding

As this is the case for each ^{1}H NMR spectrum, all spectra in library contain noise. All values below a certain threshold, ρ_l, (that can be defined by the user; default value is $\rho_l = 1$), are considered as noise and set to 0. This allows to select peak positions, a step that is critical for the next selection stage.

First selection step

A metabolite can not belong to the complex mixture if at least one peak of its spectrum does not appear in the complex mixture spectrum peaks. Using this simple property, a first selection step is performed. All spectra in the reference library for which the peaks are not included in the peaks of the complex mixture spectrum are removed. This step results in a reference library of p pre-selected reference spectra that are used in the model described in Section 2.2.3. As technical biases can yield to chemical shifts, a reference spectrum is selected if all its peaks are present in the complex mixture spectrum with an allowed shift of m_1 ppm between the two spectra. In addition, as complex mixture spectra are noisy, peaks under a threshold ρ_m are ignored for this identification step. By default, the maximum allowed shift is $m_1 = 0.02$ ppm and the threshold is $\rho_m = 0.02$. These values have been calibrated on various real datasets with the help of NMR experts. However, both values can be changed by the user, depending on his spectrometer and experimental conditions.

Translation and distortion

The alignment algorithm described in Section 2.2.1 can not be used to align reference spectra with the complex mixture spectrum. The reason is that this method is not adapted to spectra with a low number of peaks as those of the pure metabolite contained in the reference library. Compared to Tardivel et al. (2017), this step is now split into two parts: a first step was added to globally shift the spectrum before a local peak distortion is performed in a second step (Figure 2.2):

1. First, reference library spectra are aligned with the complex mixture spectrum of interest by maximizing the Fast Fourier Transform cross-correlation (Wong et al. (2005)). The algorithm that finds the best shift (with a maximum allowed shift equal to m_1) is taken from the R package **speaq** (Vu et al. (2011)).

2. Second, every peak of each library spectrum taken individually is aligned by a local linear regression centered around each peak between the spectrum of interest and the reference library spectrum. To perform local distortions of the chemical shift grid for each peak, **ASICS** uses the function $\phi(t) = at(1-t) + t$, where $t \in [0,1]$, corresponds to the rescaled initial grid, $\phi(t) \in [0,1]$ to the newly scaled grid and $a \in \left[-\frac{\frac{m_2}{\tau_2-\tau_1}}{0.5^2}, \frac{\frac{m_2}{\tau_2-\tau_1}}{0.5^2} \right] \cap [-1,1]$ is a coefficient of distortion. The definition domain of a is controlled by m_2, the maximum allowed shift (with $m_2 = \frac{m_1}{5}$), and by (τ_1, τ_2) that are the lower and upper bounds of the initial grid, respectively. For each peak, different values of a are tested within this domain and the one that minimizes the residuals of the local linear regression is selected to distort this given peak. This results into a new (distorted) reference library used in the quantification algorithm.

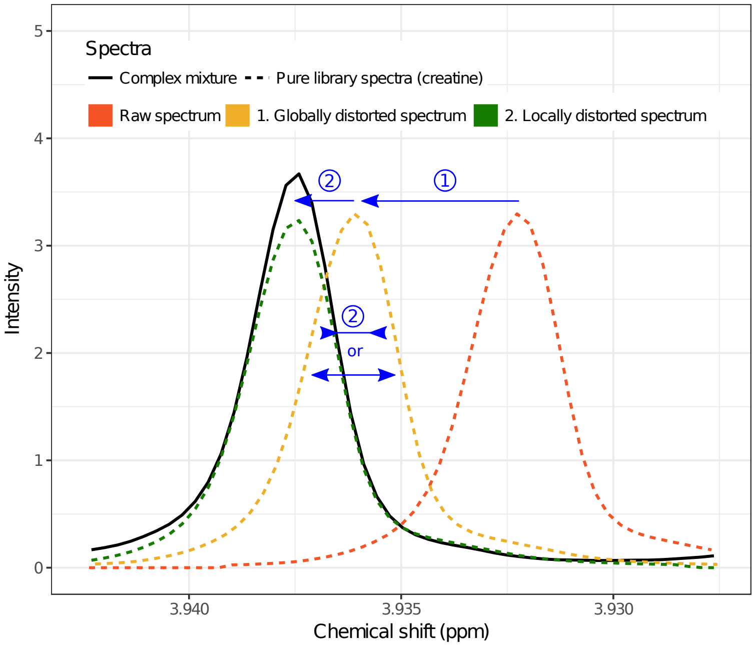


Figure 2.2 – Two steps distortion procedure for the main peak of the creatine. ① Global translation of the creatine spectrum. ② Local distortion of one of the creatine peak.

2.2.3 Metabolite quantification

Using the preprocessed complex mixture spectrum and the preprocessed spectra of the reference library, the metabolite identification and quantification in the complex mixture spectrum is performed similarly as in Tardivel et al. (2017). More precisely, the quantification methods does not use the Lasso (that gives biased estimates) anymore but it has been replaced by an faster unpenalized estimation followed by the control of the Family Wise Error Rate (FWER). The complex mixture spectrum is defined as a linear combination of the library reference spectra: $g(t) = \sum_{i=1}^{p} \beta_i f_i(\phi_i(t)) + \epsilon(t)$, with $\beta_i \geq 0$, where g corresponds to the complex mixture spectrum, $f_i \circ \phi_i$ to the p pre-selected preprocessed spectra of the reference library, $\beta = (\beta_1, \dots, \beta_p)$ to the coefficients associated with these spectra (or, equivalently, with the corresponding metabolites) and ϵ to the noise. The noise is structured so as to take into account both an additive noise, ϵ_2, and a multiplicative noise, ϵ_1: $\epsilon = \sqrt{\sum_{1 \leq i \leq p} \beta_i f_i \circ \phi_i}\, \epsilon_1 + \epsilon_2$.

A variable selection procedure is implemented to obtain a sparse β by controlling the Family Wise Error Rate (FWER) with a risk α. Usually, the threshold for rejecting $\mathcal{H}_0 : \beta_i = 0$ is the same for every i. Here, we used the procedure described in Tardivel (2017) that allows to define metabolite dependent thresholds in order to maximize the test power. More precisely, the custom thresholds ν_i are computed to minimize the volume of the acceptance region, namely $\arg\min_{(\nu_i)_i} \prod_{i=1}^{p} \nu_i$ subject to $\mathbb{P}^{\mathcal{H}_0}(|\hat{\beta}_1| \leq \nu_1, \ldots, |\hat{\beta}_p| \leq \nu_p) = 1 - \alpha$, where $(\hat{\beta}_i)_{i=1,\ldots,p}$ are MLE estimates of the previous linear model. The solution of this optimization problem is obtained by simulating a large number of realizations of the random variable $Z \sim \mathcal{N}(0, \Sigma)$, where Σ is the estimated variance of the estimates $(\hat{\beta}_i)_i$ so as to have $\mathbb{P}^{\mathcal{H}_0}(|\hat{\beta}_1| \leq \nu_1, \ldots, |\hat{\beta}_p| \leq \nu_p) = \mathbb{P}(|Z_1| \leq \nu_1, \ldots, |Z_p| \leq \nu_p)$, and the thresholds (ν_i) are obtained as the $1 - \alpha$ quantile of the random variable $\{|Z_1|, \ldots, |Z_p|\}$, that allows to control the FWER.

Once the metabolites selected, the quantifications $(\beta_i)_i$ for those selected metabolites are re-estimated by restricting the previous linear model to this subset in order to limit estimation bias. Finally, the relative quantifications are obtained by dividing $(\hat{\beta}_i)_i$ by the respective number of protons of each selected metabolite. In **ASICS**, pure library preprocessing and quantification are implemented in a unique function that can be run at once for several spectra with a parallel computing backend.

2.2.4 Post-quantification statistical analyses

On quantified metabolites (or on a subset of metabolites that are sufficiently frequently observed in the whole set of complex mixture spectra), the following analyses can be performed:

Quantification assessment

To assess the quality of **ASICS** quantification, a plot with the original complex mixture spectrum, $g(t)$, and the reconstructed spectrum, $\sum_{i=1}^{p} \beta_i f_i(\phi_i(t))$, can be obtained for a given sample (Figure 2.3). In addition, one reference spectrum for a given metabolite, and its distorted spectrum, can be superimposed to this plot in order to assess the quality of the metabolite selection for metabolites of interest.

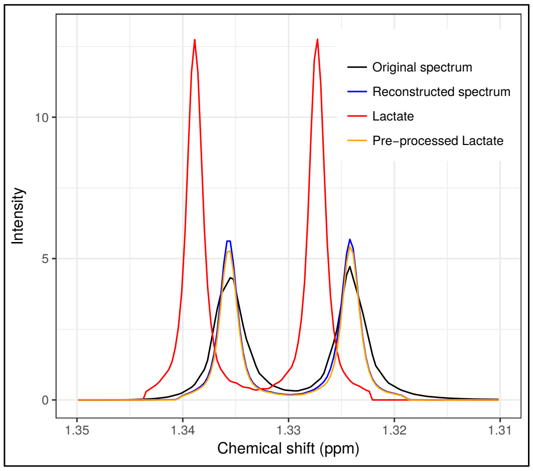


Figure 2.3 – Zoom in the diagnostic plot of the quantification to visually access the quality of the quantification of one of the lactate peaks and of the reconstructed spectrum, as compared to the original complex mixture spectrum.

Exploratory analysis

To explore results and detect outliers or batch effects, Principal Component Analysis (PCA) can be performed. Individual and variable plots are available to ease the visualisation and interpretation of PCA results (Figure 2.4).

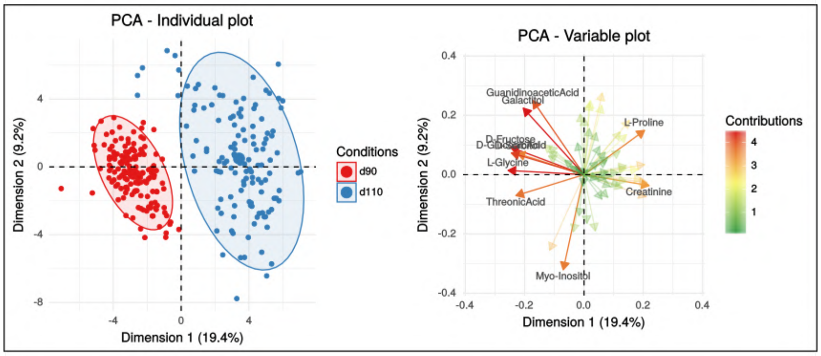


Figure 2.4 – PCA plots on **ASICS** quantifications for the study on plasmatic metabolome at the end of gestation in piglets. Left: individuals. Right: variables.

Discriminant analysis

When the samples correspond to two experimental conditions, Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA, Trygg and Wold (2002)) can be performed to find the metabolites with the highest discriminant power between these two conditions with a dedicated function based on the implementation available in the **ropls** package (Thévenot et al., 2015). Prediction error and variable importance in projection (VIP) are computed by a 10-fold cross-validation procedure, with a stability index for the VIP based on the results of the folds. Individual and variable plots are also available (Figure 2.5).

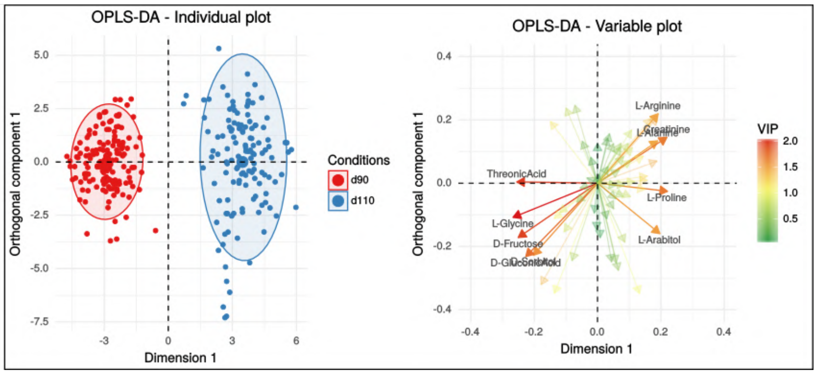


Figure 2.5 – OPLS-DA plots on **ASICS** quantifications for the study on plasmatic metabolome at the end of gestation in piglets. Left: individuals. Right: variables.

Statistical tests

To find differentially quantified metabolites, statistical tests have also been implemented. Since relative quantifications are usually non normally distributed, Kruskal-Wallis tests are used to find differences between the two (or more) groups, in combination with a correction for multiple testing, as available in the R function `p.adjust`. Boxplots showing the differences in metabolite quantification between the conditions can be displayed (Figure 2.6).

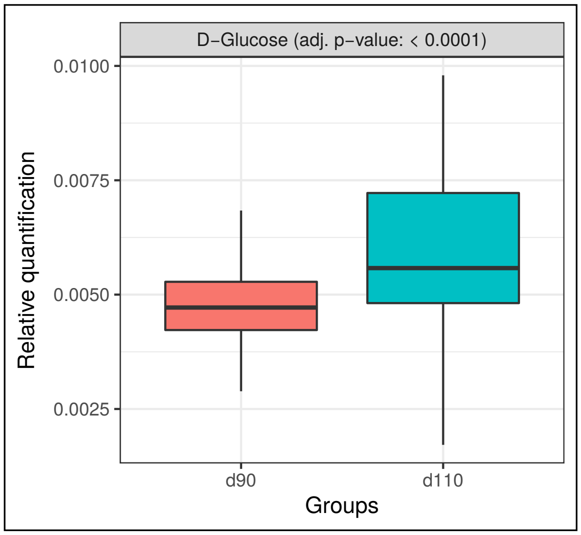


Figure 2.6 – Boxplots of the estimated glucose quantifications at the two gestational ages for the study on plasmatic metabolome in piglets.

2.3 Case studies

2.3.1 Plasma metabolome at the end of gestation in piglets

Genetic selection performed during the last decades has been associated with an increase in perinatal mortality in domestic pig, *Sus scrofa* (Canario et al., 2006, 2007). One main factor related to neonate survival is the maturation of fetal tissues and organs in late gestation (Voillet et al., 2014; Yao et al., 2017; Voillet et al., 2018; Gondret et al., 2018). In order to explore the development of the metabolic status in late gestation, an experiment was performed on pig fetuses. Metabolomic data were acquired on plasma samples collected on $n = 155$ Large White (LW) fetuses at 90 days of gestation and on $n = 128$ fetuses at 110 days of gestation (birth is around 114 days; ANR PORCINET project). All ^{1}H NMR spectra were phased and baseline corrected. Glucose, fructose, and lactate were directly quantified by standard methods (they have been chosen as indicators of carbohydrate metabolism). More details about the experimental design and data acquisition can be found in the frame “PORCINET project: design and data acquisition”.

PORCINET project: design and data acquisition

The PORCINET project (ANR-09-GENM-005) proposed to study the fetal development in late gestation in pigs. The experiment authorization number for the experimental farm GenESI (Genetics, testing and innovative systems experimental unit) is A 17661. The procedures performed in this study and the treatment of animals complied with European Union legislation (Directive 2010/63/EU) and French legislation in the Midi-Pyrénées Region of France (Decree 2001-464). The ethical committee of the Midi-Pyrénées Regional Council approved the experimental design (authorization MP/01/01/01/11). The experimental design was previously described in Voillet et al. (2014). In the present article, only Large White fetuses from PORCINET were taken into account. A total of 283 piglets collected at two gestational ages were considered. All sows ($n = 20$) were anesthetized at 90 days ($n = 155$ fetuses) or 110 days ($n = 128$ fetuses) after conception (average gestation term: 114 days). Their fetuses were quickly obtained by caesarean section. Blood (approximately 5 mL) was immediately collected from the umbilical artery via a 21-gauge needle and a 5-mL syringe and placed in heparinized tubes. Plasma was prepared by low-speed centrifugation (2,000×g for 10 min at 4°C) and stored at −20°C until further analysis. Glucose, fructose, and lactate were chosen as indicators of carbohydrate metabolism.

For proton nuclear magnetic resonance (^{1}H NMR) spectroscopy analysis, sample preparation was performed as follows: D2O (500 μL) was added to plasma (200 μL) and mixed, the sample was then centrifuged for 10 min at 3,000 × g at room temperature, and the supernatant (600 μL) was transferred to 5-mm nuclear magnetic resonance (NMR) tubes for ^{1}H NMR analysis. All ^{1}H NMR spectra were acquired on a Bruker Avance DRX-600 spectrometer (Bruker SA, Wissembourg, France) operating at 600.13 MHz for ^{1}H resonance frequency and equipped with a pulsed-field gradients z system, an inverse ^{1}H$-^{13}$C$-^{15}$N cryoprobe attached to a cryoplatform (the preamplifier cooling unit), and a temperature control unit maintaining the sample temperature at 300 ± 0.1°K. The ^{1}H NMR spectra of plasma samples were acquired at 300K using the Carr-Purcell-Meiboom-Gill (CPMG) spin-echo pulse sequence with presaturation with a total spin-echo delay (2nπ) of 240 ms to attenuate broad signals from proteins and lipoproteins, which otherwise display a wide signal and hide the narrower signals of low molecular weight metabolites. The ^{1}H signal was acquired by accumulating 128 transients over a spectral width of 20 ppm (note: chemical shift units kept ppm), collecting 32,000 data points. The interpulse delay of the CPMG sequence was set at 0.4 ms with $n = 300$ as defined in the following sequence: [90-(τ-180-τ)n acquisition]. A 2-s relaxation delay was applied. The Fourier transformation was calculated on 64,000 points. All ^{1}H NMR spectra were phased and baseline corrected. The ^{1}H chemical shifts were calibrated on the resonance of lactate at 1.33 ppm. Then plasma spectra were data-reduced before statistical analysis using AMIX software (Analysis of Mixtures version 3.8; Bruker Analytische Messtechnik; Rheinstetten, Germany). The spectral region δ 0.5 to 10.0 ppm was segmented into consecutive non overlapping regions of 0.01 ppm (buckets) and normalized according to the total signal intensity in every spectrum. The region around δ 4.8 ppm corresponding to water resonance (5.1–4.5 ppm) was excluded from the pattern recognition analysis to eliminate artifacts of residual water.

Similar analyses were performed on buckets and on relative quantifications computed with **ASICS** to assess the performance of the method. The aim of these analyses was to find metabolites best explaining the differences between the two groups: fetuses at 90 and 110 days of gestation. Lists of metabolites obtained with both approaches were compared as well as the direction of change between groups, based on two OPLS-DA, one on buckets and the other on quantifications obtained with **ASICS**. VIP thresholds for both OPLS-DA were set to 1.

Metabolites that were quantified were used to make a quantitative assessment of **ASICS** by comparing the obtained (estimated) quantifications with the dosages. Pearson correlation between quantifications and dosages were computed for every metabolite directly measured by dosage. These correlations were also compared with the correlations obtained for other quantification methods: Autofit, **batman**, Bayesil and **rDolphin**. Contrary to **ASICS**, these methods were too slow or not automated to allow the quantification for the 283 spectra. Therefore, quantifications were performed on a subsample of the original dataset that corresponded to the deciles of the lactate, fructose and glucose dosage to ensure representativity (32 spectra). Computational times were also recorded. For **ASICS** quantifications, water and urea regions were excluded and the maximum shift, m_1, was set to 0.01. In order to perform all quantifications with **batman** in a reasonable time, its library was reduced to the 160 common metabolites between **batman** and **ASICS** reference libraries and the number of iterations was set to 10,000.

2.3.2 Urinary metabolome of Type 2 diabetes mellitus

In order to test our method on data acquired with another spectrometer than the one on which the pure metabolite library included in **ASICS** has been obtained, we used the public datasets from Salek et al. (2007). The experiment has been designed to improve the understanding of early stage of type 2 diabetes mellitus (T2DM) development. ^{1}H NMR human metabolome was obtained from 84 healthy volunteers and 50 T2DM patients. Raw 1D Bruker spectral data files were found in the MetaboLights database (Haug et al. (2013); study MTBLS1). In the original study, spectra were normalized by the area under the curve after excluding water (4.24–5.04 ppm), urea (5.04–6.00 ppm) and glucose (3.19–3.99 ppm, 5.21–5.27 ppm) regions. Finally, a bucketing was performed with a 0.04-ppm width. The original study used a combination of PLS-DA and statistical tests (t-test, F-test, Kruskal-Wallis test and Kolmogorov-Smirnov test) on buckets (with a manual expert identification) to find differences between the healthy and ill individuals. This dataset allowed us to test the performance of **ASICS** on a different fluid (urine) in a different species (human).

Contrary to Salek et al. (2007), we kept glucose region for a quantification with **ASICS** because the glucose spectrum was available in the library. However, regions of water and urea were excluded. The other parameters of the different methods were set to their default values except for **ASICS** threshold that was set to $\rho_m = 0.05$, because we had observed that this dataset was noisier than the previous one. In addition, to control differences that could originate from the analysis method itself, we performed the comparison between the buckets and the ASICS quantifications with the same method, OPLS-DA, as for the study about perinatal survival (VIP thresholds set to 1.2).

2.4 Results and discussion

2.4.1 Comparison with biochemical dosages on piglets

Correlations between quantifications and biochemical dosages of the three metabolites were performed on the 32 selected spectra. We were not able to obtain quantifications with Bayesil because no chemical shift reference (TSP) has been added during spectrum acquisition. Bayesil handles spectrum from raw NMR induction-decay signal and so it requires that spectra are collected with TSP added to the sample (Ravanbakhsh et al., 2015; Beirnaert et al., 2018), TSP is sometimes used as an internal reference in samples for NMR. This procedure is not advised for plasma metabolome, and thus not routinely applied, since TSP binds to plasma proteins (Beckonert et al., 2007).

Table 2.1 provides the correlations between the quantified target metabolites and their corresponding dosages for the different quantification methods. In addition, the table includes the correlation between one bucket of the target metabolite and the corresponding dosage as a reference value. These results show that **ASICS** outperforms Autofit, **batman** and **rDolphin** for the three metabolites and provides quantification whose correlations are identical to the ones obtained with a direct comparison to the buckets. Results obtained with **batman** and the library with 160 metabolites are consistent with findings of other studies: the method is not suited for untargeted approaches (Tardivel et al., 2017; Beirnaert et al., 2018). If the quantification with **batman** is performed including only the three targeted metabolites in

Table 2.1 – Correlation between biochemical dosages of three metabolites and relative quantifications obtained with four methods and the buckets. Bucket for lactate: 1.335; bucket for fructose: 3.995; bucket for glucose: 5.235. Computational time is given for one spectrum.

	Lactate	Fructose	Glucose	Computational time	Parallel environment
ASICS	0.93	0.95	0.90	$\sim$ 1'30 min	Yes
Autofit	0.52	0.74	0.75	$<$ 1min	No
batman (with 160 metabolites)	0.46	0.56	0.22	$\sim$ 2 days	Yes
batman (with 3 metabolites)	0.55	0.70	0.82	$\sim$ 45 min	Yes
rDolphin	0.82	Not available	0.77	$\sim$ 1'30 min	No
Buckets	0.93	0.95	0.90	2 s	Yes

the reference library, correlations become similar to the ones obtained by the other methods, but are still lower than those obtained by **ASICS** with no prior selection of the reference library.

On a practical point of view, **ASICS** has other interesting features: first, it provides an easy way to handle (complement, replace, manipulate) the reference library whereas **batman** and **rDolphin** need that information on each multiplet (chemical shift position, multiplicity. . .) is specified. A biochemical expertise is thus required for the modification of the reference library in these packages. Autofit is a commercial software that requires the acquisition of a license, which strongly limits its use. Finally, the reference library cannot be modified in Bayesil and this method is only available through a web interface that makes automation of several spectra processing impossible.

In terms of computational times, the preprocessing of the library and the metabolite quantification with **ASICS** takes about 1'30 min per spectrum and can be launched at once in parallel. A parallel environment is also available for **batman** but the quantification of a single spectra takes approximately 2 days because of the use of a Bayesian framework that requires extensive MCMC simulations. Computational time needed by **rDolphin** is approximately the same than for **ASICS** but parallel implementation is not proposed in the package. Only Autofit has a lower computation time than **ASICS** (less than one minute) but spectra can only be quantified sequentially (no parallel environment).

A table summarizing capabilities of each method is available in Table 2.2.

Table 2.2 – An overview of open source NMR data processing solutions.

	ASICS	**Autofit** (Weljie et al., 2006)	**batman** (Hao et al., 2012)	**Bayesil** (Ravanbakhsh et al., 2015)	**rDolphin** (Cañueto et al., 2018)
Software	R	Chenomx	R	Web	R
Pre-processing	Yes	Yes	No	Yes	No
Alignment	Yes	No	No	No	No
Identification	Yes	Yes	Yes	Yes	Yes
Quantification	Yes	Yes	Yes	Yes	Yes
Data analysis	Yes	No	No	No	No
Parallel environment	Yes	No	Yes	No	No
Computational time	∼ 1'30 min	< 1min	∼ 2 days	∼ 10 min	∼ 1'30 min

2.4.2 Differences between gestational ages of fetuses

For the study about fetuses in late gestation, two outliers were detected on the bucket dataset in a preliminary study (Figure 2.7) and were removed from the analysis (Figures 2.8 and 2.9).

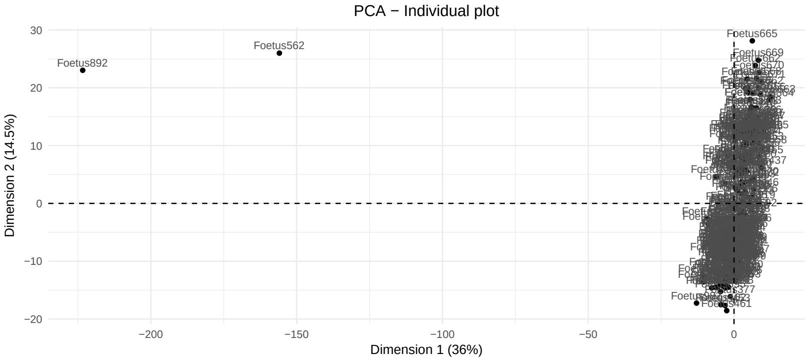


Figure 2.7 – PCA on buckets (axes 1 and 2, projection of individuals). Two outliers are identified that were removed from the analysis.

OPLS-DA was performed on quantified metabolites and on buckets. Both showed the same predicting power: all samples were perfectly separated according to their stages of gestation. For the bucket analysis, VIP values identified 268 buckets on 781 that were found influential to separate the two groups. Based on this list, a manual identification performed by an NMR expert hightlighted 21 metabolites.

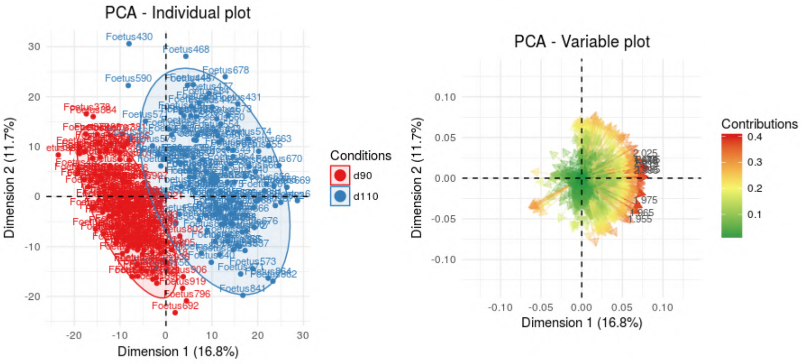


Figure 2.8 – PCA on buckets (axes 1 and 2, after the two outliers have been removed). Left: individuals. Right: variables.

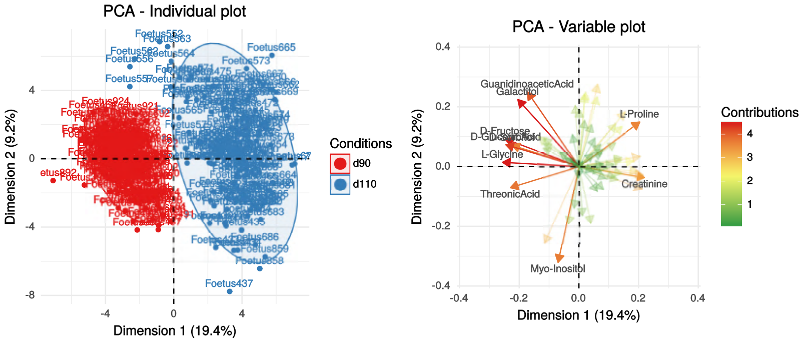


Figure 2.9 – PCA on **ASICS** quantification (axes 1 and 2, after the two outliers have been removed). Left: individuals. Right: variables.

The same analysis was performed on the **ASICS** quantifications and allowed to obtain 22 metabolites. The results obtained by **ASICS** and buckets analysis are detailed in Table 2.3 (page 47). Nine metabolites were found common to both analyses (Figure 2.10): lactate, creatinine, fructose, glucose, threonine, valine, alanine, proline and leucine.

For the metabolites which were not identified by both approaches, we observed five cases:

- metabolites only identified on buckets because the pure spectra was not present in the **ASICS** reference library: the 3-methyl-2-oxovaleric acid and the lipids;

- metabolites that were identified by **ASICS** but not selected as influ-

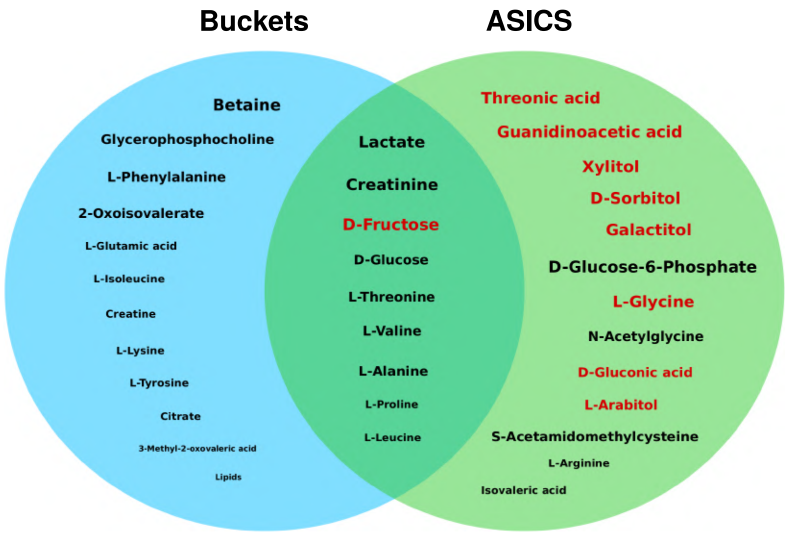


Figure 2.10 – Venn diagram comparing selected metabolites from analyses made on buckets (left) and on **ASICS** quantifications (right). Font size corresponds to average intensity of the associated buckets. A name is written in red if all peaks for this metabolite fall in the 3.5–4.2 ppm region (a region with a high density of peaks).

ential whereas the buckets corresponding to their peaks were: the betaine and the glutamic acid. Those metabolites indeed exhibited differences between the two groups (that were found significant by a Kruskal-Wallis test) but OPLS-DA did not select them as the most influential. This might be due to the fact that a fixed threshold of VIP equal to 1 is not be equivalent in the two approaches (**ASICS** quantification and direct bucket analysis). Also, dimension reduction performed with the quantification could have led to a modification of the correlation structure that determines which variables are the most influential in the OPLS-DA model;

- metabolites with low intensity peaks because **ASICS** was not able to identify and quantify smaller quantities: citrate, tyrosine, lysine, creatine and isoleucine;

- metabolites that were found by **ASICS** and not by the bucket analysis but for which all peaks corresponded to buckets that were found influential in the bucket analysis: the glycine and the guanidinoacetic acid. For this case, it is very likely that the non identification of these metabolites comes from an expertise bias (peaks are confused with glucose and fructose thus the expert does not identify it);

- metabolites for which no clear conclusion could be driven on their presence without expert knowledge in NMR or biology or the help of other technologies like 2D NMR spectrometry. For **ASICS** analysis these metabolites correspond to metabolites whose spectra have peaks only in the region with a high density of peaks (3.5 to 4.2 ppm; threonic acid, xylitol, sorbitol, galactitol, glucolic acid and arabitol), with a low concentration (N-acetylglycine, acetamidomethylcysteine, arginine and isovaleric acid) or with peaks confused with glucose peaks (glucose-6-phosphate).

The metabolites found by **ASICS** are consistent with known biological processes of late gestation in pig, especially with the fetal two-fold increase of weight during the last three weeks. It is expected to find up-regulation of the protein synthesis in late gestation, which is illustrated by the increase of amino acid abundances (alanine, proline, threonine, arginine, leucine, valine) just before birth. Also, functional analysis performed with IPA (see Figure 2.11) highlighted 13 metabolites (among the 22 identified by **ASICS**) involved in common metabolic pathways directly related to late stage gestation (survival or organism, metabolism of protein, conversion of lipid). Among these metabolites, 6 (guanidinoacetic acid, sorbitol, glucose-6-phosphate, glycine, gluconic acid and arginine) were identified only by **ASICS** and not with the bucket approach. In this study, the only weakness of **ASICS** is thus a tendency to miss low concentrated metabolites, especially if those have peaks only in the region with a high density of peaks.

2.4.3 Differences for T2DM patients

Results for the T2DM study are provided in Table 2.4 (page 48) and in Figure 2.12. The same conclusions than in Section 2.4.2 can be driven: some metabolites were extracted by both analyses (creatinine, betaine, hippuric acid, guanidinoacetic acid, alanine, glucose, indoxylsulfate, acetoacetate and trigonelline), some did not have a pure spectra available in the

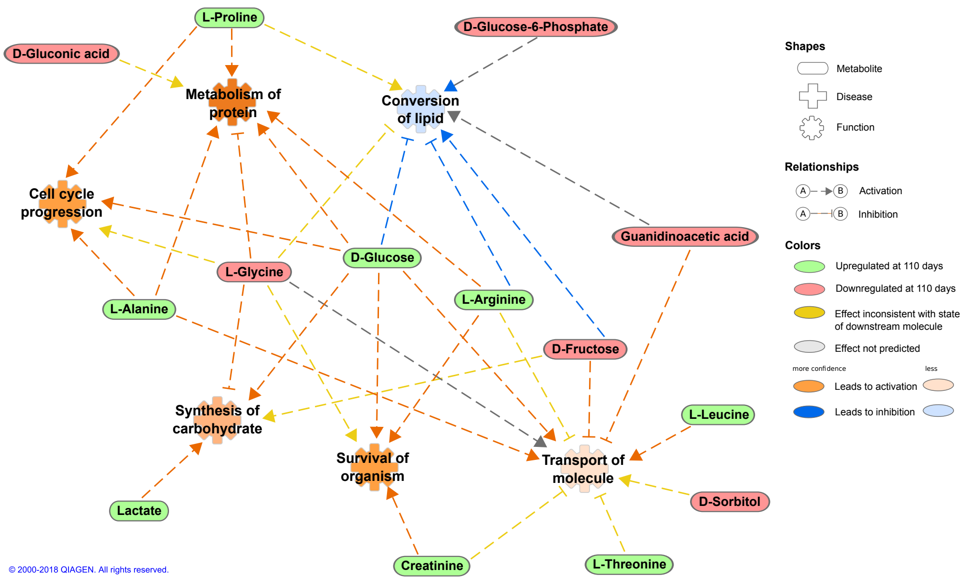


Figure 2.11 – Metabolomic pathway based on the metabolites identified by **ASICS** as obtained with Ingenuity Pathway Analysis© (IPA©, Ingenuity Systems; QIAGEN, Inc., Valencia, CA, USA, `https://analysis.ingenuity.com/pa`). IPA contains a large bibliographic database (Ingenuity Pathways Knowledge Base©). 13 out of 22 of the identified metabolites are present in the network, among which 6 (guanidinoacetic acid, sorbitol, glucose-6-phosphate, glycine, gluconic acid, and arginine) were identified only by **ASICS**.

library (phenylacetylglycine and 2PY) and the **ASICS** algorithm had difficulties to identify metabolites with low concentrations (3-hydroxybutyrate, isoleucine, 2-oxoisovalerate, fumaric acid and butyrate) or with only one proton (allantoin).

In addition, results were compared with those previously obtained by Salek et al. (2007) with the same NMR data and with those of an independent experiment realized on urine samples (among other samples) from T2DM patients with another non targeted metabolomic technology Yousri et al. (2015) (results also given in Table 2.4 page 48). Those comparisons highlighted the relevance of **ASICS** quantification that showed results consistent with previous studies and prior knowledge on Type 2 diabete: some of the metabolites were extracted by **ASICS** and by bucket quantification, like alanine or acetoacetate (Table 2.4 page 48), and have also been identified in Salek et al. (2007). We were also able to extract other metabolites, like the glucose (D-Glucose), the guanidinoacetic acid or the glycerol, that

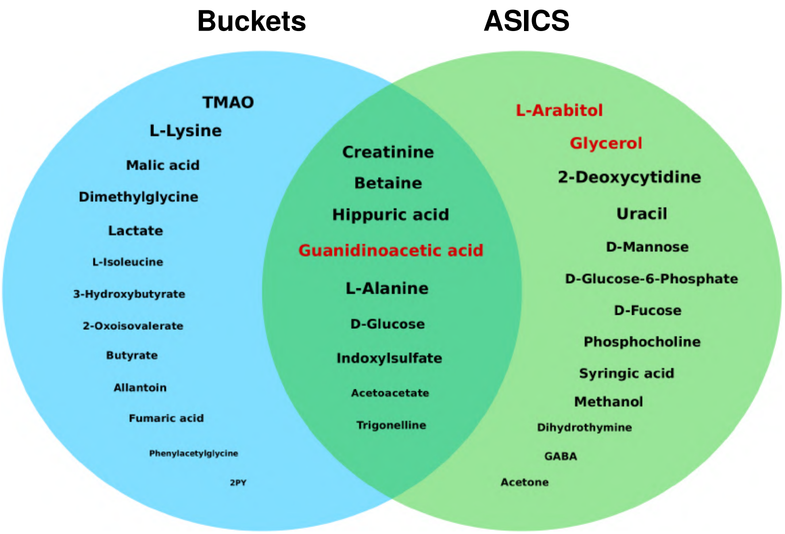


Figure 2.12 – Venn diagram comparing selected metabolites from analyses made on buckets (left) and on **ASICS** quantifications (right). Font size corresponds to average intensity of the associated buckets. A name is written in red if all peaks for this metabolite fall in the 3.5–4.2 ppm region (a region with a high density of peaks).

were not previously described because the glucose region was excluded from the study in Salek et al. (2007). The glycerol was identified both by **ASICS** and by Salek et al. (2007) in experiments on rats and mice (the glucose region was only excluded in the human dataset and not in the rat and mouse datasets). In all experiments, the glycerol increased in diabetics, which might reflect changes in fatty acids metabolism. With **ASICS**, the creatinine and its precursor, the guanidinoacetic acid (both also found with buckets), were directly quantified in urine and only the creatinine was previously described in (Salek et al., 2007; Yousri et al., 2015) as down regulated in T2DM. Both these metabolites reflect possible impairment of the renal function in diabetics.

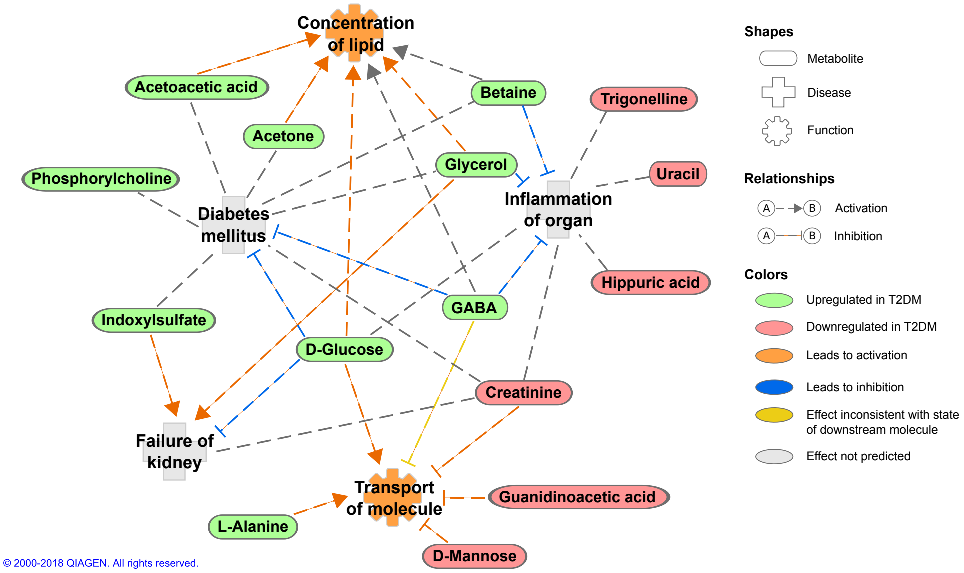
In addition, three metabolites (acetoacetate, acetone and 3-

hydroxybutyrate) reflected the presence of ketone bodies in urine when complications for diabete are likely to occur (Misra and Oliver, 2015). 3-hydroxybutyrate and acetoacetate are detected by Salek et al. (2007) and Yousri et al. (2015) together with buckets and **ASICS**. Acetone is only identified as discriminant by **ASICS** allowing the possibility to reflect the risk of acidocetose in diabetics. Another metabolite rarely identified in T2DM, arabitol (L-Arabitol), was quantified as decreasing only with **ASICS** and firstly described by Yousri et al. (2015) in urine of patients. Together with glucose-6-phosphate, also only identified with **ASICS**, these metabolites reflect the pentose pathway activity in diabetics. Metabolites associated to this pathway were also previously identified in urine as strongly associated with T2DM development in a diabetic rat model (Sun et al., 2014). Finally, only ASICS allowed the identification of GABA (γ-aminobutyric acid), a neuromediator recently identified to be increased in T2DM and related to a lower cognitive functioning observed in some diabetic patients (Van Bussel et al., 2016).

In conclusion, not only was **ASICS** able to automatically recover the main findings of the bucket and expert analysis, it was also able to extract a number of metabolites that are relevant and confirmed by other independent studies but not found by the bucket and expert analysis (glycerol, guanidinoacetic acid, acetone, arabitol, glucose-6-phosphate and GABA). This untargeted approach allowed to highlight several metabolic pathways linked to Type 2 Diabete Mellitus, as illustrated in Figure 2.13.

2.5 Conclusion

This article presents an R package, **ASICS**, integrating a complete analysis workflow of ^{1}H NMR spectra. This pipeline integrates an automatic metabolite identification and quantification method based on a reference library of pure metabolite spectra. **ASICS** showed better quantification results than existing methods and allowed to perform a complete and reproducible study on several hundreds spectra in only a few hours. Its use on two real world datasets exhibited similar results than the standard analysis on buckets followed by expert manual identification but also allowed to provide new information. For both studies, new metabolites, not extracted by expert identification, were found by **ASICS**, some of them confirmed by previous and independent studies. Obviously, as is the case for other omics data, in coming to a conclusion on whether the metabolites were really present in samples, a validation would be necessary.



© 2000-2018 QIAGEN. All rights reserved.

Figure 2.13 – Metabolomic pathway based on the metabolites identified by **ASICS** as obtained with Ingenuity Pathway Analysis© (IPA©, Ingenuity Systems; QIAGEN, Inc., Valencia, CA, USA, `https://analysis.ingenuity.com/pa`). IPA contains a large bibliographic database (Ingenuity Pathways Knowledge Base©). 15 out of 22 of the identified metabolites are present in the network, among which 6 (acetone, uracil, GABA, glycerol, phosphorylcholine, and mannose) were identified only by **ASICS**.

Finally, **ASICS** still has some limitations: the algorithm had difficulties to identify metabolites with low concentrations or with their peaks all located in a region with a high density of peaks. Future work will tackle this aspect, by trying to couple the information from the whole set of spectra to improve the individual quantification.

Funding

This project received financial support from French National Agency of Research (PORCINET grant ANR-09-GENM005). The PhD fellowship of Gaelle Lefort is supported by the Digital Agriculture Convergence Lab (#DigitAg, `http://www.hdigitag.fr/`, ANR-16-CONV-0004), by the INRA Mathematics and Computer Science Division, by the INRA Animal Genetics Division and by the INRA Animal Health Divison.

Table 2.3 – Metabolites selected by OPLS-DA as relevant to discriminate ages of gestation for both approaches. [a] buckets for metabolites extracted with the bucket approach and identified by an expert. [b] VIP for metabolites extracted with the **ASICS** approach.

Metabolite	**Buckets**[a]	**VIP**[b]	**Change at 110 days**
2-Oxoisovalerate	[1.12, 1.14]		↗
3-Methyl-2-oxovaleric acid	[1.09, 1.11]		↗
Betaine	[3.26, 3.27]		↗
Citrate	[2.51, 2.53], [2.54, 2.57], [2.67, 2.72]		↗
Creatine	[3.04, 3.05]		↗
Creatinine	[3.04, 3.05]	1.71	↗
D-Fructose	[3.55, 3.56], [3.57, 3.58], [3.69, 3.73], [3.78, 3.82], [3.89, 3.90], [3.98, 4.05]	1.96	↘
D-Gluconic acid		1.77	↘
D-Glucose	[3.23, 3.26], [3.38, 3.51], [3.74, 3.78], [5.22, 5.25]	1.04	↗
D-Glucose-6-Phosphate		1.07	↘
D-Sorbitol		1.58	↘
Galactitol		1.39	↘
Glycerophosphocholine	[4.28, 4.35]		↗
Guanidinoacetic acid		1.07	↘
Isovaleric acid		1.13	↗
L-Alanine	[1.46, 1.49]	1.48	↗
L-Arabitol		1.54	↗
L-Arginine		1.49	↗
L-Glutamic acid	[2.00, 2.09], [2.33, 2.38]		↘
L-Glycine		2.08	↘
L-Isoleucine	[0.92, 0.95], [1.00, 1.02]		↗
L-Leucine	[0.95, 0.98]	1.48	↗
L-Lysine	[1.68, 1.76], [2.99, 3.03]		↗
L-Phenylalanine	[7.31, 7.38], [7.40, 7.42]		↗
L-Proline	[1.96, 2.00], [3.32, 3.36]	1.72	↗
L-Threonine	[4.26, 4.28]	1.47	↗
L-Tyrosine	[6.88, 6.90], [7.18, 7.20]		↗
L-Valine	[0.98, 1.00], [1.03, 1.05], [2.20, 2.22], [2.25, 2.30]	1.32	↗
Lactate	[1.32, 1.35], [4.09, 4.14]	1.23	↗
Lipids	[0.88, 0.89], [0.90, 0.92]		↗
N-Acetylglycine		1.23	↗
S-Acetamidomethylcysteine		1.17	↗
Threonic acid		1.99	↘
Xylitol		1.17	↘
Unidentified buckets	72 buckets	-	-

Table 2.4 – Metabolites selected by OPLS-DA as relevant to discriminate T2DM patients for both approaches. [a] buckets for metabolites extracted with the bucket approach and identified by an expert. [b] VIP for metabolites extracted with the **ASICS** approach.

Metabolite	**Buckets**[a]	**VIP**[b]	**Change in T2DM**	**Change in T2DM in Salek et al. (2007)**	**Change in T2DM in Yousri et al. (2015)**
2PY	[6.66, 6.69], [8.32, 8.34]		↗	↗	
2-Deoxycytidine		1.67	↘		
2-Oxoisovalerate	[1.13, 1.14]		↘	↗	
3-Hydroxybutyrate	[1.20, 1.22], [4.14, 4.17]		↗	↗	↗
Acetoacetate	[2.26, 2.27]	1.59	↗	↗	↗
Acetone		1.69	↗		
Allantoin	[5.37, 5.40]		↗	↘	
Betaine	[3.25, 3.26], [3.90, 3.91]	2.13	↗		
Butyrate	[0.88, 0.89], [1.50, 1.55], [2.12, 2.13], [2.15, 2.18]		↗	↗	
Creatinine	[3.04, 3.05], [4.05, 4.07]	1.53	↘	↘	↘
D-Fucose		1.29	↘		
D-Glucose	[3.38, 3.41], [3.45, 3.55], [3.83, 3.92], [5.22, 5.25]	1.41	↗		↗
D-Glucose-6-Phosphate		1.44	↗		
D-Mannose		1.72	↘		↗
Dihydrothymine		1.23	↗		
Dimethylglycine	[2.92, 2.94]		↗	↗	
Fumaric acid	[6.52, 6.54]		↘	↘	
GABA		1.26	↗		
Glycerol		2.13	↗	↗	
Guanidinoacetic acid	[3.78,3.79]	2.41	↘		
Hippuric acid	[3.94, 3.98], [7.54, 7.58], [7.62, 7.66], [7.82, 7.85]	1.49	↘	↘	
Indoxylsulfate	[7.20, 7.24], [7.26, 7.30], [7.69, 7.70]	1.64	↗	↗	
L-Alanine	[3.76, 3.77], [3.79, 3.80]	1.38	↗	↗	↗
L-Arabitol		2.41	↘		↘
L-Isoleucine	[0.94, 0.95], [1.00, 1.01], [1.25,1.27]		↗	↘	↗
L-Lysine	[1.37, 1.44], [1.70, 1.71], [1.74,1.75], [1.93, 1.96], [3.00, 3.02]		↗		
Lactate	[4.11, 4.14]		↗	↗	↗
Malic acid	[2.33, 2.35], [2.37, 2.39], [2.65, 2.66], [4.31, 4.32]		↘	↘	↗
Methanol		1.69	↘		
Phenylacetylglycine	[3.67, 3.70], [3.72, 3.75], [7.42, 7.47]		↗	↘	
Phosphocholine		1.37	↗		
Syringic acid		1.54	↗		
TMAO	[3.25, 3.26]		↗	↗	
Trigonelline	[4.40, 4.45], [8.05, 8.06], [8.08, 8.09], [8.82, 8.86], [9.12, 9.14]	1.35	↘	↘	
Uracil		1.86	↘		
Unidentified buckets	114 buckets	-	-	-	-

Chapter 3

Joint automatic metabolite identification and quantification of a set of ^{1}H NMR spectra

Production scientifique

The content of this chapter corresponds to the article "G. Lefort, L. Liaubet, N. Marty-Gasset, C. Canlet, N. Vialaneix, R. Servien (2021). Joint automatic metabolite identification and quantification of a set of ^{1}H NMR spectra. *Analytical Chemistry*, 93(5) : 28612870". The method is implemented in the version 2.0 of the R package **ASICS** available on Bioconductor.

Abstract

Metabolomics is a promising approach to characterize phenotypes or to identify biomarkers. It is also easily accessible through NMR, which can provide a comprehensive understanding of the metabolome of any living organisms. However, the analysis of ^{1}H NMR spectrum remains difficult, mainly due to the different problems encountered to perform automatic identification and quantification of metabolites in a reproducible way. In addition, methods that perform automatic identification and quantification of metabolites are often designed to process one given complex mixture spectrum at a time. Hence, when a set of complex mixture spectra coming from the same experiment has to be processed, the approach is simply repeated independently for every spectrum, despite their resemblance. Here, we present new methods that are the first to either align spectra or to identify and quantify metabo-

lites by integrating information coming from several complex spectra of the same experiment. The performances of these new methods are then evaluated on both simulated and real datasets. The results show an improvement in the metabolite identification and in the accuracy of metabolite quantifications, especially when the concentration is low. This joint procedure is available in version 2.0 of **ASICS** package.

3.1 Introduction

Among omics, metabolomics is promising to identify potential biomarkers as the metabolites are close to the final phenotype and because of the experiment's moderate cost (Fiehn, 2002). Nuclear Magnetic Resonance (NMR) allows to obtain metabolomic profiles from easy-to-obtain fluids (*e.g.*, plasma, serum or urine), and NMR spectrometers produce a spectrum from a sample of one of these fluids. We will term such a spectrum a "complex spectrum" as it provides a profile of the quantification of all the metabolites contained in the sample (Nicholson and Wilson, 1989). However, the quantification is not direct: the complex spectrum is made of several peaks, where one peak can correspond to several metabolites, and one metabolite is described by one or several peaks –the quantity of the metabolite in the sample varying proportionally to the area under its peaks.

The classical approach to analyze such spectra consists in cutting each spectrum in small intervals, called buckets, and in computing the area under the spectrum of each bucket to perform statistical analyses (Alonso et al., 2015; Zhang et al., 2010). Since buckets are not directly connected to metabolites, this approach requires that NMR experts identify the metabolites from the buckets that are found relevant by the statistical analyses for a given biological question. This identification step is tedious, time consuming, expert dependent and, by consequence, not reproducible. It also leads to a serious loss of information since the identification of metabolites is restricted to the metabolites that correspond to extracted buckets (Considine et al., 2018).

To ease the use of NMR data, we developed a method, **ASICS**, which allows to automatically identify and quantify metabolites in NMR complex spectra (Tardivel et al. (2017); Lefort et al. (2019); R Bioconductor package at `https://bioconductor.org/packages/ASICS/`, including preprocessing steps and model fit). This method is based on a library of pure spectra (*i.e.*, spectra obtained from a single metabolite) that is used as a reference to fit a reconstruction model, limiting the effect of signal overlap between pure spectra. The model fit provides a measure of the relative quantity of metabolites in every sample (if an internal standard is used, absolute quantities can also be derived). This method has been evaluated in Lefort et al. (2019), where quantifications of metabolites were performed on urine of diabetic patients and on plasma of pig fetuses and were compared to a manual identification and quantification performed on a few targeted metabolites. Overall, the comparison showed that the automatic quantification provided results similar to the expert manual processing but in a much shorter amount of time and with an easily reproducible procedure. This makes this approach usable even for very large datasets (the overall processing of a complex mixture spectrum takes approximately 2 minutes on a standard laptop). It also showed that **ASICS** had a much better sensitivity / specificity trade-off than other automatic identification methods such as **batman** (Hao et al., 2014) or Bayesil (Ravanbakhsh et al., 2015) and improved quantification compared to targeted automatic quantification methods like **rDolphin** (Cañueto et al., 2018) or Autofit (Weljie et al., 2006).

However, we also showed that quantifications of less concentrated metabolites were of poorer quality, as is often the case in automatic methods, because these metabolites are hard to distinguish from the noise level. To improve the quantification of lowly concentrated metabolite, preprocessing steps of the analyzed complex spectrum are critical. Among critical preprocessings, one of them aims at aligning every pure spectrum of the reference library on the analyzed complex mixture. **ASICS** uses its own alignment, inspired by the alignment implemented in **speaq** (Beirnaert et al., 2018), but NMR tools include methods that were also designed to perform spectrum alignment, like icoshift (Savorani et al., 2010) or **speaq** (Beirnaert et al., 2018). However, whatever the identification and quantification tools, they are all designed to process the complex spectra one by one, independently, which is under-efficient when these come from the same experiment in closed conditions and thus share some similarities with one another.

Here, we present a new method to align pure spectra with the complex spectra of a sample of interest and to estimate quantifications that integrate information obtained from several complex spectra of the same experiment. The joint alignment is performed by automatically calibrating one of the parameters of the alignment algorithm. The joint quantification uses the joint alignment and is based on the use of a multivariate regression model incorporating a group sparse penalty. Both approaches are evaluated on simulated spectra (for which a ground truth is available) and on a real dataset of newborn piglet plasma and lead to improved identification and quantification, especially for lowly concentrated metabolites. This joint procedure is available in version 2.0 of **ASICS** package.

3.2 Methods and tools

3.2.1 General overview of the quantification strategy

Automatic identification and quantification of metabolites in a complex spectrum, g, is performed using a reference library of p pure spectra, $(f_i)_{i=1,\dots,p}$ (*e.g.*, spectra obtained from a single metabolite; Tardivel et al. (2017); Lefort et al. (2019)). The method then fits a model where the complex spectrum is decomposed into a linear combination of pure spectra in which the estimated coefficients divided by the number of proton u_i of the metabolite i, $(\beta_i)_i/(u_i)_i$, correspond to the quantification of the corresponding metabolites $i \in \{1,\dots,p\}$. To obtain valid quantifications, the coefficients $(\beta_i)_i$ are thus additionally constrained to be positive or null, which leads to the following model:

$$g(t) = \sum_{i=1}^{p} \beta_i\, f_i(t) + \epsilon(t) \quad \text{with } \beta_i \geq 0, \tag{3.1}$$

where $g(t)$ and $(f_i(t))_{i=1,\dots,p}$ respectively correspond to the complex spectrum to quantify at chemical shift t (in ppm) and to the ith spectrum in the reference library also at chemical shift t. The noise $\epsilon(t)$ is assumed to be structured such that $\epsilon(t) \perp\!\!\!\perp \epsilon(t')$ for $t \neq t'$ and includes both an additive noise $\epsilon_2(t)$ and a multiplicative noise $\epsilon_1(t)$ such that: $\epsilon(t) = \sum_{i=1}^{p} \beta_i f_i(t)\epsilon_1(t) + \epsilon_2(t)$ where $\epsilon_1 \sim \mathcal{N}(0,\omega_1^2)$, $\epsilon_2 \sim \mathcal{N}(0,\omega_2^2)$ and ω_1, ω_2 are user-defined values (`mult.noise` and `add.noise` respectively in **ASICS** R package).

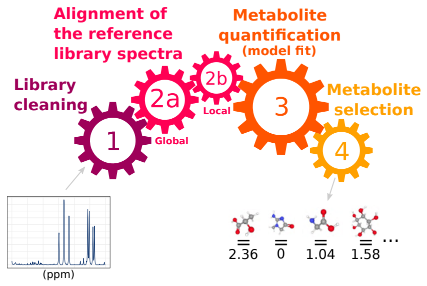

Alignment of the reference library spectra
Metabolite quantification (model fit)
Library cleaning
2a
2b
Local
3
Metabolite selection
1
Global
4
(ppm)
2.36
0
1.04
1.58
...

Figure 3.1 – Steps of the metabolite quantification of NMR spectra.

However, the model is fitted only after a number of preprocessing steps have been performed as illustrated in Figure 3.1 (see Lefort et al. (2019) for further details): a **library cleaning step** selects a limited number of relevant pure spectra in the reference library to be used in model (3.1) in order to improve its fit. Then, **two alignment steps** are performed to align the peaks of every selected pure spectrum, f_i, to the peaks of the complex spectrum g. These steps are necessary to correct peak shifts or distortions (expansion or narrowing) due to technical variations during the acquisition process (*e.g.*, pH or temperature). A global shift, s_i, is first estimated individually for every pure spectrum f_i and a refinement of this shift is then performed for every peak in f_i to estimate additional local shifts.

In addition, a postprocessing step is performed after the model of Equation (3.1) has been fitted. It aims at controlling the number of falsely selected metabolites. A **multiple testing selection procedure** based on FamilyWise Error Rate (FWER) is performed and consists in computing a threshold, ν_i, for each metabolite, which depends on all the estimated parameters $(\beta_{i'})_{i'=1,\ldots,p}$ and then in setting to 0 estimates (*i.e.*, quantifications) such that $\{\beta_i \leq \nu_i\}$.

The following two paragraphs will describe in more detail the two alignments steps, for which a joint version is proposed in this article. These alignments cannot be performed with usual NMR alignment methods because peaks are much more rare in pure spectra than in complex spectra and are thus harder to precisely bound (in complex spectra, a peak is naturally bounded by its neighbor peaks). Technical drifts are also generally larger because pure spectra usually cannot be acquired in the same batch of experiments. The solution consists in first obtaining a global shift s_i by optimizing:

$$s_i = \arg\max_{s \leq m_1} \mathrm{Cor}_{\mathrm{FFT}}(g(t), f_i(t+s)) \tag{3.2}$$

where $\mathrm{Cor}_{\mathrm{FFT}}$ is the the fast Fourier transform (FFT) cross-correlation (Wong et al., 2005) between the complex mixture g and a set of pure spectra f_i shifted by $s \leq m_1$ with m_1 a maximum shift defined by the user.

Then, each peak of the pure spectrum is independently aligned on the complex spectrum g locally, using a warping function that is constrained with a local maximum shift, $m_2 = m_1/5$. These two alignment steps result in an aligned reference library corresponding to the complex spectrum g whose quality is thus strongly conditioned on the user-defined parameter m_1.

When the quantification is performed on n complex spectra, $(g_j)_{j=1,\ldots,n}$, from the same experiments, a naive approach would be to perform all these steps *independently* for each complex spectrum. This would result in n different selections of the metabolites to be included in the model (library cleaning step) and in n aligned reference libraries. These aligned reference libraries all depend on a unique maximum allowed shift, m_1, defined by the user and that generates global shifts, $(s_{ji})_i$, and local shifts specific to the corresponding complex spectrum g_j. In addition, in Equation (3.1), the error term $\epsilon_j(t)$ and the estimated coefficients $(\beta_{ji})_j$ would all depend on the complex spectrum under study, independently from each other, as well as the thresholds, $(\nu_{ji})_j$ that control the FWER.

However, complex spectra from the same experiment share some common traits. It is thus expected that using joint steps, in which cleaning, alignment and quantification are somehow "constrained" to share similarities between all complex spectra of a same experiment or of a same condition within an experiment, has the potential to improve the overall quality of metabolite identification and quantification. In the next two sections, we describe two procedures for joint reference library alignment and joint metabolite quantification, respectively. Note that these two procedures are not meant to be used together (Figure 3.2): joint alignment aims at providing n aligned reference libraries for which the maximum shift allowed, m_1, is optimally and automatically tuned for information coming from all spectra rather than being user defined. This refined joint alignment has the potential to improve quantification of the metabolites when the model (3.1) is fitted independently for each complex spectrum (as illustrated in the Section "Results and discussion").

On the other hand, the joint quantification is a globally joint procedure that uses an aligned library that is common to the n complex spectra. It thus includes its own alignment step, derived from the joint alignment procedure and called the "common alignment" step. Advantages and drawbacks of these two joint approaches, depending on the experiment characteristics and on the user's expectations, are discussed in the Conclusion.

3.2.2 Joint alignment of the reference library

In the previously described alignment steps, the reference library is aligned independently on all complex spectra g_j and all pure spectra in the reference library f_i but this alignment depends on a unique maximum allowed shift, m_1, used for both the global and the local alignments. This parameter somehow represents the "typical maximum shift" expected for the experiment and it is critical to properly set the range of values that are maximized with the Corr_{FFT} measure as in Equation (3.2). Previous experiments have

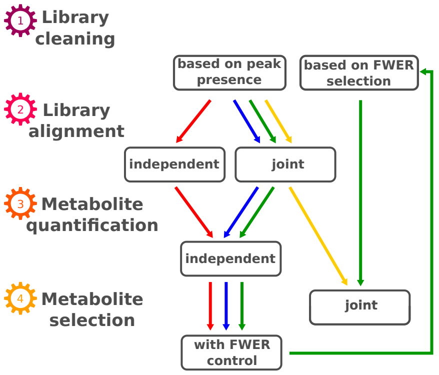


Figure 3.2 – Four different scenarios for automatic metabolite quantification (red: independent alignment and quantification, blue: joint alignment and independent quantification, yellow: joint alignment and basic joint quantification and green: joint alignment and joint quantification with FWER cleaning step). The other preprocessing steps (normalization, baseline correction, ...) are common to all approaches and described in Section 2.1 of Lefort et al. (2019).

shown a rather important sensitivity to this parameter and also that its value would be better determined depending on a given pure spectrum, f_i, because it presents high variations in relation with the range of the spectrum shifts.

The idea of the joint alignment of the reference library is therefore to automatically set a specific maximum allowed shift for each pure spectrum, m_{1i}, using information obtained from all complex spectra. This method thus increases the number of maximum shifts from 1 to p and provides more flexibility to account for the difference between pure spectra, while being more adapted to the given set of complex spectra. It is summarized in Figure 3.3 and the full method is given in Algorithm 1.

More precisely, for a given pure spectrum f_i, m_{1i} is tuned by performing a rough quantification based on several maximum shift candidates (steps 3-4 of the algorithm) and by independently computing a measure of fitness between this estimated quantification and a bucket area for all complex spectra (step 5). Even if the bucket area is a poor estimate of the true metabolite quantification, having this information from several complex spectra allows to make it usable to compute a relevant quality measure of the alignment preprocessing (step 10) and thus of each maximum candidate shift. The "best" maximum shift is therefore finally selected from this quality measure (step 12).

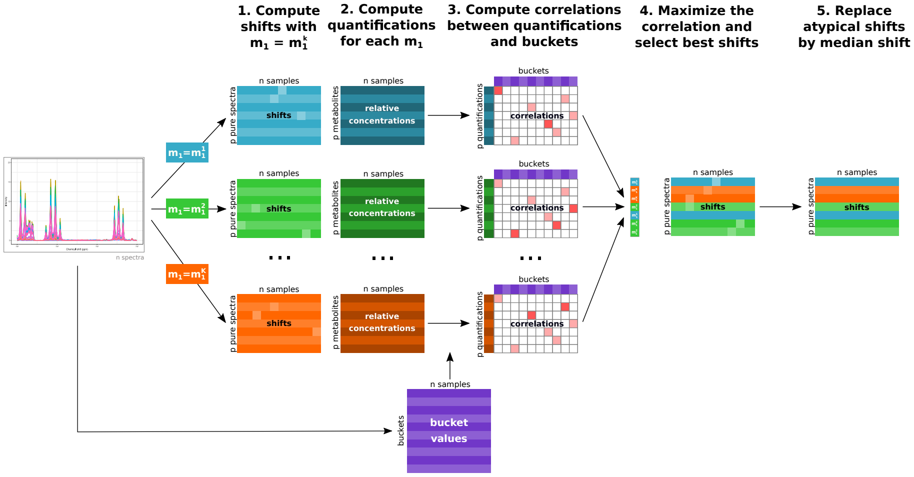

1. Compute shifts with $m_1 = m_1^k$
2. Compute quantifications for each m_1
3. Compute correlations between quantifications and buckets
4. Maximize the correlation and select best shifts
5. Replace atypical shifts by median shift
n spectra
$m_1=m_1^1$
$m_1=m_1^2$
$m_1=m_1^K$
n samples
p pure spectra
shifts
p metabolites
relative concentrations
buckets
p quantifications
correlations
...
bucket values

Figure 3.3 – Overview of the different steps of the joint alignment of the reference library.

Algorithm 1 Joint alignment of the reference library.

```
Require: set of candidate maximum shifts: $\mathcal{M} = (m_1^k)_{k=1,\dots,K}$
for all $m_1^k \in \mathcal{M}$ do
    for all $j = 1, \dots, n$ and $i = 1, \dots, p$ do    ▷ rough estimation of the quantification based on model fit and bucket areas
        find the best global shift $s_{ji}^*$ to align $f_i$ on $g_j$ by maximizing $\mathrm{Corr}_{\mathrm{FFT}}$ as in Equation (3.2)
        perform the model fit as in Equation (3.1): return quantification $\mathbf{Q}_{ji}^k = \boldsymbol{\beta}_{ji}^k / u_i$
        compute area of the bucket in $g_j$ at the position of every peak, $l$, in $f_i$: return bucket areas $(\mathbf{A}_{jil}^k)_l$
    end for
end for
for all $i = 1, \dots, p$ do
    for all $m_1^k \in \mathcal{M}$ do    ▷ assessment of the quality of maximum shift candidates
        evaluate quality of $m_1^k$ as: $\mathbf{C}_{ki} = \max_l \mathrm{Cor}(\mathbf{A}_{.il}^k, \mathbf{Q}_{.i}^k)$
    end for
    return $m_{1i}^* = \arg\max_{m_1^k \in \mathcal{M}} \mathbf{C}_{ki}$    ▷ selection of one maximum shift for every pure spectrum
end for
```

Global and local alignments of every pure spectrum f_i are performed for all complex spectra $(g_j)_j$ using the estimated maximum shift m_{1i}, and an additional joint post-processing step is then performed: the global alignment results in the computation of global shifts $(s_{ji})_j$, all smaller than m_{1i}. Outlier shifts (s_{ji} for which $|s_{ji} - \text{median}(s_{j'i})_{j'=1,\dots,n}| > 5 \times (t_2 - t_1)$) are thus further corrected and replaced by $\text{median}(s_{j'i})_{j'=1,\dots,n}$.

3.2.3 Joint metabolite quantification using a multivariate Lasso

In the standard procedure where complex spectra are all processed independently from one another, the identification of metabolites present in a given complex spectrum g_j is performed by a postprocessing step performed after the model fit. This procedure uses thresholds, ν_{ji}, based on FWER control, that are obtained independently for each complex mixture g_j and allows to decide whether the metabolite i should be selected or not. This approach allows to control the FWER of the metabolites in every complex spectrum g_j but can suffer from a lack of power. Since complex mixture spectra of a same experiment are expected to share a large fraction of common metabolites, the identification power of the procedure could be improved by using information from all spectra rather than performing the selection independently. In addition, in this independent approach, the quantification (model fit) and the identification (FWER control) are performed in two consecutive steps. The idea of the proposal described in this section is to address these two issues by designing a joint approach with a simultaneous identification and quantification that are based on all complex spectra at a time.

To do so, the idea is to fit a multi-response version of model (3.1), in which the simultaneously predicted values are G, the $(q \times n)$-matrix of columnwise complex spectra $(g_j)_{j=1,\dots,n}$. This requires to obtain an aligned reference library common to all complex spectra, F, which is made of the $(q \times p)$-matrix of columnwise aligned pure spectra $(f_i)_{i=1,\dots,p}$ and will serve as predictor of the multi-response version of model (3.1). In short, this common aligned reference library is based on the same preprocessing and postprocessing steps as the ones described in Figure 3.1 that are aggregated using, for instance, a user defined ratio of common evidence within complex spectra, r_c. It contains two cleaning steps, designed to reduce the size of the reference library, p, and global and local alignment steps. Technical details on how the common aligned library is obtained are provided in Algorithm 2.

The multi-response model is then based on a matrix version of the least

Algorithm 2 Preparation of a common aligned library.

```
Require: user defined ratio of evidence, $r_c \in ]0,1]$
for all $i = 1, \dots, p$ do                                          ▷ Cleaning step 1
    for all $j = 1, \dots, n$ do
        Perform independent cleaning steps (based on the presence of peaks of $f_i$ in $g_j$) return kept metabolites for $g_j$, $\mathcal{S}_j$
    end for                                                      ▷ End of Cleaning step 1
    Metabolites $i$ used to fit model (4) are the ones such that: $\frac{|\{i \in \mathcal{S}_j,\ j=1,\dots,n\}|}{n} \geq r_c$
end for
for all $j = 1, \dots, n$ do                                   ▷ Cleaning step 2 (optional)
    Perform alignment of the reference library and quantification of $g_j$ and FWER selection return selected metabolites for $g_j$, $\mathcal{S}'_j$
end for
for all $i = 1, \dots, p$ do
    Metabolites $i$ used to fit model (4) are the ones such that: $\frac{\left|\{i \in \mathcal{S}'_j,\ j=1,\dots,n\}\right|}{n} \geq r_c$
end for                                                          ▷ End of Cleaning step 2
for all $i = 1, \dots, p$ do                                        ▷ Global alignment
    Perform a joint alignment as described in Section "Joint alignment of the reference library" return global shifts $(s_{ji})_{j=1,\dots,n}$
    Align $f_i$ using the global shift $\tilde{s}_i = \text{median}(s_{ji})_{j=1,\dots,n}$
end for                                                      ▷ End of Global alignment
for all $i = 1, \dots, p$ do                                         ▷ Local alignment
    Perform local alignment of $f_i$ on a reference complex spectrum $g^{\text{ref}}$ defined as
        $g^{\text{ref}} = \underset{j=1,\dots,n}{\arg\max} \frac{1}{n} \sum_{j'=1}^{n} \text{Cor}_{\text{FFT}}(g_j, g_{j'}).$
end for                                                       ▷ End of Local alignment
return Common aligned reference library $F$
```

square minimization problem used to solve model (3.1), which writes:

$$\underset{\beta \in \mathbb{R}^{p \times n}}{\arg\min} \frac{1}{2} \|G - \Gamma F \beta^{\top}\|_F^2, \quad \text{st } \beta_{ij} \geq 0 \tag{3.3}$$

where Γ is the diagonal covariance matrix of the residuals and $\|.\|_F$ is the Frobenius norm. In this version, the quantification of the metabolite associated to pure spectrum f_i in complex spectrum g_j are based on the estimated coefficient β_{ij}.

In addition, the use of a Lasso-type penalty to the square loss of Equation (3.3) is known to be efficient for selecting variables (Tibshirani, 1996). This type of penalty indeed enforces the sparsity of the solution of the minimization problem, *i.e.*, the estimated coefficients $(\beta_{ij})_{i,j}$ are forced toward 0, except for those most important for the prediction quality. In our case, a desirable property would be that all $(\beta_{ij})_{j=1,\ldots,n}$ are forced toward 0 simultaneously for a given i, *i.e.*, that a given metabolite i is jointly identified or not identified for all samples. This can be performed by the use of a group-Lasso approach (Yuan and Lin, 2006), that is based on the ℓ_1-ℓ_2 norm $\sum_{i=1}^{p} \|\beta_{i.}\|^2$, with $\beta_{i.}$ the vector of length n, $(\beta_{ij})_{j=1,\ldots,n}$.

Finally, the solved minimization problem is identical to the one implemented in the R package **glmnet** (Friedman et al., 2010) and described in Simon et al. (2013):

$$\arg\min_{\beta \in \mathbb{R}^{p\times n}} \left\{ \frac{1}{2} \|G - \Gamma F \beta^\top\|_F^2 + \lambda \sum_{i=1}^{p} \|\beta_i\|_2 \right\}, \quad \text{st } \beta_{ij} \geq 0 \qquad (3.4)$$

The parameter $\lambda > 0$ is used to control the trade-off between the accuracy to the data (the error term computed with the Frobenius norm) and the model sparsity. It is usually tuned by cross-validation.

3.2.4 Implementation

Joint alignment and quantification are implemented in **ASICS** package version 2.0 (R Bioconductor package at https://bioconductor.org/packages/ASICS/). The user can define which approach to use (spectrum-dependent or joint alignment or quantification) by setting the following arguments:

`joint.alignment` to decide whether a joint alignment (if `joint.alignment=TRUE`) or an independent alignment (otherwise) is performed;

`quantif.method` to decide which type of quantification to perform. The choices are either `"FWER"` (independent quantification for every complex spectrum), `"Lasso"` (not including "Cleaning step 2" for common library alignment) or `"both"` (including "Cleaning step 2" for common library alignment). The fit of model (3.4) is performed using the R package **glmnet** (version 3.0-2) and the regularization parameter, λ, is also tuned by the cross-validation procedure available in this package. Note that if `quantif.method` is not set to `"FWER"`, the argument `joint.alignment` has no effect since the common alignment procedure of Algorithm 2 is automatically performed;

`clean.threshold` to set r_c when a joint quantification is performed.

3.3 Experimental data and design

The joint alignment and joint quantification performances were assessed separately using two datasets: a simulated dataset was first used because of the ease to obtain a ground truth (true shift or true quantification) for performance quantification. A real dataset, in which some metabolites have been directly quantified using dosages, was also used to evaluate both aspects (but with no ground truth available for the shift, the alignment quality was evaluated indirectly by its impact on the quantification quality). Our approach was also compared with state-of-the-art alternatives freely available to perform alignment and/or quantification.

3.3.1 Simulated spectra

To assess the performances of joint alignment and joint quantification, we first simulated n spectra $(g_j)_{j=1,\ldots,n}$ with metabolites in known concentrations, $\tilde{b}_{ji}$, from some of the p pure spectra $(f_i)_{i=1,\ldots,p}$ present in **ASICS** reference library. Five steps were necessary to simulate spectra:

1. a common set of metabolites was selected from the p pure spectra by using p independent Bernoulli random variables with parameter $r = 1/2$;

2. to introduce individual variations between the n simulated complex spectra, $d = 2$ additional metabolites were randomly chosen among all the metabolites, independently for each simulated complex spectra. More precisely, if the metabolite was already present in the common set of selected metabolites (respectively absent), it was removed (respectively added) in the set of selected metabolites for this specific complex spectrum. For $j = 1, \ldots, n$, this led to a maximum of four different metabolites between any two complex mixture spectra. In addition, we will denote p_j the number of metabolites present in the jth complex mixture spectrum;

3. $\forall\, j = 1, \ldots, n$ and $i = 1, \ldots, p_j$, ground truth quantifications, $(\tilde{b}_{ji})_i = (\tilde{\beta}_{ji})_i/(u_i)_i$, were then simulated using p_j independent normal distributions $\mathcal{N}(\mu_1, \sigma_1 = 0.3\mu_1)$ where μ_1 was itself generated from a log-normal distribution of parameters $\mu_2 = -8$ and $\sigma_2 = 2$. Quantifications smaller than 0 were set to 0, as well as quantifications larger than 1 that were set to 1, to avoid an unrealistically large skewness in the simulated quantifications;

4. for each metabolites, f_i global shifts were simulated independently for each spectra g_j using negative binomial distributions $s_{ji} \sim NB(2, 0.25)$ and local shifts were simulated independently using normal distributions $\tau_{jil} \sim \mathcal{N}(0, 0.09)$ with l corresponding to the lth peak of the pure spectrum f_i in the complex spectrum g_j. The final overall shift for this peak was then obtained as $r_{jil} = \min(s_{ji} + \tau_{jil}, m_1)$ with $m_1 = 0.02$. Finally, the direction of the shift (left or right), α_{jil}, was chosen using a Bernoulli distribution of parameter 0.5;

5. the simulated complex spectra $\tilde{g}_j$ were computed as follows: for all chemical shift t,

$$\tilde{g}_j(t) = \sum_{i=1}^{p_j} \tilde{b}_{ji} u_i f_i \left(t + (2\alpha_{jil(t)} - 1) r_{jil(t)}\right) \tag{3.5}$$

with $l(t)$ the peak at position t (if any), u_i the number of protons of the ith metabolite. This induces variation in line widths from one peak to another. Then, a noise was added based on Equation (1):

$$g_j = \epsilon_1 \tilde{g}_j + \epsilon_2$$

with $\epsilon_1 \sim \mathcal{N}(0, \omega_1^2 = 0.09)$ and $\epsilon_2 \sim \mathcal{N}(0, \omega_2^2 = 0.07)$.

Finally, the n complex spectra were normalized by the area under the curve. Parameters used to calibrate distributions for quantification simulations and shifts were obtained from previously analyzed real datasets.

They resulted in $n = 100$ simulated complex spectra, each composed of approximately $p_i \sim 82$ pure spectra that correspond to metabolites in known concentration. The simulation process itself was repeated to obtain 100 such datasets.

3.3.2 Plasma spectra of newborn piglets

In addition, the performances were also assessed on newborn pig metabolome, obtained during the SuBPig project (funded by INRAE GISA

2018-2019). In this project, ^{1}H NMR spectra were acquired on a Bruker Avance III HD NMR spectrometer (Bruker SA, Wissembourg, France) operating at 600.13 MHz for ^{1}H resonance frequency from plasma of 97 Large White newborns collected on umbilical cord. NMR raw spectra are available in the Metabolights database (Haug et al., 2013): MTBLS2137. The same samples were also used to obtained the concentrations of 27 targeted amino acids measured with an Ultra Performance Liquid Chromatography (UPLC). Details on the experimental protocol are available in the frame "SuBPig project: design and data acquisition" and basic statistics on amino acid dosages are provided in Table 3.1.

Table 3.1 – Minimum, maximum and median concentrations for each metabolites dosed with UPLC ($n = 97$).

Concentrations (in μmol/L)	Minimum	Maximum	Median
3-Methylhistidine	3.16	19.22	7.58
Alanine	270.08	1939.22	855.93
Arginine	25.49	150.97	69.35
Asparagine	16.57	130.55	49.66
Aspartic Acid	1.68	57.61	10.02
Carnosine	1.11	25.02	14.23
Citrulline	39.99	152.44	80.23
Cysteine	10.98	42.38	22.75
Ethanolamine	12.61	78.24	27.68
Glutamine	97.29	663.70	303.11
Glutamic Acid	44.34	567.18	153.17
Glycine	177.70	1902.22	473.74
Histidine	24.98	256.10	96.92
Hydroxyproline	42.56	140.38	70.05
Isoleucine	10.71	123.22	46.89
Leucine	24.09	229.82	79.58
Lysine	76.75	388.93	218.82
Methionine	4.97	79.77	13.43
Ornithine	11.32	70.94	34.00
Phenylalanine	9.00	107.22	55.40
Proline	86.85	384.19	169.51
Sarcosine	1.78	63.20	18.49
Serine	71.61	464.62	147.19
Taurine	19.93	214.85	59.22
Threonine	77.81	262.25	141.75
Tryptophan	11.97	26.83	19.30
Tyrosine	19.13	171.25	54.26
Valine	161.72	424.20	291.98

NMR spectra were preprocessed and quantified using **ASICS** with default procedure and parameters, except for the threshold under which the signal is considered as noise that was set at 0.01 and the multiplicative and additive noise standard deviations that were set at 0.07 and 0.09 respectively. Noises were set at realistic values using fourteen technical replicates of a pool sample. The alanine peak (1.47–1.50 ppm) was used to set the multiplicative noise and the noisy area (9.4–10.5 ppm) was used to set the additive noise. Details on the preprocessing of these spectra are available in the frame “SuBPig project: design and data acquisition”.

SuBPig project: design and data acquisition

Ethics statement

This study was conducted in accordance with the French legislation on experimentation and ethics. The French Ministry of Agriculture authorized this experiment on living animals at the INRAE facilities (UE1372 GenESI Génétique, Pig phenotyping and Innovative breeding facility, doi:10.15454/1.5572415481185847E12) with the agreement number APAFiS for animal housing and the agreement number #13648-2018020417291866 v4 for the protocol.

Plasma sample collection

Blood (approximately 5 mL) of the 97 piglets was collected individually on piglets from the umbilical cord and placed in heparinized tubes. Plasma was prepared by low-speed centrifugation (2,000 g for 10 min at 4°C) and stored at −80°C until further analysis.

NMR protocol

Each sample of plasma (200 μL) was diluted in 500 μL phosphate buffer prepared in deuterated water (0.2 M, pH 7.0) containing TSP (1.17 mM) as internal standard, vortexed, centrifuged at 5000 g for 15 min at 4°C, and 600 μL transferred into 5 mm NMR tube. All ^{1}H NMR spectra were acquired on a Bruker Avance III HD NMR spectrometer (Bruker Biospin, Rheinstetten, Germany) operating at 600.13 MHz for ^{1}H resonance frequency and at 300K, using the Carr-Purcell-Meiboom-Gill (CPMG) spin-echo pulse sequence. Spectrum preprocessing (group delay correction, solvent suppression, apodization, fourier transformation, zero order phase correction, internal referencing, baseline correction and window selection) was perform using the R package **PepsNMR** (version 1.2.1) with the TSP peak for internal reference. Finally, all spectra were aligned with each other using the method implemented in the **ASICS** package (as previously described in Section 2.1 of Lefort et al. (2019)).

UPLC protocol

Plasma amino acid concentrations were obtained using an ultra HPLC system (Waters Acquity Ultra Performance LC system, Waters, Guyancourt, France) coupled to an Acquity tunable UV detector and a mass detector (SQD detector) to identify the few coeluting chromatographic peaks. The column was a MassTrak AAA column (2.1 × 150 mm). Amino acid derivatization was performed with using an AccQůTag Ultra derivatization (MassTrak AAA Waters, Milford, MA). Norvaline was used as internal standard and a mixture of amino acids was used for calibration and quantification. The Empower 2chromatography software (Waters corporation, Milford, MA, USA) was used for instrument control and data acquisition.

3.3.3 Evaluation of the joint alignment

The joint alignment procedure was compared to independent alignment as performed in **ASICS** and in two other tools designed for that purpose: icoshift (version 3.0; Savorani et al. (2010)) and **speaq** (version 2.6.1; Beirnaert et al. (2018)). All alignment methods were run for both datasets (simulated dataset and piglet plasma dataset) and, on the simulated dataset, 100 simulations of 100 complex spectra were performed to ensure the robustness of the results. In addition, assessment of the performance was not obtained identically for both datasets.

For each simulated dataset, a cosine similarity was computed for any metabolite i between the true (unknown) contribution of its given pure spectrum, f_i, to the simulation of g_j (ground truth) and the result of the alignment of f_i on g_j. For the sake of simplicity, this similarity was computed using the alignment obtained on a single reference complex spectrum, g_{j^*}, that was the most similar (in terms of average cosine similarity) to all other complex spectra. This measure allowed to use the ground truth of the simulation to assess the quality of the alignment in a simple and efficient way.

In addition, the non-parametric Durbin test (as implemented in the R package **PMCMR**; Conover (1999); Pohlert (2014)) was used to test the significance of the differences in cosine similarity between different alignment methods. The Durbin test allows to account for the pairing of metabolites across experiments and is also able to cope with the incompleteness of block design that is due to the fact that different metabolites are used to generate the reference complex spectrum g_{j^*} across simulated complex spectra within one dataset.

Once the reference library had been aligned, it was submitted to the **ASICS** independent quantification algorithm. The effect of the quality of the alignment on the quality of the identification and on the quantification was assessed. The metabolite identification quality was evaluated by comparing the identified metabolites with the metabolites truly used in the simulation. The significance of the difference in method sensitivity and specificity was assessed using Kruskal-Wallis test followed by the post-hoc Nemenyi test.

Finally, the metabolite quantification quality was evaluated by computing the correlation between the estimated metabolite quantification b_i and the ground truth metabolite quantification $\tilde{b}_i$ across $j = 1, \ldots, n$. As for alignment quality, the significance of the differences between methods was tested using the Durbin test.

For the piglet plasma dataset, we did not know all metabolites that were truly present in the complex spectra so we could not perform the direct evaluation of the alignment quality, nor the evaluation through the quality of metabolite identification. However, we were able to assess the impact of the alignment on the quality of some metabolites' quantification. This was done by computing correlations between estimated quantifications and UPLC concentrations, which are used as reference measures here. The significance of the differences between methods was tested using the Durbin test followed by post-hoc Durbin tests.

3.3.4 Evaluation of the joint quantification

Different scenarios of the joint quantification method were evaluated: joint quantification with a single cleaning step (in yellow in Figure 3.2), joint quantification with a second cleaning step (in green in Figure 3.2), for which several values of the ratio of common evidence (r_c) were tested: $r_c \in \{1\%, 10\%, 50\%\}$. This joint quantification procedure was compared with quantifications obtained with **ASICS** independent quantification (in blue in Figure 3.2). On the piglet plasma dataset, we also compared the results with another quantification method, performed independently on each complex mixture spectrum: the one implemented in the R package **rDolphin** (Cañueto et al. (2018); which was the alternative quantification method which performed best among those tested in Lefort et al. (2019)). This method requires to provide a list of targeted metabolites for which the quantification has to be performed. This list is naturally provided by the UPLC dosages in the piglet plasma dataset, but no such natural choice is available for the simulated dataset.

The quality of the quantification was assessed as already described in Section "Evaluation of the joint alignment", by correlation between estimated quantifications and simulated ones (simulated dataset) or either by correlation between estimated quantifications and UPLC concentrations (piglet plasma dataset). Note that **rDolphin** produces a quantification for several regions of interest that it has identified in the metabolite pure spectrum. We chose to keep only the highest correlation with the UPLC concentrations in our final results in order to show the "best case scenario" of **rDolphin**.

3.4 Results and discussion

3.4.1 Evaluation of **ASICS** joint alignment procedure

Figure 3.4 provides cosine similarities between the true contribution of f_i to the simulation of g_j (ground truth) and the result of the alignment of f_i on g_{j^*} for the simulated dataset. This shows that the joint alignment outperforms the other methods. In addition, differences between methods (p-value < 0.001; Durbin test) as well as pairwise differences (p-values $<$ 0.001 for all pairs; Durbin post-hoc test) were all found significant.

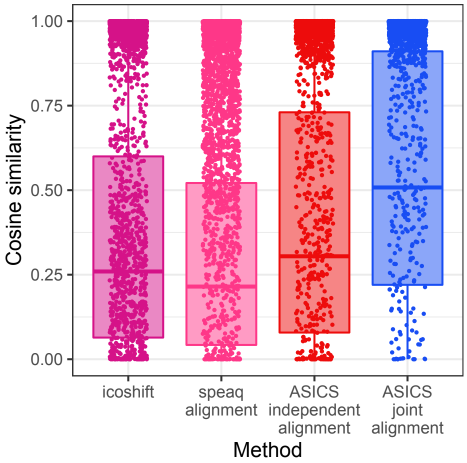


Figure 3.4 – Cosine similarity between the true contribution of f_i to the simulation of g_j (ground truth) and the result of the alignment of f_i on g_{j^*}. Alignments were performed with icoshift, **speaq** or **ASICS** (independent and joint versions) for 100 reference spectra corresponding to the 100 simulations. Points correspond to the cosine similarity of the 30 more concentrated metabolites in every simulation.

The median cosine similarity for **ASICS** joint alignment is equal to 0.51 overall but increases to 0.99 when computed on the 30 more concentrated metabolites only. This is explained by the fact that peaks of very lowly concentrated metabolites are usually under noise signal in the complex spectra and are thus not or poorly detected. For these upmost concentrated metabolites, the median cosine similarity is equal to 0.97 for icoshift and to 0.90 for speaq, both results still significantly differ from **ASICS** joint alignment performances (p-values < 0.001 in both cases; Durbin post-hoc test).

A similar positive impact of the joint alignment was also obtained on subsequent identifications and quantifications for the simulated dataset (Table 3.2 and Figure 3.5).

	icoshift	**speaq**	**indep.**
speaq	1.00	-	-
indep.	0.95	0.91	-
joint	0.84	0.90	0.52

(a) Sensitivity (global p-value $= 0.60$; Kruskal-Wallis test)

	icoshift	**speaq**	**indep.**
speaq	0.99	-	-
indep.	0.005	0.003	-
joint	0.04	0.05	< 0.001

(b) Specificity (global p-value < 0.001; Kruskal-Wallis test)

	icoshift	**speaq**	**indep.**
speaq	0.41	-	-
indep.	< 0.001	< 0.001	-
joint	0.005	< 0.001	0.83

(c) Null quantification rate[1](global p-value < 0.001; Kruskal-Wallis test)

	icoshift	**speaq**	**indep.**
speaq	< 0.001	-	-
indep.	< 0.001	< 0.001	-
joint	< 0.001	< 0.001	< 0.001

(d) Correlation between simulated and quantified metabolites (global p-value < 0.001; Durbin test)

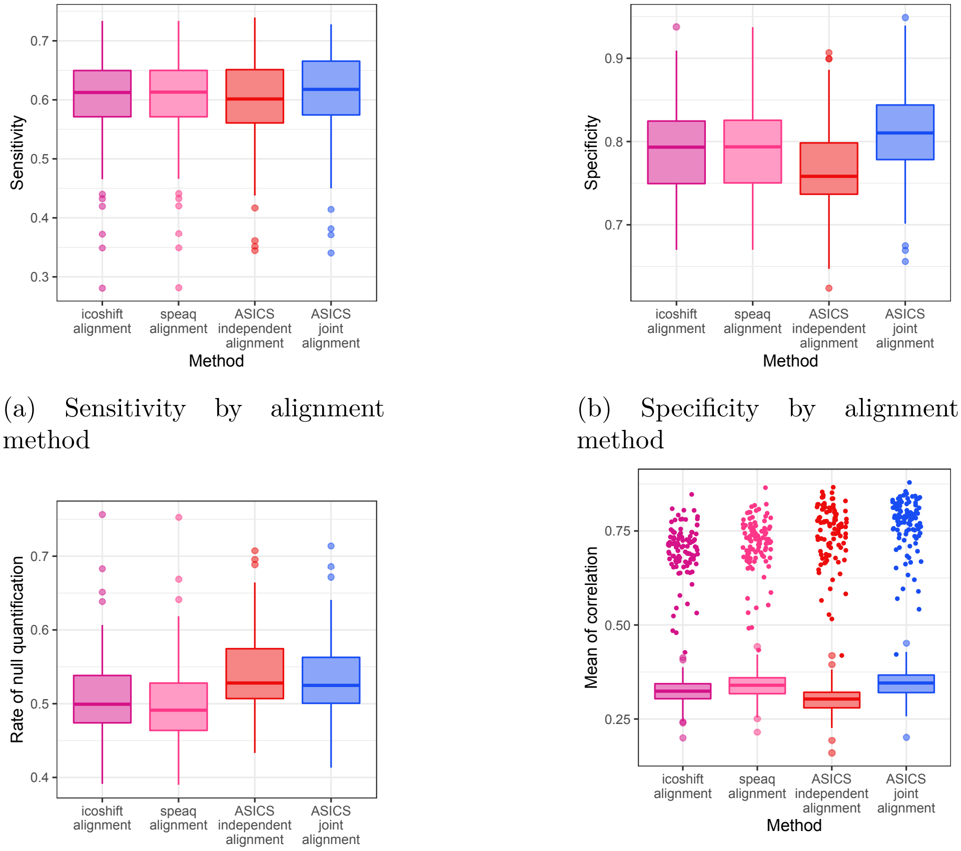
Table 3.2 – p-values of post-hoc Nemenyi tests for sensitivity, specificity and null quantification rate or Durbin tests for correlation for the comparison between each pair of alignment methods (icoshift, **speaq**, **ASICS** independent and joint alignment). **ASICS** independent quantification was performed after library alignment for all methods.

More precisely, from the identification point of view, the results showed that, even if the sensitivity across methods is not significantly different (p-value $= 0.60$; Kruskal-Wallis test), the specificity was improved by **ASICS** joint alignment (p-values < 0.001; Kruskal-Wallis test). Quantifications were also improved by **ASICS** joint alignment (p-values < 0.001; Durbin tests). Median correlations were equal to 0.30 for **ASICS** independent alignment, to 0.32 for icoshift alignment, to 0.34 for speaq alignment and to 0.35 for **ASICS** joint alignment (p-values < 0.001; Durbin post-hoc tests). Again, median correlation of **ASICS** joint alignment increased to 0.79 when considering only the 30 upmost concentrated metabolites (between 50% and 60%

[1] The rate of null quantification is computed on the metabolites identified in at least one complex mixture. It is given by the following formula

$$\text{Rate of null quantification} = 1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} \mathbf{1}_{\{\beta_{ij} > 0\}}}{n \sum_{j=1}^{p} \mathbf{1}_{\left\{\sum_{i=1}^{n} \beta_{ij} > 0\right\}}}$$

(average frequency of unidentification for metabolites that have been identified at least once). In particular, the rate of null quantification is low if the identified (resp. unidentified) metabolites are identified (resp. unidentified) in all complex spectra, *i.e.*, if the identification are consistent accross complex spectra.



(a) Sensitivity by alignment method

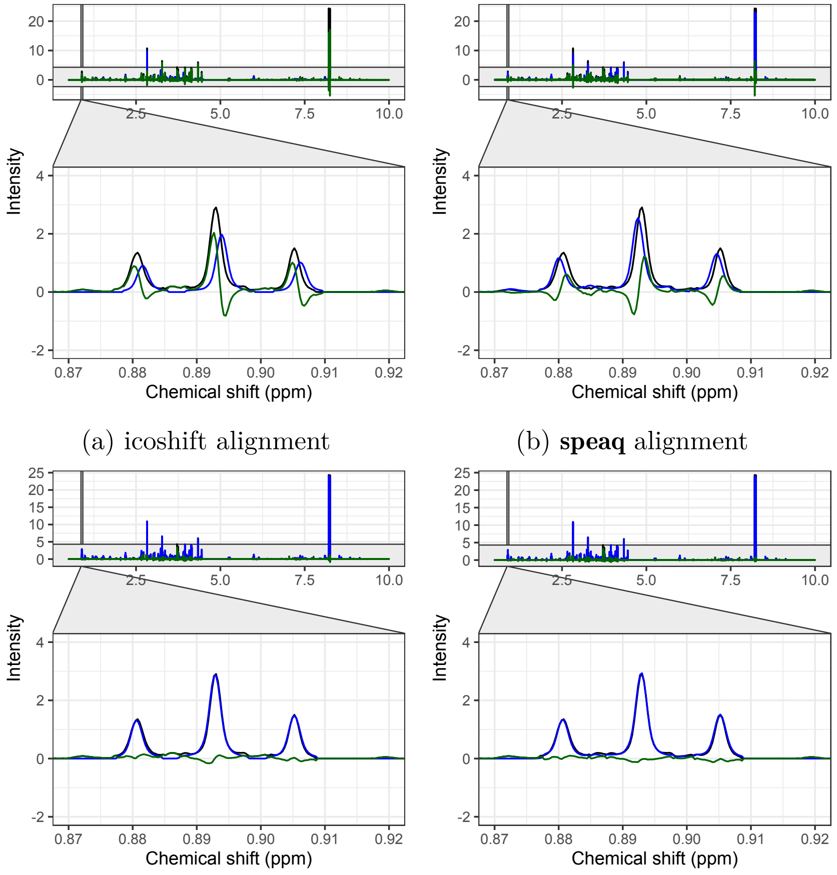
(b) Specificity by alignment method

(c) Null quantification rate by alignment method

(d) Correlation between simulated and quantified metabolites by alignment method

Figure 3.5 – Comparison of alignment methods based on four indicators. Points on Figure 3.5d correspond to the correlation obtained for the 30 most concentrated metabolites. **ASICS** independent quantification was performed after library alignment for all methods.

of estimated quantifications were equal to 0). Figure 3.6 also provides examples of one simulated complex spectrum, its corresponding reconstructed spectrum (after alignment and model fit) and the residual spectrum (the simulated complex spectrum minus its reconstructed spectrum) for different methods. This figure confirms that **ASICS** alignments lead to a better reconstruction of the complex spectrum, with smaller residuals. The difference between **ASICS** joint and independent alignments is not as visible and strong than the difference between **ASICS** alignments and other ones.



(a) icoshift alignment

(b) **speaq** alignment

(c) **ASICS** independent alignment

(d) **ASICS** joint alignment

Figure 3.6 – Comparison of impact of alignment methods on signal reconstruction and focus on one peak (valerate). Black: original simulated complex spectrum. Blue: Reconstructed spectrum after alignment and quantification (performed with **ASICS** independent quantification). Green: Residual spectrum (black spectrum minus blue spectrum).

In addition, the sensitivity of the performances of the independent and the joint alignment procedure to different magnitudes of shifts in the simulated data was also assessed. The results (see Fig. 3.7) show that the joint alignment leads to significantly improved results for the highest shift values (p-value < 0.001 overall; Wilcoxon paired tests).

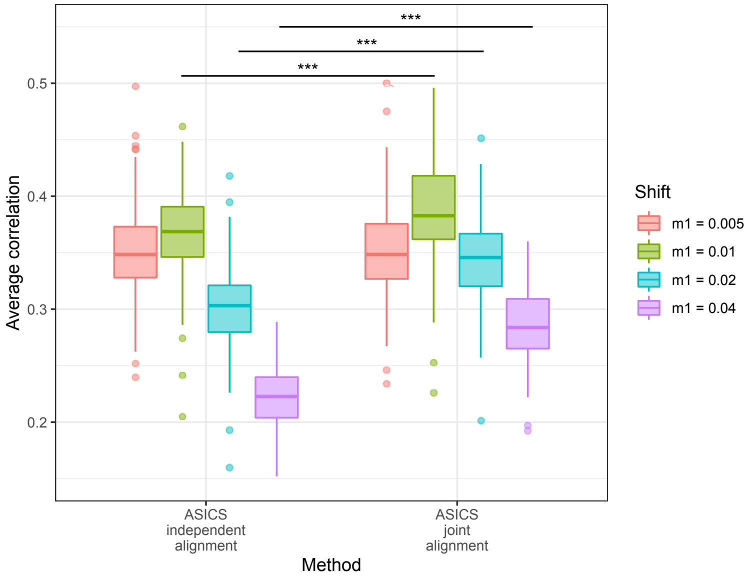


Figure 3.7 – Difference in average correlation over metabolites for 100 datasets for **ASICS** independent and joint alignments using different values of the shift magnitude in simulations (m_1 as in Section 3.3.1). ***: significant differences using paired Wilcoxon tests (p-values < 0.001 overall).

Finally, computational time for the different alignment procedures were obtained on a 24 processor (3.00GHz Intel) 256Go RAM server (with Debian 4 OS): processing of a given dataset (100 complex spectra) took 1 minute for icoshift, 2h45 for **speaq**, 18 minutes for **ASICS** independent alignment and 34 minutes for **ASICS** joint alignment.

The evaluation of the impact of the alignment on the quality of the quantification for the piglet plasma dataset exhibited a similar trend (Figure 3.8 and Table 3.3). Correlations between quantifications and UPLC dosages were found higher with **ASICS** joint alignment (median = 0.61) than with **ASICS** independent alignment (median = 0.44; p-value = 0.08; Durbin post-hoc test), **speaq** alignment (median = 0.21; p-value < 0.001, Durbin post-hoc test) or icoshift alignment (median = 0.32; p-value < 0.001, Durbin post-hoc test).

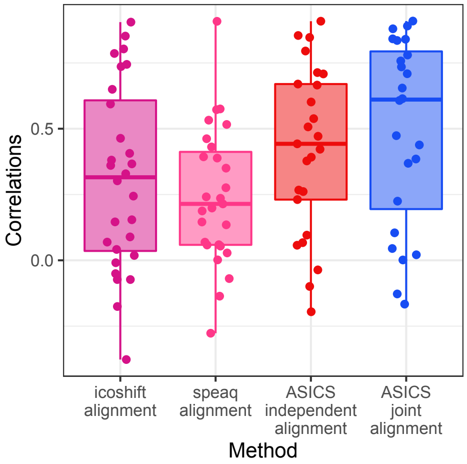


Figure 3.8 – Correlations between quantifications and UPLC dosages using three different alignment methods. **ASICS** independent quantification was performed after library alignment for all methods. Points correspond to every individual correlations.

	icoshift	**speaq**	**independent**
speaq	0.71	-	-
independent	0.003	0.007	-
joint	< 0.001	< 0.001	0.08

Table 3.3 – p-values of Durbin post-hoc tests for correlations between quantifications and UPLC dosages between each pair of alignment methods (global p-value $<$ 0.001; Durbin test). **ASICS** independent quantification was performed after library alignment for all methods.

In particular, **ASICS** joint alignment allows to improve the quality of alignment and subsequent quantification of metabolites for which the pure spectrum has a small number of peaks. For instance, the glycine has a pure spectrum with only one peak. In the complex spectra on Figure 3.9, the actual peak of glycine is at 3.57 ppm. However, with independent alignment, pure spectra of glycine were usually aligned around 3.56 ppm (red spectra). Thus, the correlation between UPLC concentrations and estimated quantifications was equal to 0.07 instead of 0.88 with a joint alignment.

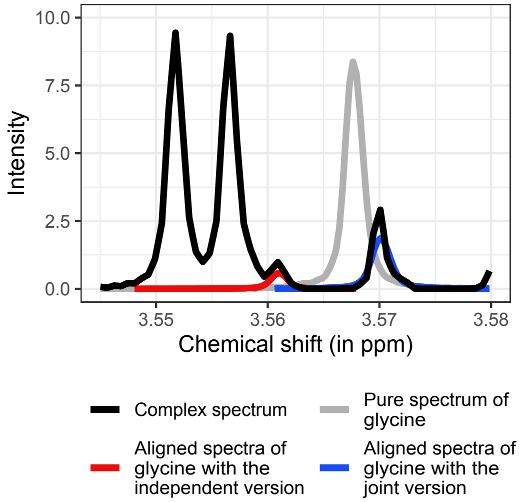
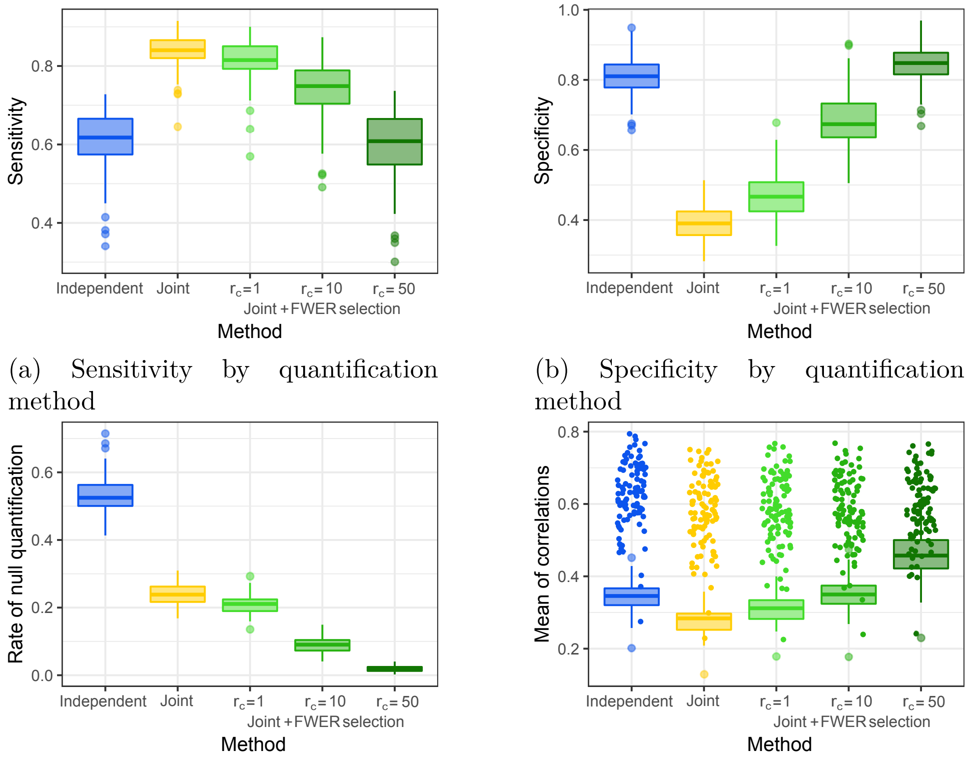


Figure 3.9 – Glycine pure spectrum aligned on every complex mixture spectrum by **ASICS** independent or joint alignments.

3.4.2 Evaluation of **ASICS** joint quantification

On the simulated dataset, the quality of metabolite identification was found to be opposite for sensitivity and specificity. The best method in terms of sensitivity was **ASICS** joint quantification with a single cleaning step and the worst methods were **ASICS** independent quantification and **ASICS** joint quantification with $r_c = 50\%$, both being very stringent on the identified metabolites (Figure 3.10a). On the contrary, these latter two methods were the ones with the best specificity, whereas **ASICS** joint quantification with a single cleaning step achieved the worst specificity (Figure 3.10b).

From the quantification point of view (Figure 3.10d), **ASICS** joint quantification with $r_c = 50\%$ is the method that achieves the best performances (median correlation equal to 0.46, whereas all the others are below 0.4; p-value < 0.001 for each pairwise comparison; Durbin post-hoc tests). When looking at the two methods with the highest specificity (**ASICS** independent quantification and **ASICS** joint quantification with $r_c = 50\%$), the quantification was found better with the joint approach (median correlation equal to 0.35 and 0.46 respectively; p-value < 0.001; Durbin post-hoc test). Indeed, the FWER selection procedure used in **ASICS** independent quantification leads to an under-efficient selection procedure that sets some quantifications



(a) Sensitivity by quantification method

(b) Specificity by quantification method

(c) Null quantification rate by quantification method

(d) Correlation by quantification method

Figure 3.10 – Comparison of quantification methods based on three indicators. Points on Fig. 3.10d correspond to correlations of the 30 most concentrated metabolites.

to 0 when the joint quantification is able to better estimate their small values (Figure 3.10c).

Finally, computational time for the different alignment procedures were obtained on a 24 processor (3.00GHz Intel) 256Go RAM server (with Debian 4 OS): automatic quantification of one dataset (100 complex mixture spectra) took 7 minutes for **ASICS** independent quantification, 7 minutes 30 for **ASICS** independent quantification without the cleaning step and 35 minutes with the cleaning step. When the reference complex mixture spectrum used for the alignment is provided by the user, **ASICS** independent quantification with the cleaning step took 15 minutes.

Correlations between quantifications and UPLC dosages for the piglet plasma dataset are displayed in Figure 3.11 for the different methods. Overall, it shows that **ASICS** joint quantification with $r_c = 50\%$ again performs the best on this dataset. In particular, **ASICS** joint quantification gives results significantly better than **rDolphin** (median correlations equal to 0.87 and 0.75, respectively; p-value < 0.001; Durbin post-hoc test). **rDolphin** performed worse despite the fact that the method was given the metabolites of interest in contrast to **ASICS** that performs its own metabolite identification.

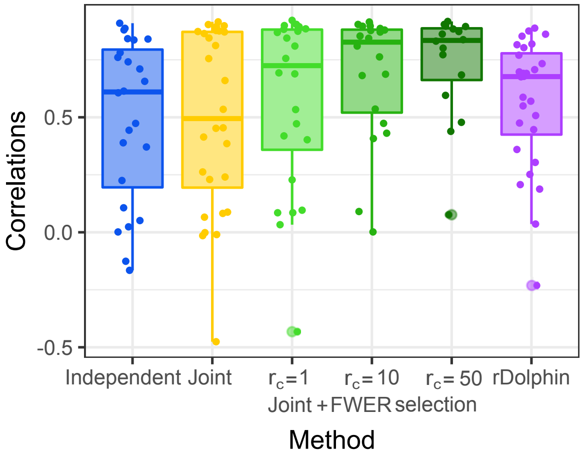


Figure 3.11 – Correlation between quantifications and UPLC dosages by quantification method. Points correspond to all correlations.

In this dataset, amino acid dosages allow to explore a wide variety of concentration values from very concentrated metabolites (alanine or glycine, more than 500μmol/L on average) to very lowly concentrated metabolites (methionine or ornithine, less than 50μmol/L on average). **ASICS** joint quantification allows to address one of the limits of **ASICS** independent quantification described in Lefort et al. (2019), where quantifications of lowly concentrated metabolites were found of poorer quality. Here, the median correlation of lowly concentrated metabolites ($< 100\mu$mol/L) was improved by the joint approach with $r_c = 50\%$: median correlations were equal to 0.77 versus 0.50 (p-value < 0.001; Durbin post-hoc test) for the same two methods (see also examples on the serine and the methionine in Figure 3.12).

In addition to these two examples, **ASICS** joint quantification also allows to more accurately quantify other types of metabolites that were not identified or were identified only in a few spectra with the FWER selection of **ASICS** independent quantification.

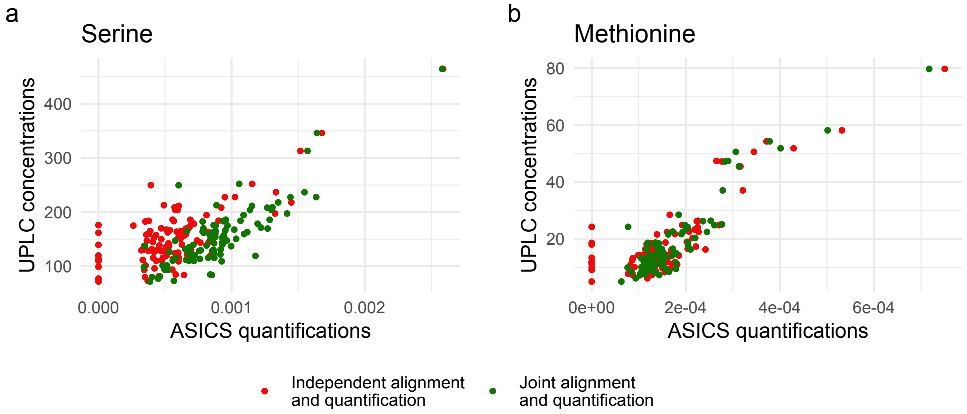


Figure 3.12 – Correlation between quantification and UPLC dosages for (a) serine (155μmol/L on average) and (b) methionine (18μmol/L on average) with independent quantification (red) or joint quantification with $r_c = 50\%$ (dark green).

Another case where **ASICS** joint quantification with $r_c = 10\%$ provides better results than **ASICS** independent quantification is the case where the pure spectrum of a metabolite has several peaks close to the noise level due to a large number of peaks in this spectrum. This is the case of the lysine, for instance, which has a correlation equal to 0.88 with **ASICS** joint quantification ($r_c = 10\%$) and to 0.42 with **ASICS** independent quantification.

3.5 Conclusion

To the best of our knowledge, **ASICS** joint alignment and quantification approaches are the only automatic approaches that allow to account for multiple samples for automatic identification quantification of metabolites in complex mixture spectra. Both joint steps lead to improved quantification accuracy and a better identification of metabolites present in the complex mixture. In particular, the joint approaches are efficient to help identify metabolites with low concentrations, which are hard to distinguish from the noise level. This is true even when using the joint approaches in combination with stringent pre-filtering steps ($r_c = 50\%$), which are necessary to control the number of false identifications. Finally, with the flexibility offered by the setting of a less stringent pre-filtering step ($r_c = 1\%$ or $r_c = 10\%$), the user can also quantify very lowly concentrated targeted metabolites that are known to be present in the complex mixture. Overall, the joint approaches allow to leverage the initial weakness of **ASICS** independent quantification as well as those of most automatic identification methods on the poor identification and quantification of lowly concentrated metabolites. Joint approaches can result in an increased computational time, especially for the quantification, but the computational time still remains acceptable (less than one hour for $\sim$ 100 complex spectra) and can result in a strong improvement of the signal reconstruction and of the quantification, especially when complex spectra were acquired with large shifts in the peak positions compared to the reference library.

Acknowledgements

The authors are grateful to the INRAE metaprogram fund "GISA" (Integrated management of animal health) for the funding of the SuBPig project (Enhancing survival at birth). The PhD fellowship of Gaelle Lefort is supported by the Digital Agriculture Convergence Lab (#DigitAg, http://www.hdigitag.fr/, ANR-16-CONV-0004), by INRAE Mathematics, Computer and Data Sciences, Digital Technologies Division, by INRAE Animal Genetics Division and by INRAE Animal Health Divison. The funders had no role in the study design, analyses, results, interpretation and decision to publish.

The authors are very grateful to the staff of experimental pig facilities (INRAE, 2018. Pig phenotyping and Innovative breeding facility, doi:10.15454/1.5572415481185847E12). They also thank all the participants from GenPhySE, PEGASE and GenESI laboratories (INRAE) and MetaToul-AXIOM plateform for sample and data collection especially Nadine Mézière and Colette Mustière (INRAE PEGASE) for performing all UPLC dosages, Hélène Quesnel (INRAE PEGASE) for useful discussions and comments and, Roselyne Gautier (MetaToul-AXIOM plateform) for her help in the acquisition of NMR spectra.

The authors are grateful to the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, doi:10.15454/1.5572369328961167E12) for providing computing resources.

The authors also thank Isabelle Pinard for English proofreading and correction.

Chapter 4

The maturity in fetal pigs using a multi-fluid metabolomic approach

Scientific production

The content of this chapter corresponds to the article "G. Lefort, R. Servien, H. Quesnel, Y. Billon, L. Canario, N. Iannuccelli, C. Canlet, A. Paris, N. Vialaneix, L. Liaubet (2020). The maturity in fetal pigs using a multi-fluid metabolomic approach. *Scientific Reports*, **10**, 19912".

Abstract

In mammalian species, the first days after birth are an important period for survival and the mortality rate is high before weaning. In pigs, perinatal deaths average 20% of the litter, with important economic and societal consequences. Maturity is one of the most important factors that influence piglet survival at birth. Maturity can be defined as the outcome of complex mechanisms of intra-uterine development and maturation during the last month of gestation. Here, we provide new insights into maturity obtained by studying the end of gestation at two different stages (three weeks before term and close to term) in two breeds of pigs that strongly differ in terms of neonatal survival. We used metabolomics to characterize the phenotype, to identify biomarkers, and provide a comprehensive understanding of the metabolome of the fetuses in late gestation in three fluids (plasma, urine, and amniotic fluid). Our results show that the biological processes related to amino acid and carbohydrate metabolisms are critical for piglet maturity. We confirm

the involvement of some previously described metabolites associated with delayed growth (*e.g.*, proline and myo-inositol). Altogether, our study proposes new routes for improved characterization of piglet maturity at birth.

4.1 Introduction

In mammalian species, the first days after birth are an important period for survival and the mortality rate is high before weaning. In humans, despite the important reduction in the mortality rate in the recent years, neonatal deaths (before one month after birth) still represent 47% of deaths before the age of five, *i.e.* about 2.5 million per year (UNICEF, WHO, World Bank Group and United Nations, 2019). In a polytocous species like swine, this rate averages 20% of the litter in commercial lines (Edwards and Baxter, 2015), and the most critical period for piglet survival is the perinatal period that includes birth and the first 24 hours of postnatal life. Many factors have been shown to influence piglet survival at and after birth (Edwards and Baxter, 2015). They have been related to maternal effects (*e.g.*, intrauterine effects, farrowing duration, parity, health status), to piglet factors (*e.g.*, genetic type, vitality at birth), to piglet characteristics that are partly under maternal control (*e.g.*, birth weight) or to a combination of these factors. Consistently, the most important factors identified for postnatal survival are birth weight, hypothermia, the latency to first suckle, or their combinations, predisposing piglets to starvation or crushing (Baxter et al., 2008; Edwards and Baxter, 2015). Piglet maturity is also likely to be an important determinant of subsequent survival and postnatal growth (Leenhouwers et al., 2002; Basso and Wilcox, 2010). Maturity at birth, which can be defined as complete development enabling survival at birth, is the outcome of complex intra-uterine development and maturation mechanisms that occur during the end of gestation (Leenhouwers et al., 2002). In pigs, the maturation period is considered to be the last month of gestation (approximately 90-114 days of gestation, dg). Together with environmental conditions, physiological maturity at birth thus has major consequences for neonatal mortality highlighting the need for a deeper understanding of maturity to effectively reduce perinatal mortality.

In this context, the aim of the present study was to provide a comprehensive description of the metabolome of pig fetuses in late gestation. Metabolomics is a promising approach to investigate health and welfare in large cohorts, for phenotype characterization and for the identification on usable biomarkers: high-throughput metabolome measurements are easy to obtain and at affordable cost by ^{1}H Nuclear Magnetic Resonance (NMR) and the metabolome enables the comprehensive characterization of the small molecules involved in metabolic chemical reactions. To this end, we compared the metabolomes of plasma, urine, and amniotic fluid, in 611 fetuses in two breeds of pigs, Large White (LW) and Meishan (MS), at two stages of late gestation (90 dg and 110 dg). The three fluids were chosen to represent different aspects of fetus metabolism: plasma reflects the regulation of fetal metabolism, the urine metabolome its excretory renal function, and amniotic fluid its nutritional function and mechanical protection as well as interactions with maternal and placental tissues. The two stages of gestation are representative of the maturation process in late gestation, 90 dg being the onset of fetal maturation in both breeds and 110 dg being close to term (Foxcroft et al., 2006). The two breeds of pigs were chosen because they strongly differ in terms of neonatal survival and can thus be used to identify differences that are possibly responsible for perinatal survival. The LW breed represents European breeds and has been genetically selected for lean growth and prolificacy. Its high rate of perinatal mortality is partly due to lower physiological maturity at birth (Canario et al., 2007). On the contrary, the MS breed presents a low rate of mortality and is considered to be more mature at birth (Herpin et al., 1993; Canario et al., 2006). LW and MS sows were inseminated with mixed (LW and MS) semen so that pure and crossed fetuses would grow in the same uterine environment. The reciprocal crossed fetuses were expected to present an intermediate degree of maturity between LW and MS fetuses. To a lesser extent, these reciprocal crossed fetuses also allowed us to observe maternal or paternal effects or heterosis.

This study, and the search for differences between the two breeds and between the two stages of gestation, completes our previous transcriptomic and proteomic studies (Voillet et al., 2014, 2018; Gondret et al., 2018; Yao et al., 2017) that were performed on muscle, intestinal and adipose tissues using the same experimental design. The present study also completes the blood parameters known to be associated with piglet maturity at birth (*e.g.*, albumin and IGF-I plasma concentrations, Canario et al. (2007)). The potential and functional new biomarkers reported here can be used for genetic selection or to improve management of sow feeding in late gestation. Even if the fatty acid metabolism could not be investigated in our study (due to technical limitations regarding lipid quantification), the study allowed us to confirm some previously described metabolites associated with delayed growth and to identify important biological processes involved in piglet maturity.

4.2 Methods

4.2.1 Ethics Statement

All the fluids from pig fetuses were obtained in the framework of the PORCINET project (ANR-09-GENM-005-01, 20102015). The experiment authorization number of the experimental farm GenESI (Pig phenotyping and Innovative breeding facility, doi:10.15454/1.5572415481185847E12) is A-17-661. The procedures and the animal management complied with European Union legislation (Directive 2010/63/EU) and the French legislation in the Midi-Pyrénées Region (Decree 2001-464). All experiments were performed in accordance with relevant guidelines and regulations and were approved by the ethical committee of the Midi-Pyrénées Regional Council (authorization MP/01/01/01/11).

4.2.2 Animals and plasma, urine and amniotic fluid sampling

Plasma, urine, and amniotic fluid samples were obtained from 611 pig fetuses at two gestational stages (90 and 110 days, average gestation term 114 days). MS and LW sows were inseminated with mixed semen (LW and MS) so that most litters were composed of purebred fetuses (LW or MS) and crossbred fetuses (LW×MS from MS sows and MS×LW from LW sows). MS and LW breeds were chosen as two extreme breeds for piglet mortality at birth, a better survival rate being observed in MS piglets. The experimental design is described in detail in Voillet et al. (2014) and is summarized in Figure 4.1.

A total of 329 fetuses had a LW mother and 282 fetuses had a MS mother. The fetuses were obtained by caesarean section. Fetuses were weighed (statistics on weights are provided in Table 4.1), and on average, LW weighed more than MS despite their lower maturity.

Table 4.1 – Fetus weights at 90 and 110 days of gestation according to genotype (mean ± standard deviation in grams).

Genotype	**90 days of gestation**	**110 days of gestation**
LW	619 ± 141	1171 ± 323
MS×LW	633 ± 91	1292 ± 197
LW×MS	579 ± 114	1092 ± 201
MS	490 ± 86	910 ± 101

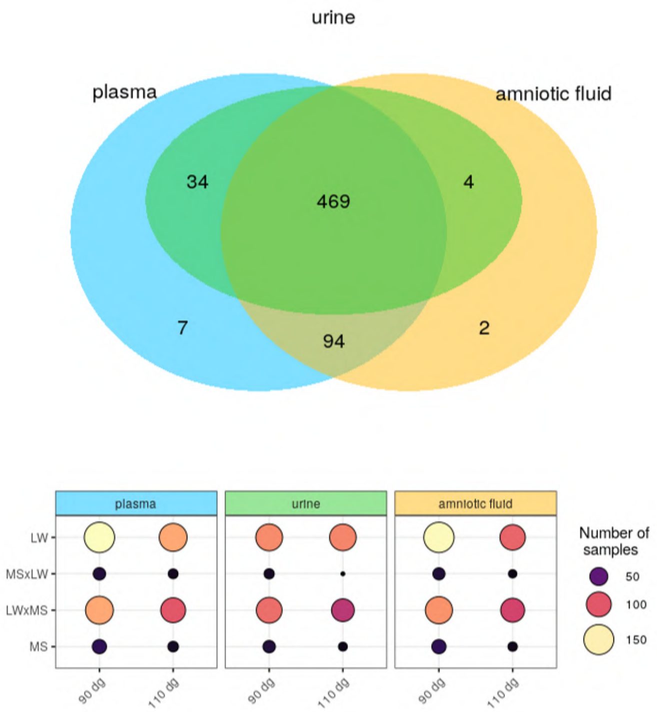


Figure 4.1 – Summary of the design experiment: number of samples by fluid, stage of gestation and genotype.

After laparotomy of the sow, blood (approximately 5 mL) was collected individually from the umbilical artery of the piglets using a 21-gauge needle and a 5 mL syringe and placed in heparinized tubes. After section of the umbilical cord, the fetus was euthanized (Yao et al., 2017). Plasma was prepared by low-speed centrifugation (2,000 g for 10 min at 4°C) and stored at −80°C until further analysis. The amniotic fluid (10 mL) was collected during the caesarean and immediately centrifuged (2,000 g for 10 min at 4°C) to discard cell debris and stored at −20°C until further analysis. The urine samples were collected directly in the bladder with a 5 mL syringe during dissection of the fetuses, immediately frozen to avoid contamination and stored at −80°C until further analysis.

4.2.3 Nuclear magnetic resonance

The detailed protocol for sample preparation, spectra acquisition and preprocessing can be found in Lefort et al. (2019). Briefly, each sample of plasma and amniotic fluid (200 μL) was diluted in 500 μL deuterated water (D_2O) and centrifuged without the addition of internal standard to improve spectra quality. For urine, 200 μL of phosphate buffer prepared in deuterated water (0.2 M, pH 7.0) were added to 500 μL of urine, vortexed, centrifuged at 5000 g for 15 min, and 600 μL transferred to 5 mm NMR tube. All ^{1}H NMR spectra were acquired on a Bruker Avance DRX-600 spectrometer (Bruker SA, Wissembourg, France) operating at 600.13 MHz for ^{1}H resonance frequency and at 300K using the Carr-Purcell-Meiboom-Gill (CPMG) spin-echo pulse sequence. The Fourier transformation was calculated on 64,000 points. All spectra were phased, baseline corrected and then calibrated on the resonance of lactate (1.33 ppm) using Topspin (V2.1, Bruker, Biospin, Munich, Germany). The regions corresponding to water resonance (5.1–4.5 ppm) and urine (6.5–6.0 ppm) were excluded to eliminate artifacts of residual water and urine.

4.2.4 Metabolite identification and quantification

To measure the concentration of metabolites in the three fluids, NMR metabolomic spectra were processed with **ASICS**, a recently developed R package (Lefort et al., 2019). Before quantification, spectra were normalized by the area under the curve and aligned with preprocessing functions available in the Bioconductor R package **ASICS** (version 2.0.0, Lefort et al. (2019)). The metabolites in all fluids were identified and quantified using the **ASICS** method available in the same package. Quantification was performed using the default reference library provided in the package and was processed independently for each fluid but the same maximum chemical shift, set at 0.02, was allowed for each. The library alignment was improved (compared to that described in Lefort et al. (2019)) by using a global quality control criterion: the correlation between quantifications and targeted buckets of the spectra was maximized to choose the best alignment between peaks. Finally, metabolites that had at least 25% of quantifications larger than 0 in at least one condition (stages of gestation and genotypes) were retained. The others were removed from the list of identified metabolites (quantification set to 0).

Note that the **ASICS** quantification method is threshold-based, meaning the estimation of the quantification is exactly 0 for some metabolites. In all cases, it means that the peaks corresponding to these metabolites are below the noise level in the corresponding complex mixture spectra. But, as a consequence, a threshold effect is visible in some of the boxplots of Figures 4.8 and 4.9, where some quantification distributions are represented by a flat horizontal line centered on zero. However, the effect of such a threshold is negligible because, in all cases, it means that the real concentration, if not really zero, is so low that it can not be distinguished from noise in the complex spectra.

4.2.5 Spectra quality control

Plasma relative quantifications were previously validated in Lefort et al. (2019) using biochemical targeted dosages of three metabolites (glucose, fructose and lactate) in a subset of the samples. The results of a an Orthogonal Projections to Latent Structures Discriminant Analysis (Trygg and Wold, 2002) (OPLS-DA) based on the standard bucket approach were compared with the results of an OPLS-DA based on metabolite quantifications. The comparison showed good reproducibility and similar discriminative power between conditions for the bucket and quantification approaches, insuring minimum loss of information during quantification preprocessing.

Principal Component Analyses (PCA) was used to detect potential outliers and batch effects due to experimental covariates: sex, breeding batch, and sow. All plots are shown in Figures 4.2 à 4.4. PCA did not identify any sex, batch or experimental effect but a sow effect was clearly visible and, whenever possible was included in the subsequent analyses.

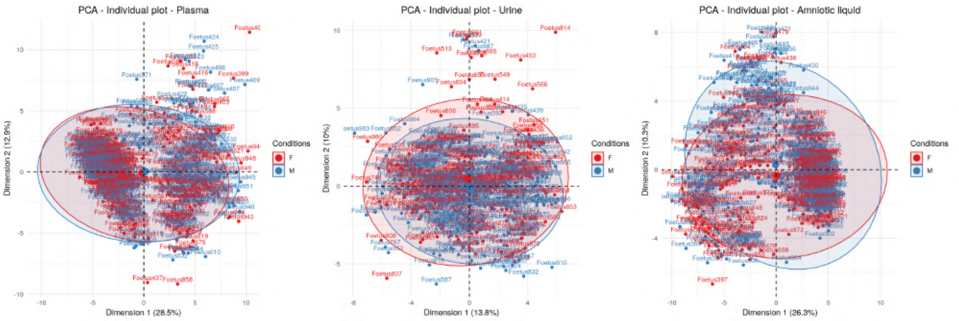


Figure 4.2 – PCA on quantifications of metabolites colored by gender (red: females, blue: males).

4.2.6 Multivariate exploratory analyses

All statistical analyses were performed with R (version 3.6.0, R Core Team (2019)). The effect on the metabolome of the stage of gestation (*i.e.*, 90 dg and 110 dg) was first investigated with OPLS-DA (Trygg and Wold, 2002).

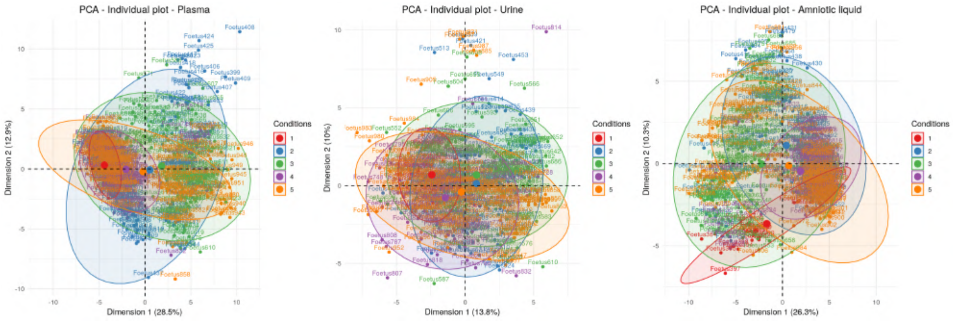


Figure 4.3 – PCA on quantifications of metabolites colored by batch effect.

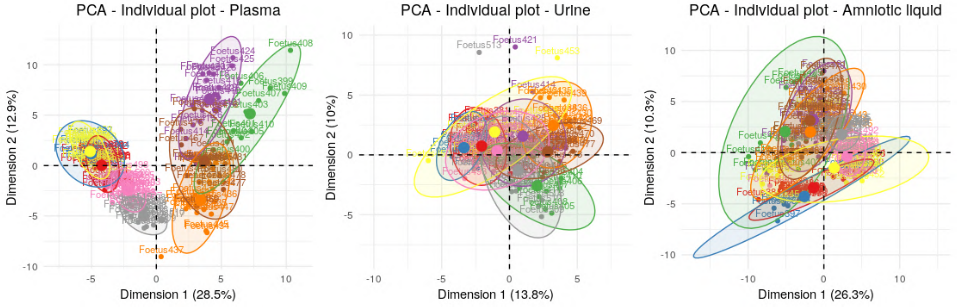


Figure 4.4 – PCA on quantifications of metabolites colored by sow. For the sake of clarity, only the first nine sows are represented.

Three OPLS-DA were performed independently on each fluid to identify the metabolites with the highest discriminant power between the two gestational stages: the most influential metabolites, *i.e.*, metabolites with a VIP index greater than 1, were extracted. The relevance of the results was insured by estimating the predictive power of each model with a 10-fold cross-validation error.

4.2.7 Univariate differential analyses with mixed models

As OPLS-DA was limited to the study of one factor with only two levels, we completed it with more complete analyses based on mixed models that can incorporate multiple effects, including the random effect of the sow used as a proxy for the effect of the uterine environment. This effect must not be mistaken with a parental effect originating from the genotype. Mixed models were used to identify metabolites with differential concentrations between conditions (gestational stages and genotypes), by fitting the following model for each fluid and each metabolite:

$$b_{ijk} = \mu + A_i + FG_j + I_{ij} + S_k + \epsilon_{ijk} \tag{4.1}$$

with b_{ijk} the vector of metabolite concentrations for gestational stage i ($i \in \{\text{d90}, \text{d110}\}$), genotype j ($j \in \{\text{LW}, \text{LW} \times \text{MS}, \text{MS} \times \text{LW}, \text{MS}\}$) and mother (sow) k. In this model, μ is the mean effect, A_i the fixed effect of the gestational stage, FG_j is the fixed effect of the genotype, I_{ij} is the effect of the interaction between the gestational stage and the genotype, $S_k \sim \mathcal{N}(0, \sigma_r^2)$ is the random effect of the sow and $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma_e^2)$ is a noise term.

For all metabolites, this model was tested against the model with only the sow effect ($b_{ijk} = \mu + S_k + \epsilon_{ijk}$) with a Fisher's test. p-values were then adjusted with the Benjamini and Hochberg (FDR) correction (Benjamini and Hochberg, 1995). Finally, using the same methodology as Voillet et al. (2014), each differential metabolite (*i.e.*, the metabolites with an adjusted p-value of less than 0.05) was associated with one of the following sub-models:

- complete: $b_{ijk} = \mu + A_i + FG_j + I_{ij} + S_k + \epsilon_{ijk}$
- additive: $b_{ijk} = \mu + A_i + FG_j + S_k + \epsilon_{ijk}$
- only stage: $b_{ijk} = \mu + A_i + S_k + \epsilon_{ijk}$
- only genotype: $b_{ijk} = \mu + FG_j + S_k + \epsilon_{ijk}$

In contrast to the approach that would have consisted in independently testing each effect of the complete model (stage effect, genotype effect and interaction), selecting the best fit sub-model avoids overfitting and produces the best set of relevant effects for each metabolite. Since the four models described above are not nested (the “only stage” and “only genotype” models are not), this selection cannot be performed using a standard Fisher's test so we used a model selection approach instead and selected the model with the minimum Bayesian Information Criterion (BIC, Schwarz (1978)).

4.2.8 Pathway enrichment

Pathway enrichment analysis was performed with the web-based tool suite MetaboAnalyst (version 4.0, Chong et al. (2018)) with the MetPA module (Xia and Wishart, 2010). *Sus scrofa* pathways were not available in MetaboAnalyst, so the *Homo sapiens* KEGG pathways were used, as a reference instead. We also checked the differences between human and pig pathways on the KEGG database and the two pathways were found to be almost identical. This confirmed the relevance of using the human pathways in MetaboAnalyst. Finally, hypergeometric tests were performed to extract pathways enriched in influential or differential metabolites and p-values were corrected for multiple testing using the Benjamini and Hochberg approach. This analysis was carried out for each fluid and on OPLS-DA and mixed model results independently.

4.3 Results

A ^{1}H NMR metabolomic analysis was performed on plasma, urine, and amniotic fluid collected from 611 fetuses at 90 and 110 days of gestation. Metabolic quantification was performed automatically with the R package **ASICS**. Among the 190 available metabolites in the reference library of the **ASICS** package, about 65 metabolites were identified in each fluid (*i.e.*, 63 in plasma, 64 in urine and 68 in amniotic fluid; Figure 4.5). Thirty-nine metabolites were identified in all three fluids including many amino acids (*e.g.*, glutamine, glycine, proline, and arginine) and many sugars (*e.g.*, glucose, fructose, glucose-6-phosphate). Other metabolites were identified in only one or two fluids, *e.g.*, leucine and isoleucine, identified only in plasma and urine, or reduced or oxidated glutathione, identified only in urine and amniotic fluid.

4.3.1 Multivariate exploratory analyses

Three Orthogonal Projections to Latent Structures Discriminant Analyses (OPLS-DA, Trygg and Wold (2002)), one for each fluid, discriminated the two stages of gestation with good accuracy (Figure 4.6), especially in the plasma where the cross-validation error was 1%. For urine and amniotic fluid, the error was slightly higher (4%) but still low, indicating a slightly less clear separation between the two groups (90 dg and 110 dg). Altogether, these results suggest that OPLS-DA can be interpreted with a high level of confidence.

Around 20 metabolites were found to be influential (Variable Influence on Projection, VIP > 1) in each fluid (23 in plasma, 21 in urine and 22 in amniotic fluid) and were consequently used for enrichment analyses of the pathways. The analyses were performed for each fluid separately, results are presented in Table 4.2 (page 92). Only three influential metabolites were common to the three fluids (glucose-6-phosphate, fructose and guanidinoacetate; Figure 4.7). Guanidinoacetate and fructose were more concentrated

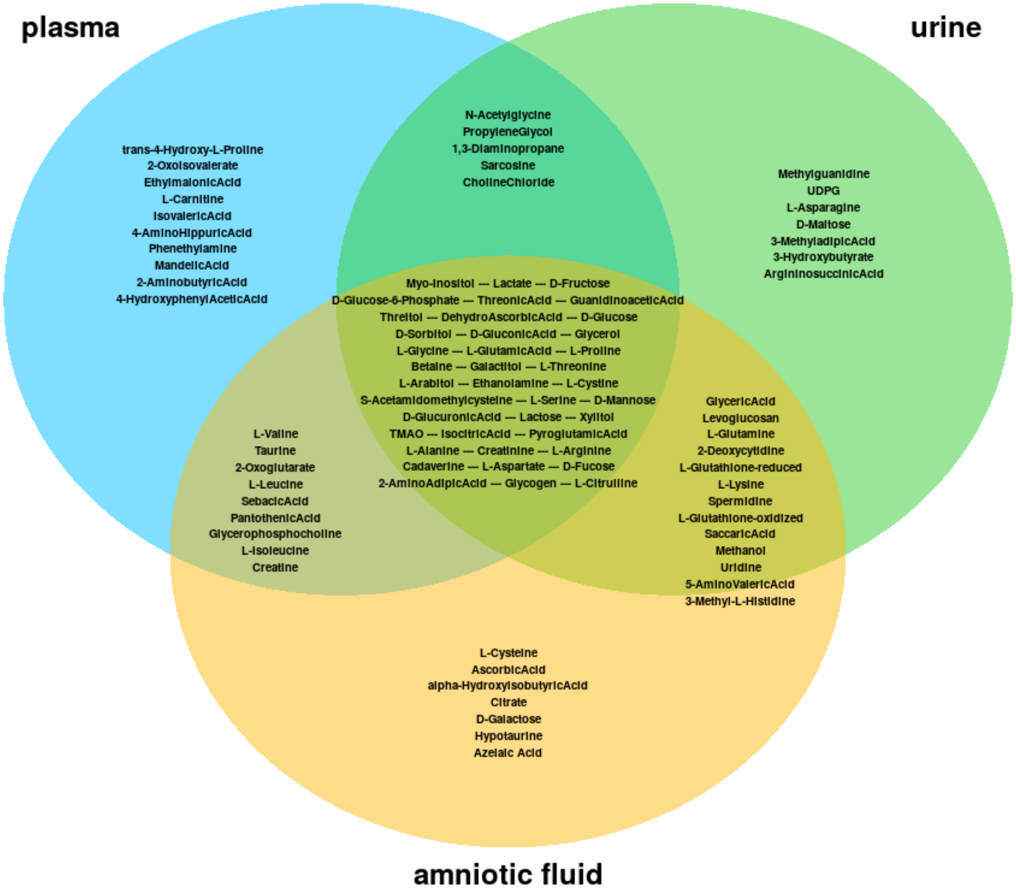
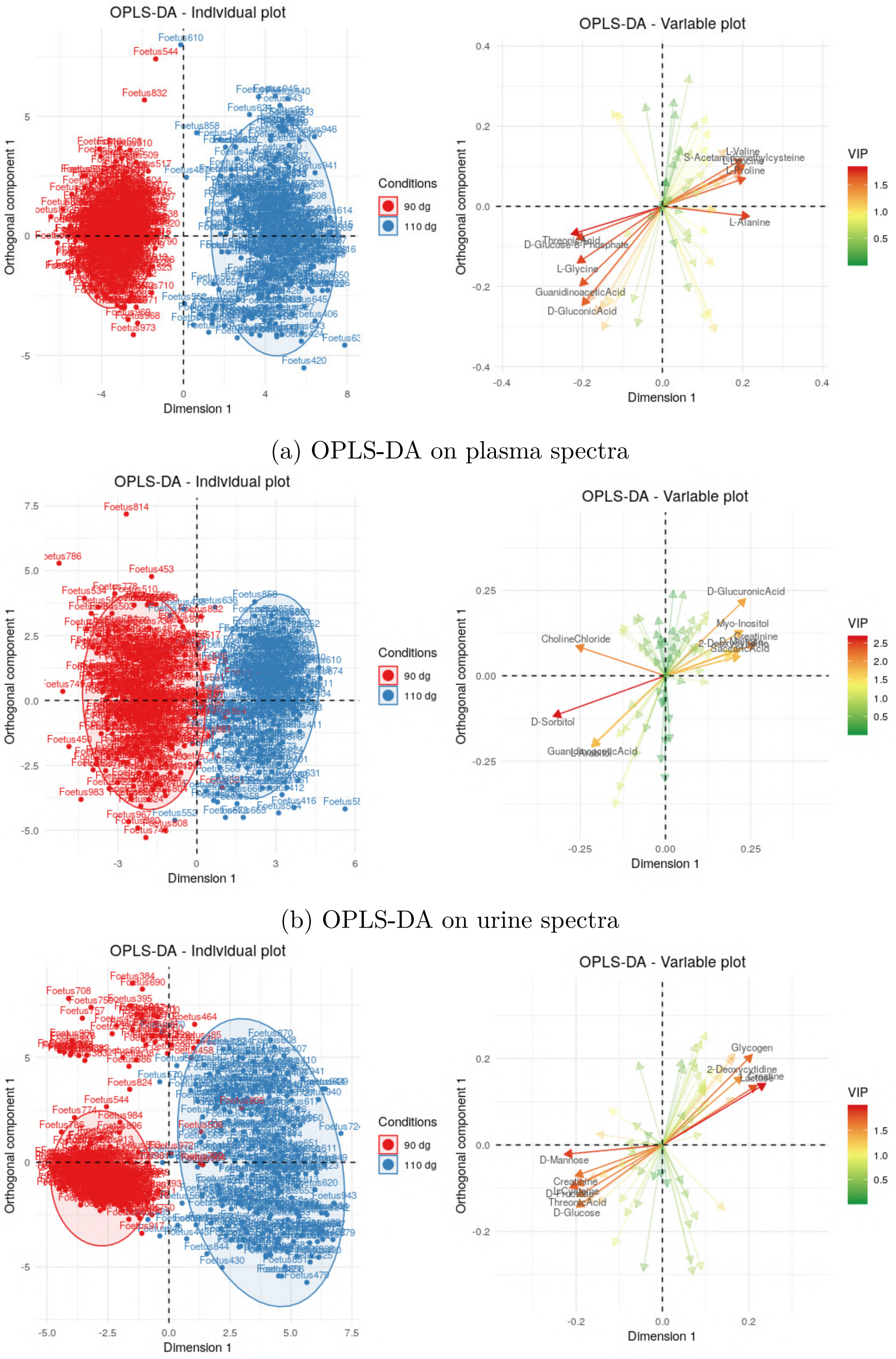


Figure 4.5 – Metabolites identified with **ASICS** package in each fluid.

at 90 dg than at 110 dg in all three fluids while glucose-6-phosphate was more concentrated at 90 dg in plasma and amniotic fluid but more concentrated at 110 dg in urine. Glucose-6-phosphate is involved in galactose metabolism, a pathway that was found to be enriched in influential metabolites in all three fluids and was also part of the pentose phosphate pathway that was enriched in influential metabolites in plasma. Along with fructose, glucose-6-phosphate is also involved in starch (*i.e.*, glycogen in animals) and sucrose metabolism. This pathway was enriched in influential metabolites in both urine and amniotic fluid. However, the concentrations of all influential metabolites of this pathway varied differently in the two fluids between the two gestational stages. In urine, most of these metabolites were present



(a) OPLS-DA on plasma spectra

(b) OPLS-DA on urine spectra

(c) OPLS-DA on amniotic fluid spectra

Figure 4.6 – Individual and variable plots for the first two axes of the Orthogonal Projections to Latent Structures Discriminant Analyses (OPLS-DA) on n = 611 fetuses. Figures were obtained using the quantifications from a plasma, b urine, c amniotic fluid spectra for both days of gestation (90 dg and 110 dg) and all genotypes (LW, MS and cross fetuses together). VIP: Variable Influence on Projection.

at higher concentrations at 110 dg (glucose-6-phosphate, glucoronate and maltose), whereas only glycogen was present at a higher concentration in amniotic fluid at 110 dg.

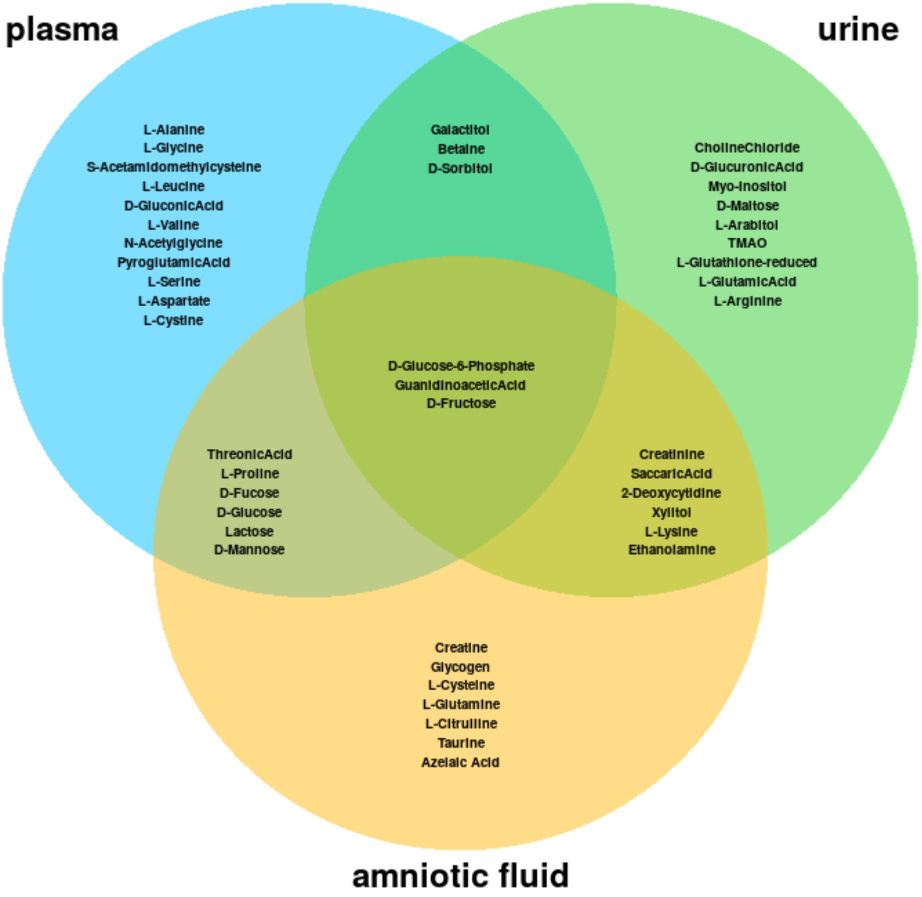


Figure 4.7 – Influential metabolites detected by OPLS-DA.

Table 4.2 – Enriched pathways in influential metabolites for OPLS-DA and in differential metabolites for mixed models. $^{<method>\_\{<fluids>\}}$ influential and/or differential metabolites for <method> (^{opls} for OPLS-DA and ^{mm} for mixed models) in <fluids> (p for plasma, u for urine and af for amniotic fluid). * Total number of metabolites in the pathway.

Pathway	**Method**	**Total metab.***	**Influential and/or differential metabolites**
Alanine, aspartate and glutamate metabolism	Mixed models^{p,u,af}	24	2-oxoglutarate$^{mm\_\{p,af\}}$, Alanine$^{mm\_\{p,u,af\}}$, Argininosuccinate$^{mm\_\{u\}}$, Asparagine$^{mm\_\{u\}}$, Aspartate$^{mm\_\{p,af\}}$, Glutamate$^{mm\_\{p,u\}}$, Glutamine$^{mm\_\{u,af\}}$,
Aminoacyl-tRNA biosynthesis	OPLS-DA^{p} and mixed models^{p,u,af}	75	Alanine$^{opls\_\{p\};\ mm\_\{p,u,af\}}$, Arginine$^{mm\_\{p,u,af\}}$, Asparagine$^{mm\_\{u\}}$, Aspartate$^{opls\_\{p\};\ mm\_\{p,af\}}$, Cysteine$^{mm\_\{af\}}$, Glutamate$^{mm\_\{p,u\}}$, Glutamine$^{mm\_\{u,af\}}$, Glycine$^{opls\_\{p\};\ mm\_\{p,u\}}$, Isoleucine$^{mm\_\{p\}}$, Leucine$^{opls\_\{p\};\ mm\_\{p\}}$, Lysine$^{mm\_\{u,af\}}$, Proline$^{opls\_\{p\};\ mm\_\{p,u,af\}}$, Serine$^{opls\_\{p\};\ mm\_\{p,af\}}$, Threonine$^{mm\_\{p,u,af\}}$, Valine$^{opls\_\{p\};\ mm\_\{p,af\}}$
Arginine and proline metabolism	OPLS-DA^{af} and mixed models^{p,u,af}	77	5-Aminopentanoate$^{mm\_\{u,af\}}$, Arginine$^{mm\_\{p,u,af\}}$, Arginosuccinate$^{mm\_\{u\}}$, Aspartate$^{mm\_\{p,af\}}$, Citrulline$^{opls\_\{af\};\ mm\_\{p,af\}}$, Creatine$^{opls\_\{af\};\ mm\_\{p,af\}}$, Creatinine$^{opls\_\{af\};\ mm\_\{p,u,af\}}$, Glutamate$^{mm\_\{p,u\}}$, Glutamine$^{opls\_\{af\};\ mm\_\{u,af\}}$, Guanidinoacetate$^{opls\_\{af\};\ mm\_\{p,u,af\}}$, Hydroxyproline$^{mm\_\{p\}}$, Proline$^{opls\_\{af\};\ mm\_\{p,u,af\}}$, Pyroglutamate$^{mm\_\{u\}}$, Sarcosine$^{mm\_\{p,u\}}$ Spermidine$^{mm\_\{u,af\}}$
Ascorbate and aldarate metabolism	Mixed models^{af}	45	2-oxoglutarate$^{mm\_\{af\}}$, Ascorbate$^{mm\_\{af\}}$, Glucuronate$^{mm\_\{af\}}$, Myo-inositol$^{mm\_\{af\}}$, Saccaric acid$^{mm\_\{af\}}$, Threonate$^{mm\_\{af\}}$
Cyanoamino acid metabolism	OPLS-DA^{p}	16	Aspartate$^{opls\_\{p\}}$, Glycine$^{opls\_\{p\}}$, Serine$^{opls\_\{p\}}$
Cysteine and methionine metabolism	OPLS-DA^{p}	56	Alanine$^{opls\_\{p\}}$, Aspartate$^{opls\_\{p\}}$, Serine$^{opls\_\{p\}}$
Galactose metabolism	OPLS-DA^{p,u,af} and mixed models^{p,u,af}	41	Galactitol$^{opls\_\{p,u\};\ mm\_\{p,u,af\}}$, Glucose$^{opls\_\{p,af\};\ mm\_\{p,af\}}$, Glucose-6-phosphate$^{opls\_\{p,u,af\};\ mm\_\{p,u,af\}}$, Glycerol$^{mm\_\{p,u,af\}}$, Lactose$^{opls\_\{p,af\};\ mm\_\{p,af\}}$, Mannose$^{opls\_\{p,af\};\ mm\_\{p,u,af\}}$, Myo-inositol$^{opls\_\{p\};\ mm\_\{p,u,af\}}$, Sorbitol$^{opls\_\{p,u\};\ mm\_\{p,u\}}$
Glutathione metabolism	Mixed models^{u,af}	38	Ascorbate$^{mm\_\{af\}}$, Cadaverine$^{mm\_\{u,af\}}$, Cysteine$^{mm\_\{af\}}$, Glutamate$^{mm\_\{u\}}$, Glycine$^{mm\_\{u\}}$, Oxidized glutathione$^{mm\_\{u,af\}}$, Pyroglutamate$^{mm\_\{u,af\}}$, Reduced glutathione$^{mm\_\{u,af\}}$, Spermidine$^{mm\_\{u,af\}}$
Glycine, serine and threonine metabolism	OPLS-DA^{p} and mixed models^{p,u,af}	48	Aspartate$^{opls\_\{p\};\ mm\_\{p,af\}}$, Betaine$^{opls\_\{p\};\ mm\_\{p,u\}}$, Choline$^{mm\_\{p,u\}}$, Creatine$^{mm\_\{p,af\}}$, Cysteine$^{mm\_\{af\}}$, Glycerate$^{mm\_\{u,af\}}$, Glycine$^{opls\_\{p\};\ mm\_\{p,u\}}$, Guanidinoacetate$^{opls\_\{p\};\ mm\_\{p,u,af\}}$, Sarcosine$^{mm\_\{p,u\}}$, Serine$^{opls\_\{p\};\ mm\_\{p,af\}}$, Threonine$^{mm\_\{p,u,af\}}$
Lysine biosynthesis	Mixed models^{af}	32	2-Aminoadipate$^{mm\_\{af\}}$, Aspartate$^{mm\_\{af\}}$, Lysine$^{mm\_\{af\}}$, 2-Oxoglutarate$^{mm\_\{af\}}$
Lysine degradation	Mixed models^{u}	47	2-Aminoadipate$^{mm\_\{u\}}$, 5-Aminopentanoate$^{mm\_\{u\}}$, Cadaverine$^{mm\_\{u\}}$, Glycine$^{mm\_\{u\}}$, Lysine$^{mm\_\{u\}}$
Pantothenate and CoA biosynthesis	Mixed models^{p,af}	27	Aspartate$^{mm\_\{p,af\}}$, 2-Oxoisovalerate$^{mm\_\{p\}}$, Cysteine$^{mm\_\{af\}}$ Pantothenate$^{mm\_\{p,af\}}$, Valine$^{mm\_\{p,af\}}$
Pentose phosphate pathway	OPLS-DA^{p} and mixed models^{af}	32	Gluconate$^{opls\_\{p\};\ mm\_\{af\}}$, Glucose$^{opls\_\{p\};\ mm\_\{af\}}$, Glucose-6-phosphate$^{opls\_\{p\};\ mm\_\{af\}}$, Glycerate$^{mm\_\{af\}}$
Starch and sucrose metabolism	OPLS-DA^{u,af}	50	Fructose$^{opls\_\{u,af\}}$, Glucose$^{opls\_\{af\}}$, Glucose-6-phosphate$^{opls\_\{u,af\}}$, Glucuronate$^{opls\_\{u\}}$, Glycogen$^{opls\_\{af\}}$, Maltose$^{opls\_\{u\}}$
Taurine and hypotaurine metabolism	Mixed models^{af}	20	Alanine$^{mm\_\{af\}}$, Cysteine$^{mm\_\{af\}}$, Hypotaurine$^{mm\_\{af\}}$, Taurine$^{mm\_\{af\}}$
Valine, leucine and isoleucine biosynthesis	Mixed models^{p}	27	2-Oxoisovalerate$^{mm\_\{p\}}$, Isoleucine$^{mm\_\{p\}}$, Leucine$^{mm\_\{p\}}$, Threonine$^{mm\_\{p\}}$, Valine$^{mm\_\{p\}}$

Finally, guanidinoacetate is involved in the metabolic pathways of several amino acids. Six amino acids were found to be influential only in plasma (alanine, aspartate, glycine, leucine, serine and valine). Pathway enrichment analysis highlighted six pathways enriched in influential metabolites in plasma, including four related to amino acids: aminoacyl-tRNA biosynthesis, glycine, serine and threonine metabolism, cyanoamino acid metabolism and cysteine and methionine metabolism. In amniotic fluid, a pathway related to amino acids, arginine and proline metabolism, was also found to be enriched in influential metabolites.

4.3.2 Differential analyses

Mixed linear models were fitted to each metabolite independently. The complete model involved two factors, gestational stage and fetal genotype, as well as their interaction (fixed effects), with sow as a random effect. In addition, all differential metabolites (*i.e.*, metabolites for which the complete model was significantly better than the model with only the sow effect) were submitted to pathway enrichment analysis. To facilitate their individual interpretation, they were then associated with one of the best-fit sub-models derived from the complete model (see Methods). All the influential metabolites extracted by the multivariate analysis were also significantly differential in one of these mixed linear models, whatever the fluid. The detailed results of the mixed models are given online (https://www.nature.com/articles/s41598-020-76709-8#Sec18, Supplementary Information 2).

The mixed models revealed that the metabolomes differed more between the two gestational stages than between genotypes in all three fluids. In plasma, 57 differential metabolites were associated with a model that included the effect of the stage gestation (complete, additive, and only stage models) whereas only 28 differential metabolites were found associated with a model that included the genotype effect (complete, additive, and only genotype models). In urine, the comparison identified 41 versus 20 metabolites and in amniotic fluid, 58 versus 6 metabolites (Table 4.3).

Differences between stages of gestation

More differential metabolites were associated with a model with the stage of gestation effect (complete, additive and, only stage models) in plasma than in urine and amniotic fluid but, in amniotic fluid, 54 differential metabolites (out of 58) were associated with the model with only the

Table 4.3 – Number of differential metabolites associated with each sub-model according to the fluid.

Sub-model	Plasma	Urine	Amniotic fluid
Complete	8	0	0
Additive	15	5	4
Only stage	34	36	54
Only genotype	5	15	2
Total for models with a stage effect	**57**	**41**	**58**
Total for models with a genotype effect	**28**	**20**	**6**
Total	**62**	**56**	**60**

stage effect, which is the highest number of differential metabolites for this model among the three fluids. In addition, temporal changes in the quantifications of metabolites associated with the only stage model differed more in plasma than in the other fluids. The majority of metabolites (28/34) were more concentrated at 110 dg than at 90 dg in plasma whereas only half the metabolites were more concentrated at 110 dg than at 90 dg in urine and amniotic fluid.

Among the 20 proteinogenic amino acids, 15 were differential in at least one fluid. All differential amino acids were associated with a stage effect model, except for five: in urine, alanine, glutamine, glycine, proline and threonine, which were associated with the only genotype model. These 15 amino acids are involved in four pathways that were enriched in differential metabolites in all three fluids (Table 4.2), alanine, aspartate and glutamate metabolism", aminoacyl-tRNA biosynthesis, arginine and proline metabolism and glycine, serine and threonine metabolism. Fourteen out of 20 metabolites in plasma and 11 out of 18 metabolites in amniotic fluid were more concentrated at 110 dg (see Figure 4.8 for a representation of arginine, creatine, creatinine, glutamine, guanidinoacetate, proline and serine).

However, differences were also identified in the three fluids. In plasma, 2-oxoisovalerate, isoleucine, leucine, threonine, and valine were all differential and associated with the only stage model. They are involved in the valine, leucine and isoleucine biosynthesis, a pathway related to amino acids that was enriched in plasma. In amniotic fluid, 2-aminoadipate, aspartate, lysine, and 2-oxoglutarate were differential and were all associated with the only stage model. They are involved in another amino acid related pathway, lysine biosynthesis, which was enriched in amniotic fluid. These four metabolites (lysine, valine, leucine, and isoleucine) are described as essential amino acids in humans and in pigs and it is widely accepted that they

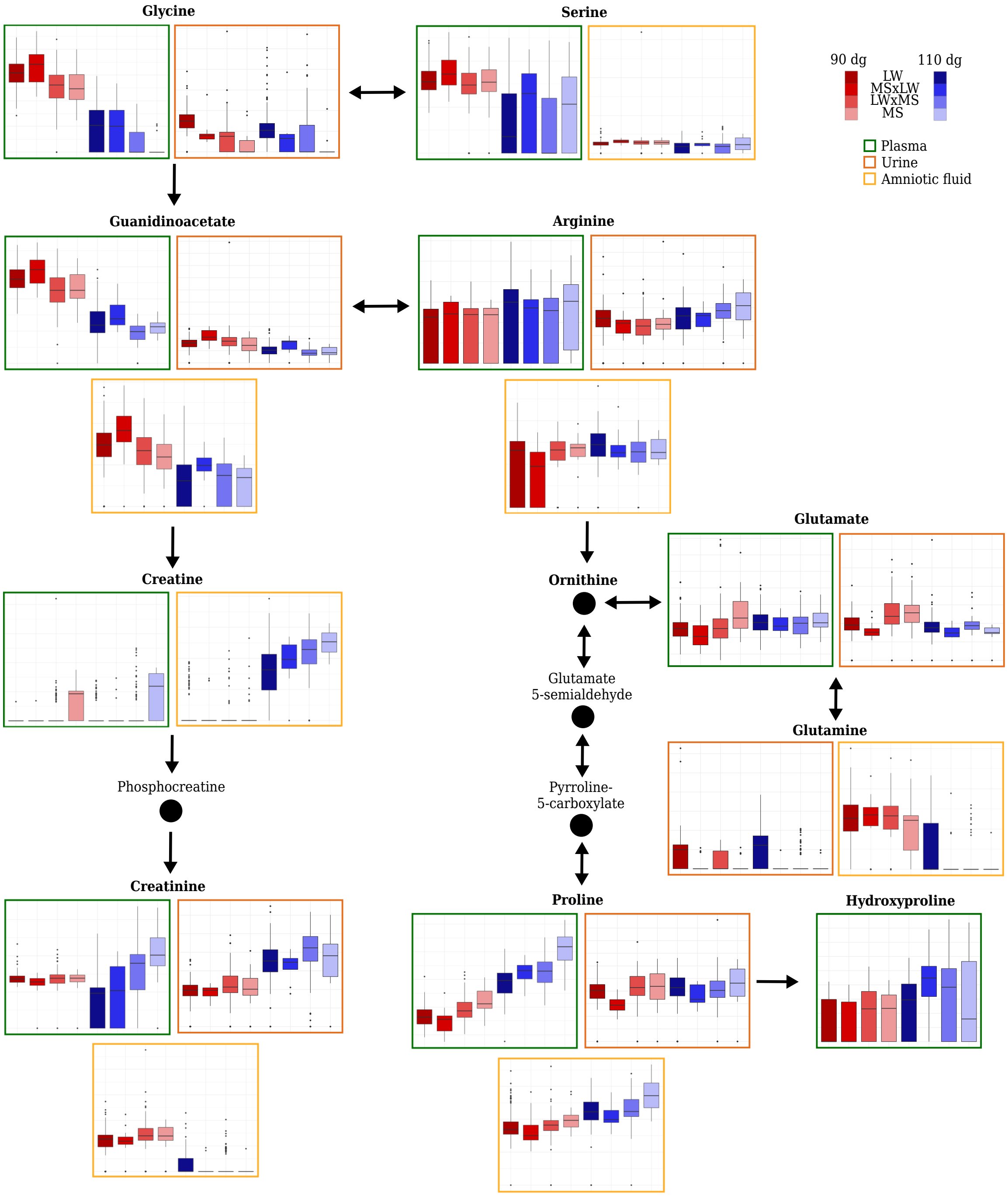


Figure 4.8 – Relative concentrations of some metabolites involved in amino acid metabolism (“arginine and proline metabolism” and “glycine, serine and threonine metabolism”) in plasma, urine and amniotic fluid at the two stages of gestation (90 dg and 110 dg, in red and blue respectively) and fetal genotypes (LW, MS×LW, LW×MS and MS, from left to right respectively). For the sake of clarity, only nine and seven metabolites out of 15 in plasma and 15 differential metabolites in urine are shown. Metabolites in bold are those included in the **ASICS** reference library. The coordinates of the y axes in boxplots can not be compared between two metabolites because the relative concentration limits of the boxplots are adapted to each metabolite.

are not synthesized by these organisms. However, metabolites of the two pathways (valine, leucine and isoleucine biosynthesis and lysine biosynthesis pathways) were all significantly more concentrated at 110 dg, which explains why they were found in our study.

In urine, fewer differential metabolites were associated with a model with the stage effect than in the other two fluids, especially amino acids. However, galactose metabolism, which was enriched in differential metabolites in the urine, contained five metabolites (myo-inositol, glucose-6-phosphate, mannose, sorbitol and galactitol), that were all associated with the only stage model.

Finally, four differential metabolites were also found in amniotic fluid, associated with a model including the stage effect: glucose-6-phosphate, gluconate, glucose and glycerate, among which three out of four (*i.e.*, except for gluconate) were associated with the only stage model. These metabolites are all involved in the pentose phosphate pathway which was found to be enriched in differential metabolites in amniotic fluid. This pathway was previously been found to be enriched in influential metabolites (as obtained by OPLS-DA) but in plasma rather than in amniotic fluid. In addition, two metabolites of this pathway, glucose, and gluconate, varied in opposite directions in the two fluids: glucose concentration was higher at 110 dg in plasma whereas it was higher at 90 dg in amniotic fluid (while the reverse was true for the concentration of gluconate).

Differences between genotypes

More differential metabolites were associated with a model that included the genotype effect (complete, additive and only genotype models) in plasma than in urine and amniotic fluid (28 metabolites compared to 20 and 6, respectively; Table 4.3).

In plasma, six differential metabolites (galactitol, glucose, glucose-6-phosphate, mannose, myo-inositol, and sorbitol) were associated with the complete or the additive model, which included a genotype effect, and two (glycerol and lactose) with the only stage model. Among these eight metabolites, some were also differential in urine and amniotic fluid but were not usually associated with a model including the genotype effect in these fluids. The only exceptions were glycerol in urine (associated with the only genotype model) and the galactitol in amniotic fluid (also associated with the only genotype model). In addition, these eight metabolites are all involved in the galactose metabolism pathway, which was enriched in differential metabolites in all fluids (Figure 4.9, 4.10 and 4.11 for plasma, urine and amniotic fluid, respectively).

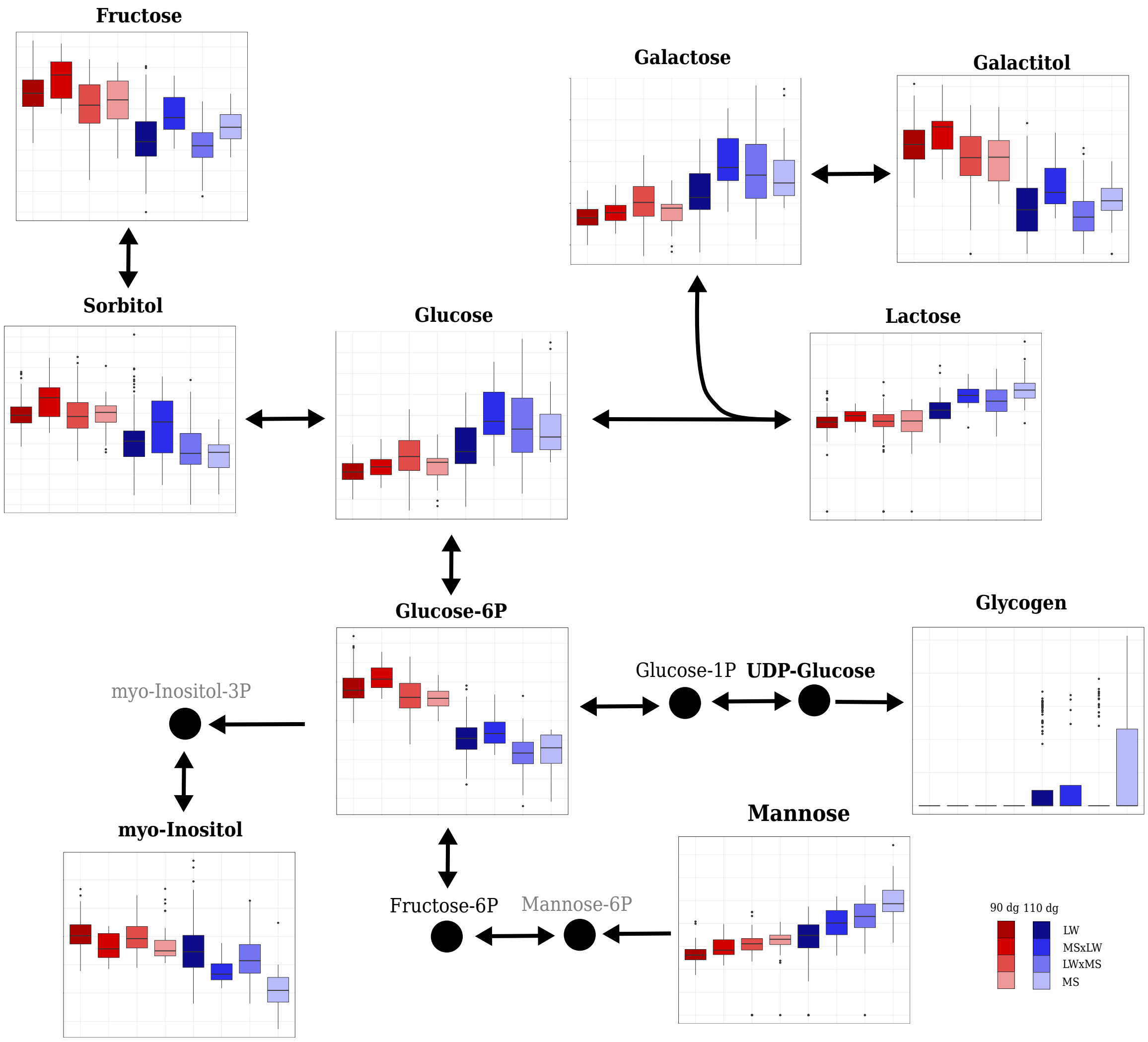


Figure 4.9 – Relative concentrations of some metabolites involved in the carbohydrate metabolism pathways ("galactose metabolism" and "starch and sucrose metabolism") in plasma according to the stage of gestation (90 dg and 110 dg, in red and blue respectively) and fetal genotypes (LW, MS×LW, LW×MS and MS, from left to right respectively). Metabolites in bold are those included in the **ASICS** reference library. The coordinates of the y axes in boxplots cannot be compared between two metabolites because the relative concentration limits of the boxplots are adapted to each metabolite.

In plasma, mannose and glucose were more concentrated at 110 dg than

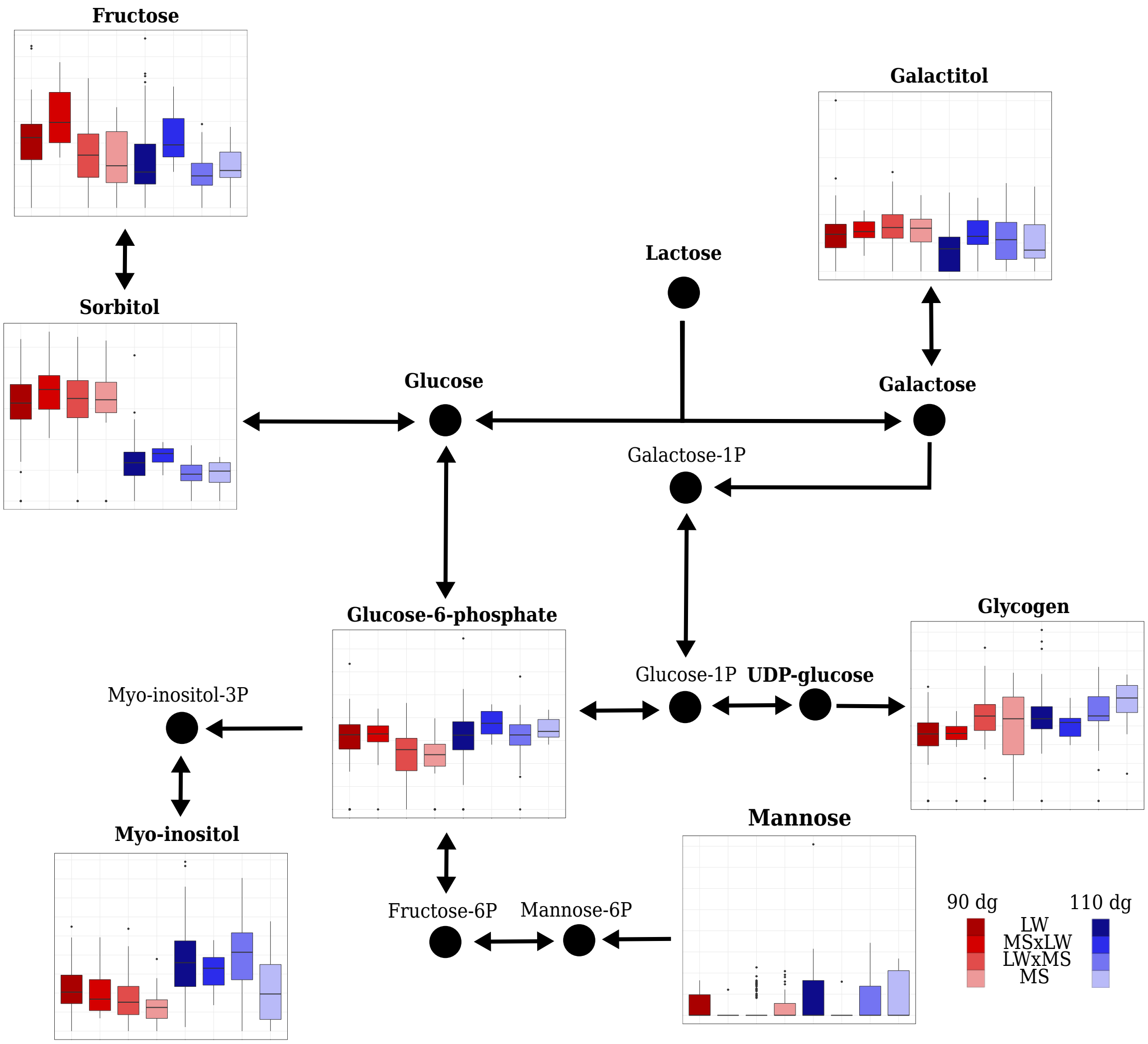


Figure 4.10 – Urine relative concentrations of some metabolites involved in the carbohydrate metabolism pathways ("galactose metabolism" and "starch and sucrose metabolism") according to stages of gestation (90 dg and 110 dg, in red and blue respectively) and to fetal genotypes (LW, MS×LW, LW×MS and MS, from left to right respectively). Metabolites in bold are the ones included in **ASICS** reference library. The coordinates of the y axes in boxplots can not be compared between two metabolites (relative concentrations limits of the boxplots are adapted to each metabolite).

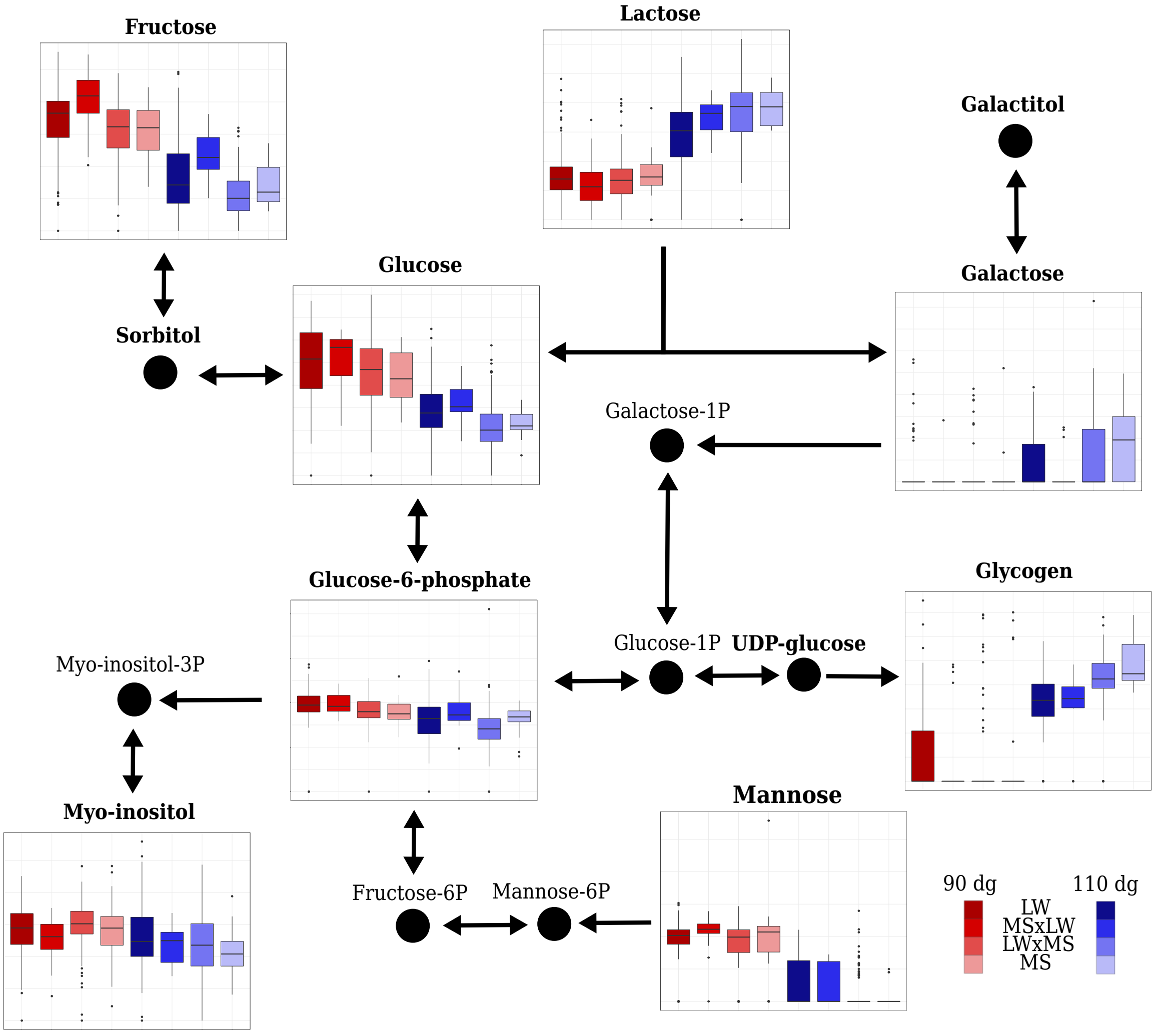


Figure 4.11 – Amniotic fluid relative concentrations of some metabolites involved in the carbohydrate metabolism pathways (“galactose metabolism” and “starch and sucrose metabolism”) according to stages of gestation (90 dg and 110 dg, in red and blue respectively) and to fetal genotypes (LW, MS×LW, LW×MS and MS, from left to right respectively). Metabolites in bold are the ones included in **ASICS** reference library. The coordinates of the y axes in boxplots can not be compared between two metabolites (relative concentrations limits of the boxplots are adapted to each metabolite).

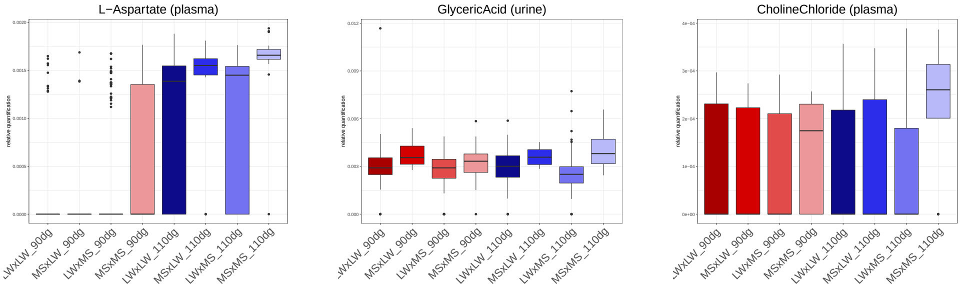
at 90 dg and were also more concentrated in MS than in LW at both 90 dg and 110 dg. On the contrary, the other three metabolites (glucose-6-phosphate, sorbitol and myo-inositol) were more concentrated at 90 dg and 110 dg in LW and the galactitol was more concentrated in MS at 110 dg. In addition, the concentration of myo-inositol was higher when the fetus had a LW father (whatever the mother genotype) and the concentration of glucose-6-phosphate, sorbitol and galactitol was higher when the fetus had a MS father. Conversely, in urine, the concentration of glycerol was higher when the fetus had a LW mother.

In plasma and urine, 12 differential metabolites of the enriched arginine and proline metabolism and glycine, serine and threonine metabolism pathways were also associated with a model including the genotype effect: creatinine (complete model), aspartate, glycine, guanidinoacetate and proline (additive model), choline and creatine (only genotype model) in plasma and glutamate and guanidinoacetate (additive model), 5-aminopentanoate, glutamine, glycerate, glycine, proline and threonine (only genotype model) in urine (see Figure 4.8 for a representation of creatine, creatinine, glutamate, glutamine, glycine, guanidinoacetate and proline). Among these metabolites, six were more concentrated in MS than in LW (5-aminopentanoate and glycerate in urine, aspartate, choline and creatine in plasma and proline in both urine and plasma). The concentrations of aspartate in plasma and of glycerate in urine were also higher when the fetus had a MS father compared to a LW father (paternal effect; Figure 4.12), while median concentrations of choline and creatine in plasma were higher than 0 only in fetuses with both a MS mother and father (effect of the pure MS genotype). Three metabolites (glutamine in urine and glycine and guanidinoacetate in plasma and urine) were more concentrated in LW than in MS.

In urine, four differential metabolites associated with the enriched glutathione pathway (Figure 4.13) were associated with the only genotype model (oxidized glutathione, glycine, and pyroglutamate) and one was associated with the additive model (glutamate). These metabolites were more concentrated in LW than in MS fetuses at 110 dg and more concentrated in MS than in LW at 90 dg, except glycine, which was still more concentrated in LW than in MS at 90 dg.

4.4 Discussion

The biological processes of fetal maturation and fetal growth retardation are of major interest in humans and in several mammalian livestock species,



(a) Aspartate concentrations in plasma

(b) Glycerate concentrations in urine

(c) Choline concentrations in plasma

Figure 4.12 – Paternal effect for aspartate in plasma and glycerate in urine and effect of the pure MS genotype for choline in plasma.

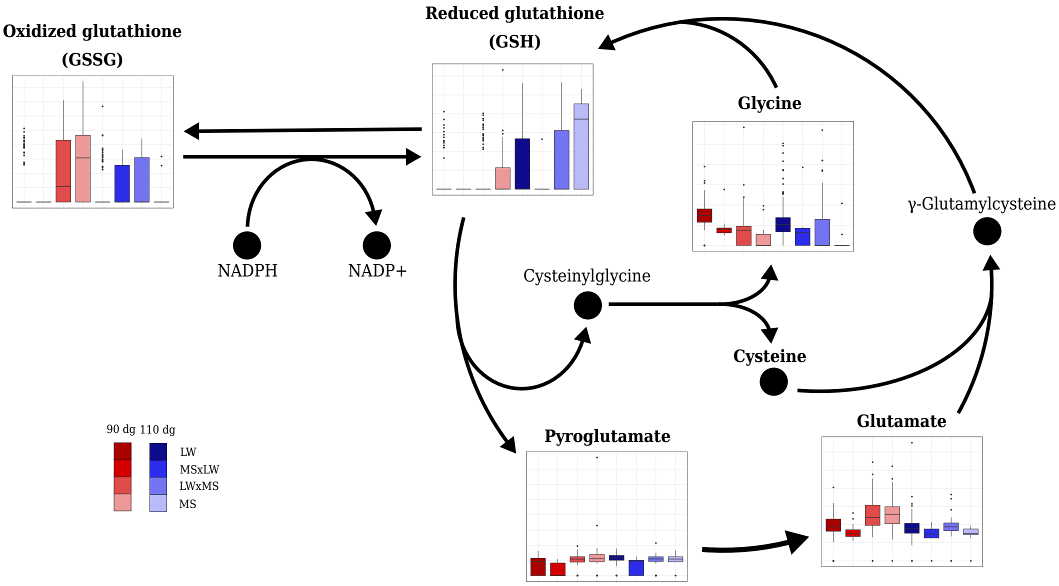


Figure 4.13 – Relative concentrations in urine according to the stage of gestation (90 dg and 110 dg, in red and blue respectively) and the fetal genotype (LW, MS×LW, LW×MS and MS, from left to right) for some metabolites in the “glutathione pathway”. For the sake of clarity, only the γ-glutamyl-cycle is represented in this figure. Metabolites in bold are those included in the **ASICS** reference library. The coordinates of the y axes in boxplots can not be compared between two metabolites because relative concentration limits of the boxplots are adapted to each metabolite.

including sheep (Goyal and Longo, 2015) and pig (Solberg et al., 2010; Wang et al., 2017). These processes are related to fetal development during late gestation, which is difficult to explore in mammalian species due to the invasiveness of experiments performed during that period. Since impaired maturation may induce postnatal developmental delay, metabolic syndrome, or early death (Gonzalez-Bulnes et al., 2016), the level of development at birth is currently evaluated by measuring birth weight (Baxter et al., 2008) as a proxy for intrauterine development. The study of the metabolome in the late gestation period is a promising way to predict immediate or later outcomes as well as to evaluate fetal growth retardation (Leite and Cecatti, 2019).

In humans, some metabolomic studies have already been performed on amniotic fluid collected during amniocentesis in the second or the third trimester of pregnancy (Graca et al., 2010; Fotiou et al., 2018). However, most other metabolomic surveys in humans have been performed later at birth, especially on cord blood (Favretto et al., 2012; Sanz-Cortés et al., 2013; El-Wahed et al., 2017), or on urine (Dessì et al., 2011). In non-human mammalian species, only a few studies related to the metabolome during late developmental processes have been published so far, including one plasmatic NMR study on pig fetuses in late gestation (Nissen et al., 2011).

Using NMR techniques, we acquired untargeted metabolomic measurements on three fluids (plasma, urine, and amniotic fluid) in 611 pig fetuses from four different genotypes at two different gestational stages. It should be noted that one limit of this technique is that it is not appropriate for quantifying lipids that form a heap of peaks in NMR spectra in water soluble fluids such that amniotic fluid, plasma and urine. No other lipidiomic approach has been investigated in this study. In addition, only a small number of metabolites involved in the glycolysis pathway were found in our study because most were not present in the **ASICS** reference library. Hence, the “glycolysis pathway” and the lipid related metabolisms were not found in our study, but are known to be important in late gestation. For example, Fainberg et al. (2012) investigated the lipidome of hepatic tissues to compare MS and commercial piglets immediately after birth. They identified five fatty acids that differentiated MS from commercial breeds and suggested that these differences may explain the better adaptability of MS piglets to the energetic demand for thermoregulation.

Despite these limitations, major differences were found in the three fluids at the two gestational stages pointing to a dramatic change in fetal metabolism between 90 and 110 dg. Such a metabolic switch was also recently reported in the metabolome of the amniotic fluid in humans between the second and the third trimesters of pregnancy (Orczyk-Pawilowicz et al., 2016). The metabolic switch observed in the present study is also consistent with our previous findings obtained using the same experimental design that highlighted important variations in the muscle and intestinal transcriptomes and in the muscle and adipose tissue proteomes in a smaller subset of the fetuses of both breeds (Voillet et al., 2014; Yao et al., 2017; Voillet et al., 2018; Gondret et al., 2018). More precisely, the first study of Voillet et al. (2014) identified important changes in the muscle transcriptome of both breeds between the two gestational ages (90 and 110 dg) by focusing on interaction effects between breed and gestational age. Notably, the study demonstrated that genes involved in muscular development are up-regulated around 90 dg and genes linked to metabolic functions, like gluconeogenesis, are up-regulated at 110 dg, whatever the genotype. Other later studies (Yao et al., 2017; Voillet et al., 2018; Gondret et al., 2018) confirmed this finding using similar analyses of the intestinal transcriptome and of the muscle and adipose tissue proteomes.

The present study identified the metabolomic pathways involved in the regulation of carbohydrates, amino acids, and glutathione metabolisms, which were found to be enriched in influential or differential metabolites. Many of the metabolites we identified are directly related to cellular energy levels and metabolism. It is indeed critical that carbohydrate metabolism is efficient at birth to provide the newborn piglet with the energy needed to overcome hypothermia due to birth, and subsequently with the energy required for maintenance, thermoregulation and growth (Le Dividich et al., 2005; Edwards and Baxter, 2015). In addition, the different pathways related to carbohydrate metabolisms have been shown to be altered during fetal development in neonates with IUGR (Mota-Rojas et al., 2011). In our study, galactose metabolism was the only pathway enriched in influential or differential metabolites whatever the fluid or the statistical method used. This pathway is essential in mammalian species, especially during fetal and neonatal development, because it plays an important role in energy delivery (Coelho et al., 2015). All seven metabolites (galactitol, glucose, glucose-6-phosphate, lactose, mannose, sorbitol, and myo-inositol) included in the galactose metabolism were identified by **ASICS** in the three fluids.

Among the metabolites identified in the galactose metabolism pathway, the concentration of myo-inositol has already been proposed as a marker of the development of obesity and type 2 diabetes in human adults (Dessì et al., 2011; Harder et al., 2007) and as a marker of IUGR in both humans (Dessì and Fanos, 2013; Barberini et al., 2014) and pigs (Nissen et al., 2011).

In these studies, a higher concentration of myo-inositol in plasma or urine was associated with a higher risk of IUGR and, thus, with lower maturity. In pigs, NMR metabolomic profiling performed on fetuses at 110 days of gestation demonstrated that low-weight fetuses had higher concentrations of myo-inositol in the plasma than high-weight fetuses (Nissen et al., 2011). Dessì and Fanos (2013) suggested that, in fetuses with IUGR, higher concentrations of myo-inositol in the plasma may reflect altered glucose metabolism and showed that fetuses with IUGR were also associated with a decrease in lipid synthesis and cell proliferation due to the reduction in insulin secretion. Such effects lead to lower birth weight. Consistently, we observed lower plasma concentrations of myo-inositol in MS fetuses, considered as more mature, at both 90 and 110 dg, despite the lower concentration of myo-inositol in urine at 90 dg compared to 110 dg (no genotype effect in this fluid). This finding suggests that more efficient glucose metabolism may partly explain the greater maturity at birth of MS piglets compared to LW piglets.

Glucose, another metabolite, is essential for the provision of the energy required for fetal growth and development. The concentration of glucose in pig plasma has already been shown to be lower in IUGR than in non-IUGR newborns (Lin et al., 2012; Staarvik et al., 2019). In our study, glucose was indeed differential and more concentrated in MS than in LW, both at 90 and 110 dg. Therefore, both myo-inositol (less concentrated in MS) and glucose (more concentrated in MS) may partly explain the better maturity of MS at birth. Moreover, the concentrations of these two metabolites were more influenced by the paternal genotype, which is consistent with a parental imprinting mechanism (Bischoff et al., 2009; Piedrahita, 2011). The role of genes, such as *IGF2* under parental imprinting during gestation has already been described (Giannoukakis et al., 1993) and its role in the fetal glycogen synthesis has also been demonstrated (Liang et al., 2010). However, the parental imprinting phenomenon has never previously been studied using metabolomic data before the current study in which we demonstrated that the concentration of some metabolites (*e.g.*, myo-inositol and glucose) depends on the paternal or maternal genotype in the reciprocal crossed fetuses.

In addition, during the last month of gestation, glucose is stored in fetal tissues, particularly in muscle and liver, in its polymerized form, *i.e.*, glycogen. The storage of glycogen just before birth has been known since centuries (Young, 1957): comparison of glycogen contents at birth and a few hours after birth showed that muscle and liver glycogen contents dropped dramatically in mammalian species a few hours after birth. In our study, glycogen was detected in urine, amniotic fluid, and to a lesser extent, in plasma, and its concentration was significantly higher at 110 dg than at 90 dg in all three fluids. At birth, piglets mainly rely on glycogen as an energy-yielding substrate before colostrum consumption (Mellor and Cockburn, 1986; van der Lende et al., 2001). Studying pig fetuses close to term in relation to their value for survival at birth, Leenhouwers et al. (2002) and Voillet et al. (2018) showed that glycogen content in liver and muscle increased with increased chance of survival. Since glycogen is a multibranched polysaccharide of glucose described as a reserve in tissues, it was surprising to find it in the three fluids we studied (*i.e.*, plasma, urine, and amniotic fluid). One possible explanation is that, as glycogen synthesis in tissues is intense, some polymers of glucose may have been released into the fluids just before birth.

Concerning carbohydrates, many amino acids were highlighted in our analyses. Nine amino acid metabolism pathways were found to be enriched in influential and differential metabolites at the end of gestation: alanine, aspartate and glutamate metabolism, aminoacyl tRNA biosynthesis, arginine and proline metabolism, cyanoamino acid metabolism, cysteine and methionine metabolism, glycine, serine and threonine metabolism, lysine biosynthesis, lysine degradation and valine, leucine and isoleucine biosynthesis. They all respond to the need for fetal development and maturation since amino acids play nutritional, physiological and regulatory roles. Twenty amino acids are known to be involved in these pathways (Salway, 2017). In our study, these pathways were enriched in 15 differential amino acids, including five essential amino acids (*i.e.*, amino acids that cannot be synthetized by animals) and 10 non essential amino acids (*i.e.*, that can be synthetized by animals). Amino acids of the arginine and proline metabolism (arginine, asparagine, aspartate, glutamate, glutamine, ornithine, and proline) are already well studied during gestation because of their essential role in fetal growth and development both in humans and pigs (Lin et al., 2014).

The arginine concentration in the amniotic fluid in early pregnancy has been described as being positively correlated with birth weight, body length and head circumference of babies (Bjørke-Jenssen et al., 2017). In addition, Wu et al. (2013) showed that arginine supplementation of sows during gestation reduced the stillbirth rate and the risk of IUGR. These two studies support an important role of arginine in fetal maturation and their results are also consistent with ours: arginine was identified by mixed models as differential. It was more concentrated at 110 dg than at 90 dg in all three fluids, although an earlier study (Lin et al., 2012) showed a reduction in the concentration of arginine in plasma between 90 and 110 dg.

Like arginine, glutamine is also highly concentrated in amniotic fluid mainly in early gestation (Wu et al., 1996). At the end of gestation, the amniotic fluid serves as a nutritional reservoir for the fetus and, as a result, uptake of glutamine by the fetus may reduce the concentration of glutamine in amniotic fluid (Lin et al., 2014). Hence, the glutamine concentration is usually considered as a limiting factor of fetal growth and a lower concentration is known to be associated with an increased risk of IUGR risk. This was confirmed by our study in which glutamine was not found at 110 dg in amniotic fluid. Proline, which is also involved in the arginine and proline metabolism pathway, is less frequently used in sow nutrition than arginine and glutamine (Liu et al., 2019). However, its important role in polyamine synthesis during the pig gestation has already been demonstrated (Wu et al., 2008). As expected, the higher concentration of proline in MS than in LW at 90 and 110 dg in plasma could be related to the lower maturity of LW piglets and to a potential delay in development already identified in these fetuses (Voillet et al., 2014, 2018; Gondret et al., 2018; Yao et al., 2017). In addition, in plasma, crossbred fetuses have intermediate concentrations of proline, with no specific maternal or paternal effect.

Serine, glycine and guanidinoacetate metabolites are involved in the “one-carbon metabolism” (this metabolism is not a KEGG pathway and its enrichment was consequently not analyzed here). This metabolism is involved in DNA methylation by providing methyl groups (Lin et al., 2014). Like for imprinting genes, DNA methylation is an important epigenetic mechanism of fetal gestation. Previous studies have shown the association between IUGR and epigenetic alterations (Lin et al., 2014). Lin et al. (2012) also showed that the concentration of serine significantly decreased between 90 dg and 110 dg in pigs. A higher concentration of serine and glycine in plasma has also been reported in IUGR rat fetuses compared to normal weight fetuses (Parimi et al., 2004). This is consistent with our findings: glycine and serine concentrations were differential and in plasma a higher concentration was found at 90 dg than at 110 dg. The concentration of glycine was also more concentrated in LW than in MS at both 90 and 110 dg. Guanidinoacetate exhibited a maternal effect: concentrations of this metabolite in fetuses with a LW mother were higher than in fetuses with a MS mother whatever the stage of gestation. For glycine, the same maternal effect was observed at 90 dg. These metabolites are precursors of creatine (see Figure 4.8), which is known to be involved in energy metabolism and development of skeletal muscles (Brosnan et al., 2009; Wang et al., 2013b). In plasma, both creatine and creatinine were differential and were more concentrated in MS than in LW at 110 dg. In contrast to our findings, other studies on IUGR reported a higher concentration of creatine and creatinine in IUGR fetal pigs (Lin et al., 2012) or newborns (Dessì et al., 2011). However, in our study, it should be noted that the concentration of creatinine in plasma changed differently according to the genotype. Indeed, at 90 dg, the concentration was approximately the same, whatever the genotype, but in LW, the concentration then decreased sharply whereas in MS, it increased sharply and was higher at 110 dg than at 90 dg.

As the production of oxidants increases during gestation due to cell proliferation, it is necessary that the glutathione metabolism is efficient because it plays a role in oxidative defense (Wu et al., 2004, 2006). An increase in oxidative stress has already been associated with IUGR or preterm infants (He et al., 2011; Wu et al., 2006; Hracsko et al., 2008; Jackson et al., 1987). As expected, in our study, glutathione metabolism was enriched in differential metabolites in urine. Reduced glutathione is formed from glutamate, cysteine, and glycine and protects cells against oxidative damage by removing hydrogen peroxide (Salway, 2017). Oxidized glutathione was only detected in MS and was more concentrated at 90 dg than at 110 dg. Conversely, reduced glutathione was more concentrated at 110 dg than at 90 dg in MS and was almost undetectable in LW. Taken together, these results suggest a better oxidative defense in MS than in LW. Interestingly, the concentration of pyroglutamate, another metabolite involved in the glutathione metabolism, has already been shown to be more concentrated in the plasma of IUGR fetuses than in a normal birth weight group, likely due to reduced glutathione synthesis (Wellner et al., 1974). In addition, pyroglutamate has been suggested as a biomarker for IUGR in fetal plasma (Lin et al., 2012). However, this is still not clearly established fact. For instance, Jackson et al. (1987) showed that the pyroglutamate/creatinine ratio in urinary excretion just after birth was higher in preterm infants. We observed the same trend during late gestation: pyroglutamate in plasma was more concentrated at 110 dg than at 90 dg and the pyroglutamate/creatinine ratio was higher in LW than in MS in urine.

4.5 Conclusions

Our study of changes in metabolism in late gestation in two contrasted pig breeds provide useful insights into potential biomarkers and metabolism pathways associated with survival at birth. In particular, proline and myo-inositol are two promising metabolites for the characterization of piglet maturity. They illustrate the importance of amino acid and carbohydrate metabolisms for fetal development in late gestation.

However, the relative quantification of metabolites we used in this study might not be sufficient to derive biomarkers with thresholds based on absolute quantifications. To achieve this goal, other targeted studies will be necessary, along with the training of an adequate prediction method to set appropriate thresholds. The comprehensive view of fetal metabolome we provided paves the way for the design of such studies.

Acknowledgements

This project received financial support from French National Agency of Research (PORCINET grant ANR-09-GENM005). The PhD fellowship of Gaelle Lefort work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004, by INRAE Mathematics and Computer Science Division, by INRAE Animal Genetics Division and by INRAE Animal Health Divison. The funders had no role in the study design, analyses, results interpretation and decision to publish. The authors are very grateful to the staff of experimental pig facilities (INRA, 2018. Pig phenotyping and Innovative breeding facility, doi:10.15454/1.5572415481185847E12) and Marie-Christine Père (PEGASE, INRAE) for expert assistance in surgeries. They also thank all the participants from GenPhySE, PEGASE and GenESI laboratories (INRAE) for sample and data collection.The authors thank Jasha Leenhouwers for the sharing of expertise on piglet survival and his valuable recommendations for data collection. The authors also thank Anne Poujol for her veterinarian assistance and Daphne Goodfellow for English proofreading and correction.

Data avaibility

The data supporting the results of this article are available in the MetaboLights database (Haug et al., 2020): MTBLS1541 (www.ebi.ac.uk/metabolights/MTBLS1541).

Chapitre 5

Conclusions et perspectives

Dans cette thèse, nous avons proposé plusieurs contributions portant sur l'identification et la quantification automatique des métabolites à partir d'un ou plusieurs spectres complexes de Résonance Magnétique Nucléaire. Dans des travaux précédents, une première version de la méthode d'identification et de quantification ASICS a été développée et a permis d'obtenir de meilleurs résultats que les autres méthodes existantes (Tardivel et al., 2017) en se focalisant principalement sur l'étape d'identification. Nous nous sommes, ici, intéressés à l'ensemble des différentes étapes afin d'améliorer la fiabilité des quantifications finales.

Avec ASICS, deux étapes préliminaires sont nécessaires pour quantifier les métabolites de manière fiable à l'aide d'une librairie de spectres de métabolites purs : l'alignement des spectres de cette librairie sur le spectre complexe puis l'identification et la quantification des métabolites par reconstruction du signal avec un modèle linéaire. La première étape d'alignement permet de corriger les décalages de pics entre les spectres dus à des différences de conditions expérimentales. Même si les différences sont faibles, il est important que les pics des spectres purs soient correctement alignés avec ceux du mélange complexe. Nous avons donc proposé une méthode d'alignement des spectres purs en deux phases : un alignement global du spectre suivi d'un alignement local pic par pic. À l'inverse des méthodes d'alignement existantes, celle-ci permet de gérer la différence entre le nombre de pics du spectre complexe (qui est élevé) et le nombre de pics des spectres purs (qui est beaucoup plus faible). Il reste cependant des erreurs d'alignement de certains pics qui entraînent des erreurs de quantification. Toujours pour la phase d'alignement, nous avons proposé une méthode permettant d'utiliser l'information d'un ensemble de spectres complexes d'une même expérience. En effet, ces spectres, qui sont supposés globalement semblables, permettent de calibrer de manière efficace le décalage global dans l'alignement du spectre pur sur l'ensemble des spectres complexes en éliminant les erreurs qui peuvent intervenir sur un faible nombre de spectres complexes.

Dans la même idée d'utiliser conjointement l'information provenant de plusieurs spectres complexes, nous avons proposé l'utilisation d'un modèle *group-Lasso* à réponses multiples plutôt que des quantifications indépendantes pour chacun des spectres complexes. Cela permet, d'une part, d'estimer conjointement les quantifications pour obtenir des valeurs plus fiables. D'autre part, grâce à la pénalité ajouté au modèle, cela élimine les métabolites qui étaient identifiés dans un très petit nombre de spectres (et qui étaient, de ce fait, probablement des faux positifs). La quantification avec ces différentes méthodes sur des spectres simulés a montré l'apport des différentes versions de l'approche tenant compte de l'ensemble des spectres sur les quantifications obtenues. Plus précisément, elle améliore surtout les identifications et quantifications des métabolites présent en faibles quantités. De plus, même si le temps de calcul des quantifications est légèrement plus élevé, l'approche jointe permet de gérer des décalages importants entre les positions des pics sur les spectres complexes et celles sur les spectres purs.

Cependant, même si les résultats sont satisfaisants et ont été validés sur différents jeux de données, il reste des marges de progression car des métabolites sont encore identifiés à tort. Généralement, lors d'une identification des métabolites dans les spectres RMN par un expert, celui-ci s'appuie sur des spectres RMN en deux dimensions. Ces spectres permettent de mettre en évidence des corrélations plus complexes entre les différents noyaux d'une même molécule (Marchand et al., 2017). Il existe plusieurs types de spectres 2D parmi lesquels la J-résolue (Aue et al., 1976), la *COrrelation SpectroscopY* (COSY ; Bax and Freeman (1981)) et l'*Heteronuclear Single Quantum Correlation* (HSQC ; Bodenhausen and Ruben (1980)). Comme pour les spectres en une dimension, l'identification des métabolites est fastidieuse et manuelle. Une perspective d'amélioration de notre méthode serait d'utiliser ces spectres de manière automatisée pour valider les identifications. Pour cela, il faudrait tout d'abord développer une méthode d'identification automatique des métabolites dans les spectres 2D. Ensuite, resterait la question de la combinaison de cette information avec celle issue de la RMN 1D.

Un second volet de cette thèse a concerné l'application de cette méthode au problème de mortalité néonatale des porcelets et plus précisément à la description des mécanismes impliqués dans la mise en place de la maturité durant le dernier tiers de gestation. Grâce à l'utilisation de la méthode ASICS, la quantification de spectres RMN de plasma, d'urine et de liquide amniotique de fœtus en fin de gestation a permis d'identifier un grand nombre de métabolites présents dans ces fluides. Des modèles mixtes suivis d'une étude d'enrichissement des voies métaboliques ont ensuite pu mettre en évidence des voies métaboliques potentiellement liées à une plus faible ou une plus forte maturité de certains porcelets. Ces voies métaboliques impliquent de nombreux acides aminés et sucres (croissance et apport d'énergie) ainsi que le métabolisme du glutathion (stress oxydatif).

Sur les mêmes animaux, plusieurs autres types de données omiques ont également été mesurés et restent à exploiter. Une première perspective à ce travail serait d'intégrer ces différents types de données pour avoir une description plus précise de la mise en place de la maturité avant la naissance. Comme l'objectif est d'obtenir des biomarqueurs à la naissance, il faudrait aussi valider l'utilisation de ces métabolites comme potentiel biomarqueurs sur des échantillons obtenus à la naissance. Durant cette thèse, des échantillons obtenus à la naissance ont été utilisés pour valider les quantifications obtenues avec ASICS mais, pour l'instant, les résultats biologiques de cette étude sont encore en cours d'approfondissement.

5

Bibliographie

Alonso, A., Marsal, S., and Julià, A. (2015). Analytical methods in untargeted metabolomics : state of the art in 2015. *Frontiers in Bioengineering and Biotechnology*, 3 :23.

Astle, W., De Iorio, M., Richardson, S., Stephens, D., and Ebbels, T. (2012). A bayesian model of NMR spectra for the deconvolution and quantification of metabolites in complex biological mixtures. *Journal of the American Statistical Association*, 107(500) :1259–1271.

Aue, W., Karhan, J., and Ernst, R. (1976). Homonuclear broad band decoupling and two-dimensional j-resolved nmr spectroscopy. *The Journal of Chemical Physics*, 64(10) :4226–4227.

Barberini, L., Noto, A., Fattuoni, C., Grapov, D., Casanova, A., Fenu, G., Gaviano, M., Carboni, R., Ottonello, G., Crisafulli, M., Fanos, V., and Dessì, A. (2014). Urinary metabolomics (GC-MS) reveals that low and high birth weight infants share elevated inositol concentrations at birth. *The Journal of Maternal-Fetal & Neonatal Medicine*, 27 Suppl 2 :20–26.

Basso, O. and Wilcox, A. (2010). Mortality risk among preterm babies : immaturity versus underlying pathology. *Epidemiology*, 21 :521–527.

Bax, A. and Freeman, R. (1981). Investigation of complex networks of spin-spin coupling by two-dimensional nmr. *Journal of Magnetic Resonance (1969)*, 44(3) :542–561.

Baxter, E. M., Jarvis, S., D'Eath, R. B., Ross, D. W., Robson, S. K., Farish, M., Nevison, I. M., Lawrence, A. B., and Edwards, S. A. (2008). Investigating the behavioural and physiological indicators of neonatal survival in pigs. *Theriogenology*, 69(6) :773–783.

Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, 2(11) :2692–2703.

Beirnaert, C., Meysman, P., Vu, T. N., Hermans, N., Apers, S., Pieters, L., Covaci, A., and Laukens, K. (2018). speaq 2.0 : a complete workflow for high-throughput 1D NMR spectra processing and quantification. *PLoS Computational Biology*, 14(3) :e1006018.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57(1) :289–300.

Bingol, K. (2018). Recent advances in targeted and untargeted metabolomics by nmr and ms/nmr methods. *High-Throughput*, 7(2).

Bischoff, S. R., Tsai, S., Hardison, N., Motsinger-Reif, A. A., Freking, B. A., Nonneman, D., Rohrer, G., and Piedrahita, J. A. (2009). Characterization of conserved and non-conserved imprinted genes in swine. *Biology of Reproduction*, 81 :906–920.

Bjørke-Jenssen, A., Ueland, P. M., and Bjørke-Monsen, A.-L. (2017). Amniotic fluid arginine from gestational weeks 13 to 15 is a predictor of birth weight, length, and head circumference. *Nutrients*, 9.

Bodenhausen, G. and Ruben, D. J. (1980). Natural abundance nitrogen-15 nmr by enhanced heteronuclear spectroscopy. *Chemical Physics Letters*, 69(1) :185–189.

Brosnan, J. T., Wijekoon, E. P., Warford-Woolgar, L., Trottier, N. L., Brosnan, M. E., Brunton, J. A., and Bertolo, R. F. P. (2009). Creatine synthesis is a major metabolic process in neonatal piglets and has important implications for amino acid metabolism and methyl balance. *The Journal of Nutrition*, 139 :1292–1297.

Cañueto, D., Gómez, J., Salek, R. M., Correig, X., and Cañellas, N. (2018). rDolphin : a GUI R package for proficient automatic profiling of 1D 1H-NMR spectra of study datasets. *Metabolomics*, 14(3) :24.

Canario, L., Cantoni, E., Le Bihan, E., Caritez, J. C., Billon, Y., Bidanel, J. P., and Foulley, J. L. (2006). Between-breed variability of stillbirth and its relationship with sow and piglet characteristics. *Journal of Animal Science*, 84(12) :3185–3196.

Canario, L., Père, M. C., Tribout, T., Thomas, F., David, C., Gogué, J., Herpin, P., Bidanel, J. P., and Le Dividich, J. (2007). Estimation of genetic trends from 1977 to 1998 of body composition and physiological state of Large White pigs at birth. *Animal*, 1 :1409–1413.

Canet, D., Boubel, J., and Canet Soulas, E. (2002). *La RMN : Concepts, méthodes et applications, 2e édition.*

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D. S., and Xia, J. (2018). MetaboAnalyst 4.0 : towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research*, 46(W1) :W486–W494.

Coelho, A. I., Berry, G. T., and Rubio-Gozalbo, M. E. (2015). Galactose metabolism and health. *Current Opinion in Clinical Nutrition and Metabolic Care*, 18(4) :422–427.

Conover, W. (1999). *Practical Nonparametric Statistics*, volume 350 of *Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics*. John Wiley & Sons, New York, NY, USA.

Considine, E., Thomas, G., Boulesteix, A., Khashan, A., and Kenny, L. (2018). Critical review of reporting of the data analysis step in metabolomics. *Metabolomics*, 14(1) :7.

Craig, A., Cloarec, O., Holmes, E., Nicholson, J., and Lindon, J. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7) :2262–2267.

De Souza, D. P., Saunders, E. C., McConville, M. J., and Liki, V. A. (2006). Progressive peak clustering in gc-ms metabolomic experiments applied to leishmania parasites. *Bioinformatics*, 22 :1391–1396.

Dessì, A., Atzori, L., Noto, A., Visser, G. H. A., Gazzolo, D., Zanardo, V., Barberini, L., Puddu, M., Ottonello, G., Atzei, A., De Magistris, A., Lussu, M., Murgia, F., and Fanos, V. (2011). Metabolomics in newborns with intrauterine growth retardation (IUGR) : urine reveals markers of metabolic syndrome. *The Journal of Maternal-Fetal & Neonatal Medicine*, 24 Suppl 2 :35–39.

Dessì, A. and Fanos, V. (2013). Myoinositol : a new marker of intrauterine growth restriction ? *Journal of Obstetrics and Gynaecology*, 33 :776–780.

Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in 1h nmr metabonomics. *Analytical chemistry*, 78 :4281–4290.

Dunn, W. B., Broadhurst, D. I., Atherton, H. J., Goodacre, R., and Griffin, J. L. (2011). Systems level studies of mammalian metabolomes : the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews*, 40 :387–426.

Edwards, S. and Baxter, E. (2015). Piglet mortality : causes and prevention. In Farmer, C., editor, *The Gestating and Lactating Sow*, pages 253–278. Wageningen Academic Publishers, Wageningen, The Netherlands.

El-Wahed, M. A., El-Farghali, O., ElAbd, H., El-Desouky, E., and Hassan, S. (2017). Metabolic derangements in IUGR neonates detected at birth using UPLC-MS. *Egyptian Journal of Medical Human Genetics*, 18(3) :281–287.

Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G. A. N., Raftery, D., Alahmari, F., Jaremko, L., Jaremko, M., and Wishart, D. S. (2019). Nmr spectroscopy for metabolomics research. *Metabolites*, 9.

Emwas, A.-H. M. (2015). The strengths and weaknesses of nmr spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods in Molecular Biology*, 1277 :161–193.

Fainberg, H. P., Bodley, K., Bacardit, J., Li, D., Wessely, F., Mongan, N. P., Symonds, M. E., Clarke, L., and Mostyn, A. (2012). Reduced neonatal mortality in Meishan piglets : a role for hepatic fatty acids ? *PLoS ONE*, 7(11) :e49101.

Favretto, D., Cosmi, E., Ragazzi, E., Visentin, S., Tucci, M., Fais, P., Cecchetto, G., Zanardo, V., Viel, G., and Ferrara, S. D. (2012). Cord blood metabolomic profiling in intrauterine growth restriction. *Analytical and Bioanalytical Chemistry*, 402 :1109–1121.

Fiehn, O. (2002). Metabolomics–the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2) :155–171.

Filntisi, A., Fotakis, C., Asvestas, P., Matsopoulos, G. K., Zoumpoulakis, P., and Cavouras, D. (2017). Automated metabolite identification from biological fluid 1H NMR spectra. *Metabolomics*, 13(12) :146.

Fotiou, M., Fotakis, C., Tsakoumaki, F., Athanasiadou, E., Kyrkou, C., Dimitropoulou, A., Tsiaka, T., Chatziioannou, A. C., Sarafidis, K., Menexes, G., Theodoridis, G., Biliaderis, C. G., Zoumpoulakis, P., Athanasiadis, A. P., and Michaelidou, A.-M. (2018). ^{1}H NMR-based metabolomics reveals the effect of maternal habitual dietary patterns on human amniotic fluid profile. *Scientific Reports*, 8 :4076.

Foxcroft, G., Dixon, W., Novak, S., Putman, C., Town, S., and Vinsky, M. (2006). The biological basis for prenatal programming of postnatal perfomance in pigs. *Journal of Animal Science*, 84(Suppl 13) :E105–E112.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1).

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., and Zhang, J. (2004). Bioconductor : open software development for computational biology and bioinformatics. *Genome Biology*, 5(10) :R80.

Giannoukakis, N., Deal, C., Paquette, J., Goodyer, C. G., and Polychronakos, C. (1993). Parental genomic imprinting of the human IGF2 gene. *Nature Genetics*, 4(1) :98–101.

Giraudeau, P. (2017). Challenges and perspectives in quantitative nmr. *Magnetic resonance in chemistry*, 55 :61–69.

Giskeødegård, G. F., Bloemberg, T. G., Postma, G., Sitter, B., Tessem, M.-B., Gribbestad, I. S., Bathen, T. F., and Buydens, L. M. C. (2010). Alignment of high resolution magic angle spinning magnetic resonance spectra using warping methods. *Analytica chimica acta*, 683 :1–11.

Gondret, F., Guével, B., Père, M.-C., Quesnel, H., Billon, Y., Com, E., Canario, L., Louveau, I., and Liaubet, L. (2018). Proteomic analysis of adipose tissue during the last weeks of gestation in pure and crossbred Large White or Meishan fetuses gestated by sows of either breed. *Journal of Animal Science and Biotechnology*, 9(1) :28.

Gonzalez-Bulnes, A., Astiz, S., Ovilo, C., Lopez-Bote, C. J., Torres-Rovira, L., Barbero, A., Ayuso, M., Garcia-Contreras, C., and Vazquez-Gomez, M. (2016). Developmental origins of health and disease in swine : implications for animal production and biomedical research. *Theriogenology*, 86 :110–119.

Goyal, R. and Longo, L. D. (2015). Metabolic profiles in ovine carotid arteries with developmental maturation and long-term hypoxia. *PloS One*, 10 :e0130739.

Graca, G., Duarte, I. F., Barros, A. S., Goodfellow, B. J., Diaz, S. O., Pinto, J., Carreira, I. M., Galhano, E., Pita, C., and Gil, A. M. (2010). Impact of prenatal disorders on the metabolic profile of second trimester amniotic fluid : a nuclear magnetic resonance metabonomic study. *Journal of Proteome Research*, 9 :6016–6024.

Guitton, Y., Tremblay-Franco, M., Le Corguillé, G., Martin, J.-F., Pétéra, M., Roger-Mele, P., Delabrière, A., Goulitquer, S., Monsoor, M., Duperier, C., Canlet, C., Servien, R., Tardivel, P., Caron, C., Giacomoni, F., and Thévenot, E. A. (2017). Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. *The International Journal of Biochemistry & Cell Biology*, 93 :89–101.

Gómez, J., Brezmes, J., Mallol, R., Rodríguez, M. A., Vinaixa, M., Salek, R. M., Correig, X., and Cañellas, N. (2014). Dolphin : a tool for automatic targeted metabolite profiling using 1d and 2d (1)h-nmr data. *Analytical and bioanalytical chemistry*, 406 :7967–7976.

Hao, J., Astle, W., de Iorio, M., and Ebbels, T. (2012). BATMAN – an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics*, 28(15) :2088–2090.

Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, J. G., and Ebbels, T. M. D. (2014). Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nature Protocols*, 9(6) :1416–1427.

Harder, T., Rodekamp, E., Schellong, K., Dudenhausen, J. W., and Plagemann, A. (2007). Birth weight and subsequent risk of type 2 diabetes : a meta-analysis. *American Journal of Epidemiology*, 165 :849–857.

Haug, K., Cochrane, K., Nainala, V. C., Williams, M., Chang, J., Jayaseelan, K. V., and O'Donovan, C. (2020). MetaboLights : a resource evolving in response to the needs of its scientific community. *Nucleic Acids Research*, 48 :D440–D444.

Haug, K., Salek, R., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., González-Beltrán, A., Sansone, S., Griffin, J., and Steinbeck, C. (2013). MetaboLights – an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, 41(D1) :D781–D786.

He, Q., Ren, P., Kong, X., Xu, W., Tang, H., Yin, Y., and Wang, Y. (2011). Intrauterine growth restriction alters the metabonome of the serum and jejunum in piglets. *Molecular Biosystems*, 7 :2147–2155.

Herpin, P., Le Dividich, J., and Amaral, N. (1993). Effect of selection for lean tissue growth on body composition and physiological state of the pig at birth. *Journal of animal science*, 71 :2645–2653.

Hore, P. (1995). *Nuclear Magnetic Resonance*. Oxford University Press.

Hracsko, Z., Orvos, H., Novak, Z., Pal, A., and Varga, I. S. (2008). Evaluation of oxidative stress markers in neonates with intra-uterine growth retardation. *Redox Report*, 13 :11–16.

Jackson, A. A., Badaloo, A. V., Forrester, T., Hibbert, J. M., and Persaud, C. (1987). Urinary excretion of 5-oxoproline (pyroglutamic aciduria) as an index of glycine insufficiency in normal man. *The British Journal of Nutrition*, 58 :207–214.

Jacob, D., Deborde, C., and Moing, A. (2013). An efficient spectra processing method for metabolite identification from 1h-nmr metabolomics data. *Analytical and Bioanalytical Chemistry*, 405 :5049–5061.

Kazmi, S. A., Ghosh, S., Shin, D.-G., Hill, D. W., and Grant, D. F. (2006). Alignment of high resolution mass spectra : development of a heuristic approach for metabolomics. *Metabolomics*, 2(2) :75–83.

Le Dividich, J., Rooke, J., and Herpin, P. (2005). Nutritional and immunological importance of colostrum for the new-born pig. *The Journal of Agricultural Science*, 143(6) :469–485.

Leenhouwers, J. I., Knol, E. F., de Groot, P. N., Vos, H., and van der Lende, T. (2002). Fetal development in the pig in relation to genetic merit for piglet survival. *Journal of Animal Science*, 80 :1759–1770.

Lefort, G., Liaubet, L., Canlet, C., Tardivel, P., Père, M.-C., Quesnel, H., Paris, A., Iannuccelli, N., Vialaneix, N., and Servien, R. (2019). ASICS : an R package for a whole analysis workflow of 1D ^{1}H NMR spectra. *Bioinformatics*, 35 :4356–4363.

Leite, D. F. B. and Cecatti, J. G. (2019). New approaches to fetal growth restriction : The time for metabolomics has come. *Revista brasileira de ginecologia e obstetricia*, 41 :454–462.

Liang, L., Guo, W. H., Esquiliano, D. R., Asai, M., Rodriguez, S., Giraud, J., Kushner, J. A., White, M. F., and Lopez, M. F. (2010). Insulin-like growth factor 2 and the insulin receptor, but not insulin, regulate fetal hepatic glycogen synthesis. *Endocrinology*, 151 :741–747.

Lin, G., Liu, C., Feng, C., Fan, Z., Dai, Z., Lai, C., Li, Z., Wu, G., and Wang, J. (2012). Metabolomic analysis reveals differences in umbilical vein plasma metabolites between normal and growth-restricted fetal pigs during late gestation. *The Journal of Nutrition*, 142 :990–998.

Lin, G., Wang, X., Wu, G., Feng, C., Zhou, H., Li, D., and Wang, J. (2014). Improving amino acid nutrition to prevent intrauterine growth restriction in mammals. *Amino Acids*, 46 :1605–1623.

Liu, N., Dai, Z., Zhang, Y., Chen, J., Yang, Y., Wu, G., Tso, P., and Wu, Z. (2019). Maternal L-proline supplementation enhances fetal survival, placental development, and nutrient transport in mice. *Biology of Reproduction*, 100(4) :1073–1081.

Marchand, J., Martineau, E., Guitton, Y., Dervilly-Pinel, G., and Giraudeau, P. (2017). Multidimensional nmr approaches towards highly resolved, sensitive and high-throughput quantitative metabolomics. *Current opinion in biotechnology*, 43 :49–55.

Mariette, J. and Villa-Vialaneix, N. (2018). Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, 34 :1009–1015.

Martin, M., Legat, B., Leenders, J., Vanwinsberghe, J., Rousseau, R., Boulanger, B., Eilers, P. H. C., De Tullio, P., and Govaerts, B. (2018). Pepsnmr for 1h nmr metabolomic data pre-processing. *Analytica Chimica Acta*, 1019 :1–13.

Mellor, D. J. and Cockburn, F. (1986). A comparison of energy metabolism in the new-born infant, piglet and lamb. *Quarterly Journal of Experimental Physiology*, 71 :361–379.

Mercier, P., Lewis, M. J., Chang, D., Baker, D., and Wishart, D. S. (2011). Towards automatic metabolomic profiling of high-resolution one-dimensional proton nmr spectra. *Journal of biomolecular NMR*, 49 :307–323.

Misra, B. B. (2018). New tools and resources in metabolomics : 2016-2017. *Electrophoresis*, 39(7) :909–923.

Misra, S. and Oliver, N. (2015). Utility of ketone measurement in the prevention, diagnosis and management of diabetic ketoacidosis. *Diabetic Medicine*, 32(1) :14–23.

Mota-Rojas, D., Orozco-Gregorio, H., Villanueva-Garcia, D. Bonilla-Jaime, H., Suarez-Bonilla, X., Hernandez-Gonzalez, R., Roldan-Santiago, P., and Trujillo-Ortega, M. (2011). Foetal and neonatal energy metabolism in pigsand humans : a review. *Veterinarni Medicina*, 5(56) :215–225.

Nicholson, J. K., Lindon, J. C., and Holmes, E. (1999). 'metabonomics' : understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological nmr spectroscopic data. *Xenobiotica*, 29 :1181–1189.

Nicholson, J. K. and Wilson, I. D. (1989). High resolution proton magnetic resonance spectroscopy of biological fluids. *Prog. Nucl. Magn. Reson. Spectrosc.*, 21(4-5) :449–501.

Nissen, P. M., Nebel, C., Oksbjerg, N., and Bertram, H. C. (2011). Metabolomics reveals relationship between plasma inositols and birth weight : possible markers for fetal programming of type 2 diabetes. *Journal of Biomedicine & Biotechnology*, 2011.

Orczyk-Pawilowicz, M., Jawien, E., Deja, S., Hirnle, L., Zabek, A., and Mlynarz, P. (2016). Metabolomics of human amniotic fluid and maternal plasma during normal pregnancy. *PloS One*, 11 :e0152740.

Parimi, P. S., Cripe-Mamie, C., and Kalhan, S. C. (2004). Metabolic responses to protein restriction during pregnancy in rat and translation initiation factors in the mother and fetus. *Pediatric Research*, 56 :423–431.

Patti, G. J., Yanes, O., and Siuzdak, G. (2012). Innovation : Metabolomics : the apogee of the omics trilogy. *Nature Reviews. Molecular Cell Biology*, 13 :263–269.

Piedrahita, J. A. (2011). The role of imprinted genes in fetal growth abnormalities. *Birth Defects Research*, 91 :682–692.

Pohlert, T. (2014). *The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR)*. R package.

R Core Team (2019). *R : A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ravanbakhsh, S., Liu, P., Bjordahl, T., Mandal, R., Grant, J., Wilson, M., Eisner, R., Sinelnikov, I., Hu, X., Luchinat, C., Greiner, R., and Wishart, D. (2015). Accurate, fully-automated NMR spectral profiling for metabolomics. *PLOS ONE*, 10(5) :e0124219.

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017). mixomics : An r package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13 :e1005752.

Salek, R., Maguire, M., Bentley, E., Rubtsov, D., Hough, T., Cheeseman, M., Nunez, D., Sweatman, B., Haselden, J., Cox, R., Connor, S., and Griffin, J. (2007). A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. *Physiological Genomics*, 29(2) :99–108.

Salway, J. G. (2017). *Metabolism at a Glance, 4th Edition*. John Wiley & Sons, Ltd, Chichester, UK.

Sanz-Cortés, M., Carbajo, R. J., Crispi, F., Figueras, F., Pineda-Lucena, A., and Gratacós, E. (2013). Metabolomic profile of umbilical cord blood plasma from early and late intrauterine growth restricted (IUGR) neonates with and without signs of brain vasodilation. *PloS One*, 8 :e80121.

Savorani, F., Tomasi, G., and Engelsen, S. (2010). icoshift : a versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2) :190–202.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464.

Simon, N., Friedman, J. H., and Hastie, T. (2013). A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. Preprint arXiv 1311.6529v1.

Smolinska, A., Blanchet, L., Buydens, L. M. C., and Wijmenga, S. S. (2012). Nmr and pattern recognition methods in metabolomics : from data acquisition to biomarker discovery : a review. *Analytica chimica acta*, 750 :82–97.

Solberg, R., Enot, D., Deigner, H.-P., Koal, T., Scholl-Bürgi, S., Saugstad, O. D., and Keller, M. (2010). Metabolomic analyses of plasma reveals new insights into asphyxia and resuscitation in pigs. *PloS One*, 5 :e9606.

Staarvik, T., Framstad, T., Heggelund, M., Brynjulvsrud Fremgaarden, S., and Kielland, C. (2019). Blood-glucose levels in newborn piglets and the associations between blood-glucose levels, intrauterine growth restriction and pre-weaning mortality. *Porcine Health Management*, 5 :22.

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J., and Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis chemical analysis working group (cawg) metabolomics standards initiative (msi). *Metabolomics : Official journal of the Metabolomic Society*, 3 :211–221.

Sun, H., Zhang, S., Zhang, A., Yan, G., Wu, X., Han, Y., and Wang, X. (2014). Metabolomic analysis of diet-induced type 2 diabetes using UPLC/MS integrated with pattern recognition approach. *PLoS ONE*, 9(3) :e93384.

Sévin, D. C., Kuehne, A., Zamboni, N., and Sauer, U. (2015). Biological insights through nontargeted metabolomics. *Current Opinion in Biotechnology*, 34 :1–8.

Tardivel, P. (2017). *Représentation parcimonieuse et procédures de tests multiples : application à la métabolomique*. PhD thesis, Universié Toulouse 3 Paul Sabatier.

Tardivel, P., Canlet, C., Lefort, G., Tremblay-Franco, M., Debrauwer, L., Concordet, D., and Servien, R. (2017). ASICS : an automatic method for identification and quantification of metabolites in complex 1D 1H NMR spectra. *Metabolomics*, 13(10) :109.

Thévenot, E. A., Roux, A., Xu, Y., Ezan, E., and Junot, C. (2015). Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *Journal of Proteome Research*, 14(8) :3322–3335.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc., Ser. B*, 58(1) :267–288.

Tofts, P. S. and Wray, S. (1988). A critical assessment of methods of measuring metabolite concentrations by nmr spectroscopy. *NMR in Biomedicine*, 1 :1–10.

Tomasi, G., Van Den Berg, F., and Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics : A Journal of the Chemometrics Society*, 18(5) :231–241.

Tredwell, G. D., Behrends, V., Geier, F. M., Liebeke, M., and Bundy, J. G. (2011). Between-person comparison of metabolite fitting for NMR-based quantitative metabolomics. *Analytical Chemistry*, 83(22) :8683–8687.

Trygg, J. and Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3) :119–128.

Tulpan, D., Léger, S., Belliveau, L., Culf, A., and Čuperlović-Culf, M. (2011). MetaboHunter : an automatic approach for identification of metabolites from 1 H-NMR spectra of complex mixtures. *BMC Bioinformatics*, 12(1) :400.

Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent Wenger, R., Yao, H., and Markley, J. L. (2008). Biomagresbank. *Nucleic Acids Research*, 36 :D402–D408.

UNICEF, WHO, World Bank Group and United Nations (2019). Levels and trends in child mortality 2019. Estimates developed by the UN Inter-agency group for child mortality estimation. Technical report, UN-IGME. https://www.unicef.org/reports/levels-and-trends-child-mortality-report-2019.

Van Bussel, F. C., Backes, W. H., Hofman, P. A., Puts, N. A., Edden, R. A., Van Boxtel, M. P., Schram, M. T., Stehouwer, C. D., Wildberger, J. E., and Jansen, J. F. (2016). Increased GABA concentrations in type 2 diabetes mellitus are related to lower cognitive functioning. *Medicine*, 95(36).

van der Lende, T., Knol, E. F., and Leenhouwers, J. I. (2001). Prenatal development as a predisposing factor for perinatal losses in pigs. *Reproduction Supplement*, 58 :247–261.

Voillet, V., San Cristobal, M., Père, M.-C., Billon, Y., Canario, L., Liaubet, L., and Lefaucheur, L. (2018). Integrated analysis of proteomic and transcriptomic data highlights late fetal muscle maturation process. *Molecular & Cellular Proteomics*, 17 :672–693.

Voillet, V., SanCristobal, M., Lippi, Y., Martin, P. G. P., Iannuccelli, N., Lascor, C., Vignoles, F., Billon, Y., Canario, L., and Liaubet, L. (2014). Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity. *BMC Genomics*, 15 :797.

Vu, T., Valkenborg, D., Smets, K., Verwaest, K., Dommisse, R., Lemière, F., Verschoren, A., Goethals, B., and Laukens, K. (2011). An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics*, 12(1) :405.

Wang, J., Feng, C., Liu, T., Shi, M., Wu, G., and Bazer, F. W. (2017). Physiological alterations associated with intrauterine growth restriction in fetal pigs : Causes and insights for nutritional optimization. *Molecular Reproduction and Development*, 84 :897–904.

Wang, K., Wang, S., Kuo, C., and Tseng, Y. (2013a). Distribution-based classification method for baseline correction of metabolomic 1D proton nuclear magnetic resonance spectra. *Analytical Chemistry*, 85(2) :1231–1239.

Wang, W., Wu, Z., Dai, Z., Yang, Y., Wang, J., and Wu, G. (2013b). Glycine metabolism in animals and humans : implications for nutrition and health. *Amino Acids*, 45 :463–477.

Weljie, A., Newton, J., Mercier, P., Carlson, E., and Slupsky, C. (2006). Targeted profiling : quantitative analysis of 1 H NMR metabolomics data. *Analytical Chemistry*, 78(13) :4430–4442.

Wellner, V. P., Sekura, R., Meister, A., and Larsson, A. (1974). Glutathione synthetase deficiency, an inborn error of metabolism involving the gamma-glutamyl cycle in patients with 5-oxoprolinuria (pyroglutamic aciduria). *Proceedings of the National Academy of Sciences of the United States of America*, 71 :2505–2509.

Wishart, D. S. (2007). Current progress in computational metabolomics. *Briefings in bioinformatics*, 8 :279–293.

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D. D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., Macinnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B. D., Vogel, H. J., and Querengesser, L. (2007). Hmdb : the human metabolome database. *Nucleic Acids Research*, 35 :D521–D526.

Wong, J., Durante, C., and Cartwright, H. (2005). Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Analytical Chemistry*, 77(17) :5655–5661.

Wu, G., Bazer, F. W., Datta, S., Johnson, G. A., Li, P., Satterfield, M. C., and Spencer, T. E. (2008). Proline metabolism in the conceptus : implications for fetal growth and development. *Amino Acids*, 35 :691–702.

Wu, G., Bazer, F. W., Satterfield, M. C., Li, X., Wang, X., Johnson, G. A., Burghardt, R. C., Dai, Z., Wang, J., and Wu, Z. (2013). Impacts of arginine nutrition on embryonic and fetal development in mammals. *Amino Acids*, 45 :241–256.

Wu, G., Bazer, F. W., Tuo, W., and Flynn, S. P. (1996). Unusual abundance of arginine and ornithine in porcine allantoic fluid. *Biology of Reproduction*, 54 :1261–1265.

Wu, G., Bazer, F. W., Wallace, J. M., and Spencer, T. E. (2006). Board-invited review : intrauterine growth retardation : implications for the animal sciences. *Journal of Animal Science*, 84 :2316–2337.

Wu, G., Fang, Y.-Z., Yang, S., Lupton, J. R., and Turner, N. D. (2004). Glutathione metabolism and its implications for health. *The Journal of Nutrition*, 134(3) :489–492.

Xia, J. and Wishart, D. S. (2010). MetPA : a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics*, 26(18) :2342–2344.

Yao, Y., Voillet, V., Jegou, M., SanCristobal, M., Dou, S., Romé, V., Lippi, Y., Billon, Y., Père, M.-C., Boudry, G., Gress, L., Iannucelli, N., Mormède, P., Quesnel, H., Canario, L., Liaubet, L., and Le Huërou-Luron, I. (2017). Comparing the intestinal transcriptome of Meishan and Large White piglets during late fetal development reveals genes involved in glucose and lipid metabolism and immunity as valuable clues of intestinal maturity. *BMC Genomics*, 18(1) :647.

Young, F. (1957). Claude Bernard and the discovery of glycogen ; a century of retrospect. *British Medical Journal*, 1 :1431–1437.

Yousri, N. A., Mook-Kanamori, D. O., Selim, M. M. E.-D., Takiddin, A. H., Al-Homsi, H., Al-Mahmoud, K. A. S., Karoly, E. D., Krumsiek, J., Do, K. T., Neumaier, U., Mook-Kanamori, M. J., Rowe, J., Chidiac, O. M., McKeon, C., Al Muftah, W. A., Kader, S. A., Kastenmüller, G., and Suhre, K. (2015). A systems view of type 2 diabetes-associated metabolic perturbations in saliva, blood and urine at different timescales of glycaemic control. *Diabetologia*, 58(8) :1855–1867.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Royal Stat. Soc., Ser. B*, 68 :49–67.

Zhang, S., Nagana Gowda, G., Ye, T., and Raftery, D. (2010). Advances in NMR-based biofluid analysis and metabolite profiling. *Analyst*, 135(7) :1490–1498.