



HAL
open science

Approche intégrative du développement musculaire afin de décrire le processus de maturation en lien avec la survie néonatale

Valentin Voillet

► **To cite this version:**

Valentin Voillet. Approche intégrative du développement musculaire afin de décrire le processus de maturation en lien avec la survie néonatale. Sciences agricoles. Institut National Polytechnique de Toulouse - INPT, 2016. Français. NNT : 2016INPT0067 . tel-02797293v2

HAL Id: tel-02797293

<https://hal.inrae.fr/tel-02797293v2>

Submitted on 19 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Université
de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (INP Toulouse)

Discipline ou spécialité :

Pathologie, Toxicologie, Génétique et Nutrition

Présentée et soutenue par :

M. VALENTIN VOILLET

le jeudi 29 septembre 2016

Titre :

APPROCHE INTEGRATIVE DU DEVELOPPEMENT MUSCULAIRE AFIN
DE DECRIRE LE PROCESSUS DE MATURATION EN LIEN AVEC LA
SURVIE NEONATALE

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Génétique, Physiologie et Systèmes d'Elevage (GenPhySE)

Directeur(s) de Thèse :

MME MAGALI SAN CRISTOBAL

MME LAURENCE LIAUBET

Rapporteurs :

M. MARIE-LAURE MARTIN-MAGNIETTE, AGROPARISTECH

Mme KIM-ANH LE CAO, UNIVERSITY OF QUEENSLAND

Membre(s) du jury :

M. HERVE REMIGNON, INP TOULOUSE, Président

M. JEAN-FRANÇOIS HOCQUETTE, INRA SAINT GENES CHAMPANELLE, Membre

M. LOUIS LEFAUCHEUR, INRA SAINT GILLES, Membre

Mme ANDREA RAU, INRA JOUY EN JOSAS, Membre

Mme LAURENCE LIAUBET, INRA TOULOUSE, Membre

Mme MAGALI SAN CRISTOBAL, INRA TOULOUSE, Membre

Résumé

Depuis plusieurs années, des projets d'intégration de données omiques se sont développés, notamment avec objectif de participer à la description fine de caractères complexes d'intérêt socio-économique. Dans ce contexte, l'objectif de cette thèse est de combiner différentes données omiques hétérogènes afin de mieux décrire et comprendre le dernier tiers de gestation chez le porc, période influençant la mortalité porcine. Durant cette thèse, nous avons identifié les bases moléculaires et cellulaires sous-jacentes de la fin de gestation, en particulier au niveau du muscle squelettique. Ce tissu est en effet déterminant à la naissance car impliqué dans l'efficacité de plusieurs fonctions physiologiques comme la thermorégulation et la capacité à se déplacer. Au niveau du plan expérimental, les tissus analysés proviennent de fœtus prélevés à 90 et 110 jours de gestation (naissance à 114 jours), issus de deux lignées extrêmes pour la mortalité à la naissance, Large White et Meishan, et des deux croisements réciproques.

Au travers l'application de plusieurs études statistiques et computationnelles (analyses multidimensionnelles, inférence de réseaux, clustering et intégration de données), nous avons montré l'existence de mécanismes biologiques régulant la maturité musculaire chez les porcelets, mais également chez d'autres espèces d'intérêt agronomique (bovin et mouton). Quelques gènes et protéines ont été identifiées comme étant fortement liées à la mise en place du métabolisme énergétique musculaire durant le dernier tiers de gestation. Les porcelets ayant une immaturité du métabolisme musculaire seraient sujets à un plus fort risque de mortalité à la naissance.

Un second volet de cette thèse concerne l'imputation de données manquantes (tout un groupe de variables pour un individu) dans les méthodes d'analyses multidimensionnelles, comme l'analyse factorielle multiple (AFM) (ou *multiple factor analysis* (MFA)). Dans notre contexte, l'AFM fut particulièrement intéressante pour l'intégration de données d'un ensemble d'individus sur différents tissus (deux ou plus). Afin de conserver ces individus manquants pour tout un groupe de variables, nous avons développé une méthode, appelée MI-MFA (multiple imputation - MFA), permettant l'estimation des composantes de l'AFM pour ces individus manquants.

Mots-clés : intégration de données omiques, réseaux biologiques, analyses multidimensionnelles, porc, maturité, mortalité néonatale

Abstract

Over the last decades, some omics data integration studies have been developed to participate in the detailed description of complex traits with socio-economic interests. In this context, the aim of the thesis is to combine different heterogeneous omics data to better describe and understand the last third of gestation in pigs, period influencing the piglet mortality at birth. In the thesis, we better defined the molecular and cellular basis underlying the end of gestation, with a focus on the skeletal muscle. This tissue is specially involved in the efficiency of several physiological functions, such as thermoregulation and motor functions. According to the experimental design, tissues were collected at two days of gestation (90 or 110 days of gestation) from four fetal genotypes. These genotypes consisted in two extreme breeds for mortality at birth (Meishan and Large White) and two reciprocal crosses.

Through statistical and computational analyses (descriptive analyses, network inference, clustering and biological data integration), we highlighted some biological mechanisms regulating the maturation process in pigs, but also in other livestock species (cattle and sheep). Some genes and proteins were identified as being highly involved in the muscle energy metabolism. Piglets with a muscular metabolism immaturity would be associated with a higher risk of mortality at birth.

A second aspect of the thesis was the imputation of missing individual row values in the multidimensional statistical method framework, such as the multiple factor analysis (MFA). In our context, MFA was particularly interesting in integrating data coming from the same individuals on different tissues (two or more). To avoid missing individual row values, we developed a method, called MI-MFA (multiple imputation - MFA), allowing the estimation of the MFA components for these missing individuals.

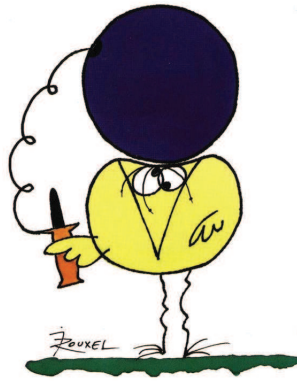
Keywords: omics data integration, biological networks, multidimensional analysis, pigs, maturity, neonatal mortality

Laboratoire de Génétique, Physiologie et Systèmes d'Élevage,
UMR1388 GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet Tolosan - France



Financement: Région Midi-Pyrénées Languedoc-Roussillon, Département de Génétique Animale (GA),
Département de Physiologie Animale et Système d'Élevage (PHASE)

Les devises Shadok



EN ESSAYANT CONTINUELLEMENT
ON FINIT PAR RÉUSSIR. DONC:
PLUS ÇA RATE, PLUS ON A
DE CHANCES QUE ÇA MARCHE.

Remerciement

Je tiens tout d'abord à remercier mes deux directrices de thèse, Magali San Cristobal et Laurence Liaubet, pour leur encadrement de très grande qualité, alliance idéale entre statistique et biologie, avec qui j'ai eu énormément de plaisir à travailler durant cette thèse, et auparavant pendant mon stage de fin d'étude. Merci pour ce que vous m'avez apporté professionnellement et humainement durant ces trois années, et la confiance que vous m'avez accordée.

Je remercie par ailleurs les rapporteurs de cette thèse, Kim-Anh Lê Cao et Marie-Laure Martin-Magniette pour l'intérêt qu'ils ont aimablement porté à mon travail. Un grand merci également à Jean-François Hocquette, Andrea Rau, Hervé Remignon et Louis Lefaucheur pour avoir accepté de faire partie du jury.

Je souhaite aussi remercier Laurianne Canario, Nathalie Viguerie, Nathalie Villa-Vialaneix, Alain Paris et Louis Lefaucheur, membres de mon comité de thèse, pour m'avoir conseillé durant ces trois années.

Je remercie ensuite l'ensemble des membres de mon laboratoire d'accueil, en particulier les membres des équipes GenoRobust et DYNAGEN. Merci aussi au groupe de travail biopuce pour leurs discussions toujours intéressantes. Merci aux fabuleux gestionnaires de l'unité ! Merci à tous les stagiaires et doctorants croisés durant cette thèse, avec une pensée particulière pour Maria, toujours motivée et de bonne humeur ! Un très grand merci à mes amis de l'INRA : Manu, Nathalie, Maguy, Yann, Manuela, Valérie et Léo, Morgane et Maël, Marjorie, Pauline et Yves, Jason, Mathilde, Maxime, Laure, Maria et Steph', Lisa copine de concert, (et tous ceux que j'ai pu oublier...) pour ces moments au bureau et à l'extérieur, vous allez me manquer !

Un grand merci à Yvette Lhabib-Manset, Nathalie Villa-Vialaneix et Ignacio González. J'ai beaucoup apprécié travailler avec vous sur les réseaux, Nathalie et Yvette, et sur ces fameuses données manquantes avec toi Ignacio ! J'espère que nous aurons tous l'occasion de continuer à travailler ensemble !

Bien entendu, un immense merci à Louis Lefaucheur avec qui nous avons fortement collaboré. Je suis très satisfait de cette coopération qui a été très enrichissante !

J'aimerais aussi remercier l'équipe avec laquelle j'ai travaillé durant ma mobilité de 4 mois au CSIRO à Brisbane. *First, I would like to thank Brian Dalrymple and his team, particularly Quan, Raidong, Bryce and Aaron for their welcome and help during my stay. It was a real pleasure to work in such a high quality environment!*

J'ajoute un remerciement à Dominique Pantalacci de l'école doctorale SEVAB pour son excellente gestion administrative.

Enfin, j'aimerais finir ces remerciements par une note plus personnelle pour mes proches. Merci à tous mes incroyables amis pour leur soutien : Sylvain et Adèle, Julien, Noémie et Antoine, Seb et Chacha (bonne chance pour votre nouvelle vie à trois), les *housemates* d'Australie (Matt, Kayo, Jenn, Karen, Stepan, John, Kat, Andy and Shiori), John, Florian, Bertrand, Jérémy et Morgane, Manu, Lisa, Maria, Laure !

Merci également à ma famille pour son soutien et son amour ! Je remercie tout particulièrement mes parents à qui je dois tant, et qui ont toujours été là pour m'encourager et me conseiller. Vous êtes des parents extraordinaires, je vous admire et suis très fier de vous ! Nous avons beaucoup de chance de vous avoir. Merci aussi à mes deux grandes soeurs et mon petit frère ! Merci aux rapportés :) et mes fantastiques neveux et nièces ! Vous êtes tous géniaux !

Pour finir, merci à Doudou pour TOUT : ton soutien, ton aide, ta patience et bien sûr pour tout ton amour ! Je suis très fier de continuer l'aventure avec toi de l'autre côté de l'Atlantique.

Table des Matières

Table des Matières	ix
Liste des figures	xii
Liste des tables	xiv
Abréviations	xvi
1 Introduction	1
1.1 Description de la fin du développement fœtal chez le porc	2
1.1.1 Un taux de mortalité élevé à la naissance	2
1.1.2 La maturité : un des facteurs impliqués dans la mortalité à la naissance	4
1.1.3 Le rôle du muscle à la naissance	8
1.2 Biologie des systèmes et intégration de données omiques	10
1.2.1 Les différents types de données omiques	10
1.2.2 Pourquoi intégrer des données omiques ?	12
1.2.3 Quelle stratégie statistique choisir ? Un cadre conceptuel pour l'intégration des données omiques	14
1.2.3.1 Intégration de type hiérarchique	15
1.2.3.2 Intégration de type méta-dimensionnelle	17
1.2.4 Limites et considérations	19
1.3 Intégration de données omiques pour mieux décrire le dernier tiers de gestation : contribution de la thèse	21
2 Analyse du transcriptome musculaire	24
2.1 Introduction	24
2.2 Article 1 : Voillet et al., <i>BMC Genomics</i> , 2014	26

2.2.1	Quelques mises à jour et commentaires	44
3	Intégration de données omiques musculaires	47
3.1	Introduction	47
3.2	Les réseaux en biologie	49
3.2.1	Introduction	49
3.2.2	Méthodes d'inférence des réseaux biologiques	50
3.2.2.1	Réseaux de type relevance network	51
3.2.2.2	Réseaux de type modèle graphique gaussien	53
3.2.3	L'algorithme PCIT (Partial Correlation with Information Theory)	56
3.2.4	Définitions et propriétés des réseaux biologiques	58
3.3	Article 2 : Voillet et al., <i>En préparation</i> , 2016	61
4	Intégration de données hétérogènes	100
4.1	Intégration de données musculaires pour différentes espèces	100
4.1.1	Introduction	100
4.1.2	Article 3 : Voillet et al., <i>En préparation</i> , 2016	102
4.2	Intégration de données issues de différents tissus pour une espèce	121
4.2.1	Introduction	121
4.2.2	Aperçu de quelques méthodes d'analyse multidimensionnelle	122
4.2.2.1	Rappels sur l'analyse en composantes principales (ACP)	123
4.2.2.2	Analyse canonique des corrélations (ACC)	124
4.2.2.3	Régression des moindres carrés partiels (PLS)	126
4.2.2.4	Analyse factorielle multiple (AFM)	128
4.2.3	Contribution de la thèse : les valeurs manquantes dans le cadre de l'AFM	131
4.2.3.1	Les valeurs manquantes en biologie	131
4.2.3.2	Article 4 : Voillet et al., <i>BMC Bioinformatics</i> , 2016	133
4.2.3.3	Illustration de la méthode MI-MFA	150
5	Discussion - Perspectives	158
5.1	Discussion biologique	159
5.1.1	Remaniement de l'expression durant la fin de gestation	160
5.1.2	Influence de la sélection génétique sur le métabolisme énergétique musculaire	162

TABLE DES MATIÈRES

5.2	Discussion méthodologique	168
5.2.1	Le choix de la méthode d'inférence des réseaux protéiques . . .	168
5.2.2	Intégration des données protéomiques et transcriptomiques : autres stratégies	171
5.2.3	Perspectives de la MI-MFA	172
	Liste des articles et communications	176
	Références	180

Liste des figures

1.1	Evaluation du nombre de porcelets nés totaux et morts par portée . . .	3
1.2	Description de la fin de gestation chez le porc	5
1.3	Schéma du développement du muscle squelettique chez le porc	9
1.4	Vue globale des systèmes biologiques	11
1.5	Vue schématique de l'intégration de type hiérarchique	15
1.6	Méthodes d'intégration multi-dimensionnelles	17
1.7	Les différents niveaux d'intégration abordés dans cette thèse	22
3.1	Exemple d'un réseau	50
3.2	Exemple de réseaux, non-orienté et orienté	50
3.3	Etapes pour la construction d'un réseau de type relevance network . . .	52
3.4	Illustration des différents types de dépendance entre variables	54
4.1	Vue schématique de l'ACP	124
4.2	Vue schématique de l'ACC	125
4.3	Vue schématique de la PLS	127
4.4	Vue schématique de l'AFM	130
4.5	Principe de la superposition des représentations de l'AFM	130
4.6	Diagrammes de Venn des individus et des variables entre les quatre transcriptomes étudiés	151
4.7	Diagrammes de Venn des variables sélectionnées entre les quatre transcriptomes	153
4.8	Visualisation de l'incertitude des individus manquants	155
4.9	Visualisation de la projection globale MI-MFA avec la projection partielle des données musculaire (A) ou hépatique (B).	156
4.10	Visualisation de la projection globale MI-MFA avec la projection partielle des données des surrénales (A) ou du sang (B).	157
5.1	Remaniement de l'expression des 12 326 sondes	161

5.2	Représentation schématique des voies des métabolismes énergétiques musculaires	166
5.3	Caractérisation des fibres musculaires	167
5.4	Expression de GPD1 au niveau génique et protéique	167
5.5	Distribution des corrélations entre variables protéiques à 110 jours de gestation	169
5.6	Exemple de corrélation entre deux variables protéiques	170
5.7	Exemple de réseau global (protéome, transcriptome et phénotypes) à 110 jours de gestation	175

Liste des tables

2.1	Nombre d'échantillons par condition du transcriptome musculaire . . .	44
4.1	Nombre de sondes déclarées différentielles par tissu après correction pour la multiplicité des tests	152

Abréviations

ACC	<u>A</u> nalyse <u>c</u> anonique des <u>c</u> orrélations
ACOM	<u>A</u> nalyse <u>c</u> o-inertie <u>m</u> ultiple
ACO	<u>A</u> nalyse <u>c</u> o-inertie
ACP	<u>A</u> nalyse en <u>c</u> omposante <u>p</u> ricipale
Acyl-CoA	Hydroxy- <u>a</u> cy-l- <u>C</u> oA déshydrogénase
ADN	<u>A</u> cide <u>d</u> éoxyribo <u>n</u> ucléique
AFM	<u>A</u> nalyse <u>f</u> actorielle <u>m</u> ultiple
ARN	<u>A</u> cide <u>r</u> ibo <u>n</u> ucléique
ATP5A1	<u>A</u> TP synthase, H ⁺ transporting, mitochondrial F1 complex, <u>a</u> lpha subunit <u>1</u>
ATP	<u>A</u> denosine <u>t</u> ri <u>p</u> hosphate
CCA	<u>C</u> anonical <u>c</u> orrelation <u>a</u> nalysis
ChIP	<u>C</u> hromatin <u>i</u> mmunoprecipitation
CKMT2	<u>C</u> reatine <u>k</u> inase, <u>m</u> itochondrial <u>2</u>
CPT1	<u>C</u> arnitine <u>p</u> almitoyl- <u>t</u> ransférase <u>1</u>
CS	<u>C</u> itrate <u>s</u> ynthase
DPI	<u>D</u> ata <u>p</u> rocessing <u>i</u> nequality
ESR1	<u>E</u> strogen <u>r</u> eceptor <u>1</u>
GLASSO	<u>G</u> raphical <u>l</u> east <u>a</u> bsolute <u>s</u> hrinkage and <u>s</u> election <u>o</u> perator
GPD1	<u>G</u> lycerol-3- <u>p</u> hosphate <u>d</u> ehydrogenase <u>1</u>
GS	<u>G</u> lycogen <u>s</u> ynthase
GTTT	<u>G</u> estion <u>t</u> echnique des <u>t</u> roupeaux de <u>t</u> ruies
IFIP	<u>I</u> nstitut <u>f</u> rançais du <u>p</u> orc
IGF2	<u>I</u> nsulin-like <u>g</u> rowth <u>f</u> actor <u>2</u>
LASSO	<u>L</u> east <u>a</u> bsolute <u>s</u> hrinkage and <u>s</u> election <u>o</u> perator
LDHA	<u>L</u> actate <u>d</u> éshydrogénase (<u>A</u>
LDH	<u>L</u> actate <u>d</u> éshydrogénase

LW	<u>L</u> arge <u>w</u> hite
MAGEL2	<u>M</u> elanoma <u>a</u> ntigen family <u>L</u> 2
MFA	<u>M</u> ultiple <u>f</u> actor <u>a</u> nalysis
miARN	<u>M</u> icro <u>a</u> cide <u>r</u> ibonucléique
MI	<u>M</u> ultiple <u>i</u> mputation
MS	<u>M</u> eishan
MyHC	<u>M</u> yosin <u>h</u> eavy <u>c</u> hain
NIPALS	<u>N</u> on-linear <u>i</u> terative <u>p</u> artial <u>l</u> east <u>s</u> quares
PCIT	<u>P</u> artial <u>c</u> orrelation <u>i</u> nformation <u>t</u> heory
PCK2	<u>P</u> hosphoenolpyruvate <u>c</u> arboxy <u>k</u> inase <u>2</u> , mitochondrial
PGK1	<u>P</u> hosphoglycerate <u>k</u> inase <u>1</u>
PLS	<u>P</u> artial <u>l</u> east <u>s</u> quare
PPARGC1A	<u>P</u> eroxisome <u>p</u> roliferator- <u>a</u> ctivated <u>r</u> eceptor <u>g</u> amma, <u>c</u> oactivator <u>1</u> <u>a</u> lpha
QTL	<u>Q</u> uantitative <u>t</u> rait <u>l</u> oci
rACC	<u>r</u> égularisée <u>A</u> nalyse <u>c</u> anonique des <u>c</u> orrélations
sCCA	sparse <u>C</u> anonical <u>c</u> orrelation <u>a</u> nalysis
SNP	<u>S</u> ingle <u>n</u> ucleotide <u>p</u> olymorphism
sPLS	sparse <u>P</u> artial <u>l</u> east <u>s</u> quare
STATIS	<u>S</u> tructuraton des <u>t</u> ableaux <u>à</u> <u>t</u> rois <u>i</u> ndices de la <u>s</u> tatistique
WGCNA	<u>W</u> eight <u>g</u> ene <u>c</u> o-expression <u>n</u> etwork <u>a</u> nalysis
YS	<u>Y</u> orkshire

Chapitre 1

Introduction

Depuis plusieurs années, le développement des biotechnologies a permis d'évoluer d'études très ciblées vers des recherches prenant en compte l'intégralité d'un système biologique. Il est possible de mesurer l'information contenue dans des dizaines de milliers de variables biologiques (omiques), allant du gène au phénotype. Dans ce contexte de biologie systémique, des perspectives biologiques (compréhension du système) et de nouveaux défis statistiques et computationnels (déluge de données hétérogènes) s'offrent à la communauté scientifique. Des projets d'intégration de données omiques sont développés avec notamment pour objectif de participer à la description fine de caractères complexes d'intérêts socio-économiques dans le domaine agronomique. Récemment, un programme ANR PORCINET (ANR-09-GENM005) a été financé. Ce projet s'intéresse à la combinaison de données biologiques hétérogènes afin de mieux comprendre et de décrire le dernier tiers de gestation chez le porc, période influençant la mortalité porcine à la naissance. Cette mortalité représente une perte très importante pour le secteur porcin. Au cours de cette thèse, je vais détailler et préciser les bases moléculaires et cellulaires sous-jacentes de la fin de gestation à l'aide des données hétérogènes issues de ce programme ANR. Ma thèse a été effectuée au sein de l'équipe GenoRobust, de l'UMR1388 GenPhySE (Génétique, Physiologie et Systèmes d'Élevage) à l'INRA de Toulouse, en collaboration avec l'UMR1348 PEGASE (Physiologie, Environnement et Génétique pour l'Animal et les Systèmes d'Élevage) de l'INRA de Rennes. L'introduction de cette thèse sera divisée en deux parties : la première partie consistera en la description du contexte biologique étudié, ici le processus de maturité en lien avec la survie néonatale chez le porc, alors que la seconde partie sera dédiée à l'état de l'art des

méthodes et stratégies existantes pour l'intégration des données omiques.

1.1 Description de la fin du développement foetal chez le porc

1.1.1 Un taux de mortalité élevé à la naissance

Durant les dernières décennies, pour répondre aux demandes croissantes en production porcine, les objectifs de la sélection ont été d'augmenter la prolificité et/ou le contenu en viande maigre. Malheureusement, cette sélection s'est accompagnée d'une augmentation substantielle de la mortalité périnatale des porcelets chez le porc domestique (*Sus scrofa* (Canario, 2006)). Tuchscherer *et al.* (2000) ont ainsi constaté que 80% de la mortalité des porcelets intervenaient durant la période périnatale ; période allant de la mise bas jusqu'à 48-72h après la naissance selon les études. En France, au cours des dernières décennies, l'augmentation de plus de trois porcelets par truie et par an fut accompagnée d'une augmentation de la mortalité de 10 à 20% (Figure 1.1). En 2014, l'Institut Français du Porc (IFIP, données de la Gestion Technique des Troupeaux de Truies GTTT) indiquait que la mortalité des porcelets s'élevait à 20%. Cette tendance s'observe également à travers le monde où la mortalité pré-sevrage des porcelets dans les troupeaux commerciaux de porcs varie entre 10 à 20% (Kirkden *et al.*, 2013), comme au Royaume-Uni avec 11,8% de décès de porcelets durant les premières 72h (Baxter *et al.*, 2008) ou encore au Japon avec 10.7% en moyenne de décès pré-sevrages (Koketsu *et al.*, 2006). La mortalité périnatale n'affecte pas uniquement l'industrie porcine, mais également d'autres mammifères. Chez l'homme, le risque de décès juste après la naissance reste relativement élevé. En 2013, 44% des 6 millions de décès d'enfants au cours de leurs 5 premières années de vie ont été dus à des problèmes durant la période néonatale (jusqu'à 1 mois après la naissance, UNICEF (2014)). Les problèmes de prématurité sont d'ailleurs la deuxième catégorie majeure de causes de décès néonatal des nouveau-nés, après les infections (Lawn *et al.*, 2005). Trois-quarts des morts néonatales se produisent durant la première semaine de vie (Lawn *et al.*, 2005). Par ailleurs, chez le mouton, Miller *et al.* (2010) soulignent que 20 à 30% des agneaux meurent à la naissance. Ce fait est d'ailleurs reconnu comme étant l'un des problèmes clés au niveau de l'efficacité de reproduction de cette espèce. Ces

constatations invitent donc à penser que l'adaptation à la vie extra-utérine est un enjeu majeur pour la survie des mammifères.

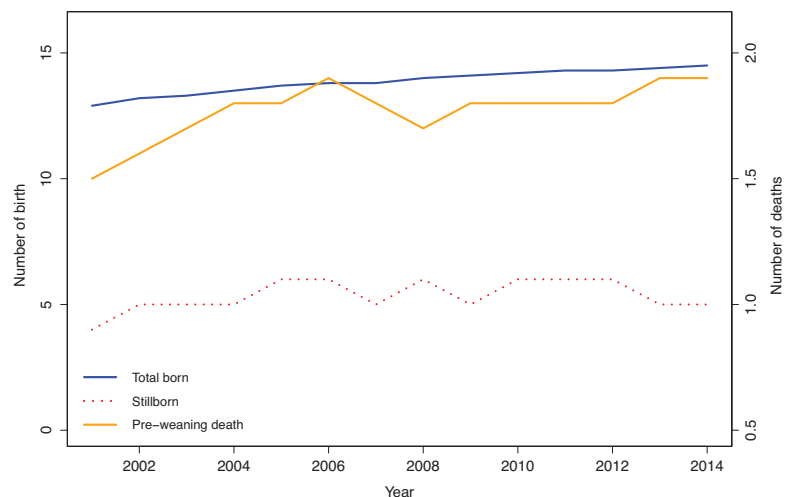


Figure 1.1 – **Evaluation du nombre de porcelets nés totaux et morts par portée au cours de la dernière décennie.** En bleu, le nombre de nés totaux selon l'axe des ordonnées de gauche. En orange et rouge, respectivement le nombre de morts pré-sevrage et le nombre de mort-nés selon l'axe des ordonnées de droite (source : IFIP).

Il a donc déjà été démontré que la mortalité des porcelets survient principalement au cours des premières heures après la naissance (Strange *et al.*, 2013). Cette mortalité des porcelets affecte fortement l'économie de l'industrie porcine, car le porc est l'une des plus importantes espèces productrices de viande dans le monde. Elle est aussi une préoccupation éthique, notamment sur le plan du bien-être animal, les consommateurs se sentant aujourd'hui de plus en plus concernés par la qualité de vie des animaux (Tuchscherer *et al.*, 2000; Serenius & Stalder, 2007; Edwards, 2011). Par exemple, au Danemark, le ministère de l'Agriculture a déclaré inadmissible une mortalité de près de 25% des porcelets avant sevrage. En outre, la législation européenne considère inacceptables les problèmes liés au bien-être des animaux en élevage et la commission européenne impose notamment aux propriétaires d'animaux de prendre «toutes les mesures appropriées en vue de garantir le bien-être de leurs animaux, ainsi que de veiller à ce que les animaux ne soient pas soumis à des douleurs, souffrances ou dommages inutiles, y compris au cours des processus de reproduction». Par conséquent, la mortalité des porcelets à la naissance altérant l'image du secteur porcin, l'attente des producteurs est grande autour de la description des mécanismes impliqués dans la néomortalité des porcelets et des mammifères en général.

1.1.2 La maturité : un des facteurs impliqués dans la mortalité à la naissance

Divers facteurs contribuant à la survie des porcelets à la naissance ont déjà été identifiés. Ils dépendent des aptitudes maternelles de la truie (durée de mise bas, santé ou comportement de la truie), des caractéristiques inhérentes au porcelet lui-même (poids et taille à la naissance, génotype ou vitalité), ou encore des interactions extérieures (nourriture, température, etc.) (van der Lende *et al.*, 2001; Baxter *et al.*, 2008; Panzardi *et al.*, 2013). La maturité est l'un de ces facteurs. Le processus de maturation mène à l'état de plein développement du porcelet, permettant ainsi sa survie à la naissance. Il se déroule durant le dernier tiers de la gestation, entre 90 jours et 110 jours (Figure 1.2), la mise bas se situant à 114 jours (Leenhouwers *et al.*, 2002; Canario, 2006; Foxcroft *et al.*, 2006). Leenhouwers *et al.* (2002) ont souligné que les différences biologiques entre différents génotypes pour la survie à la naissance pourraient être déterminées par des différences entre ces génotypes lors de la fin du développement foetal, ces différences étant reliées à la capacité du porcelet à s'adapter à de nombreuses variations environnementales associées à la transition entre la vie intra-utérine et extra-utérine. Les situations, plus ou moins avancées de la fin du développement foetal et de la maturité, sont donc des facteurs importants pour la mortalité périnatale.

La fin du développement foetal est caractérisée par différents critères tels que la taille, le poids à la naissance, les caractéristiques des organes et les réserves énergétiques comme le glycogène et les lipides (Leenhouwers *et al.*, 2001, 2002; van der Lende *et al.*, 2001). La maturité est également couplée à l'efficacité de plusieurs fonctions physiologiques comme la thermorégulation, la capacité à se nourrir et se déplacer (van der Lende *et al.*, 2001). Actuellement, le poids à la naissance est très utilisé pour caractériser la survie néonatale. Il peut notamment varier considérablement entre les porcelets d'une même portée ; la sélection génétique a probablement induit une hétérogénéité des poids au sein de la portée à la naissance (Quesnel *et al.*, 2008). Il est suspecté que cette hétérogénéité serait liée à une augmentation de la mortinatalité (Milligan *et al.*, 2002). Les porcelets de faible poids ont de moins bonnes chances de survie, alors que les porcelets de poids élevé sont moins susceptibles de mourir lors de la mise bas. Par ailleurs, les porcelets naissant avec une faible vitalité sont aussi plus susceptibles de souffrir du froid et de problèmes

de thermorégulation (Herpin *et al.*, 2002a). La vitalité, définie comme étant la vigueur ou la force physique, est directement liée à la capacité du porcelet à entrer en compétition pour un mamelon et ingérer le colostrum.

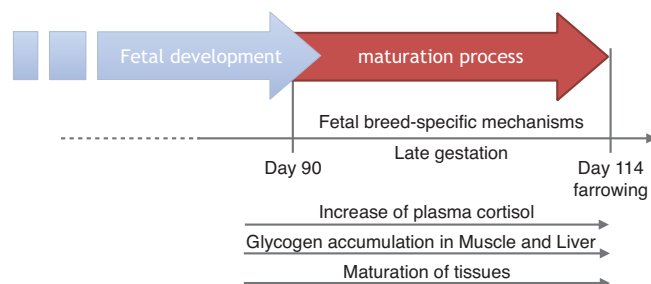


Figure 1.2 – **Description de la fin de gestation chez le porc.** Le processus de maturation a été décrit comme ayant lieu durant la fin de la gestation, entre 90 jours et la naissance (environ 114 jours). Certains mécanismes ont été identifiés durant la fin de gestation, tels qu’une augmentation du cortisol plasmatique, l’accumulation du glycogène dans le muscle et le foie, ainsi que la maturation des différents tissus. Des résultats expérimentaux ont également démontré que le génome foetal semblerait avoir une influence en fin de développement (Leenhouwers *et al.*, 2002; Canario, 2006; Foxcroft *et al.*, 2006).

Bien que le poids à la naissance ait une influence sur la survie périnatale et sur les performances de croissance du porcelet par la suite, certaines études suggèrent qu’il n’est pas le seul indicateur de la maturité (van der Lende *et al.*, 2001). Par exemple, van der Lende *et al.* (2001) ont souligné que le maintien de l’homéostasie du glucose au sein de l’organisme, via la régulation de la glycémie par plusieurs hormones (comme l’adrénaline et le cortisol) ainsi que plusieurs organes (comme le foie et le muscle), et les réserves de glycogène sont également étroitement liées à la survie du porcelet.

Le glycogène est la principale forme de stockage des polysaccharides dans les cellules animales ; son stockage permet d’emmagasiner de l’énergie et de libérer rapidement du glucose. Chez le porc, le glycogène est principalement stocké en grande quantité dans le foie et le muscle squelettique en fin de gestation. Durant la dernière période de gestation (entre 90 et 110 jours), une augmentation du taux de cortisol plasmatique est observée, celle-ci entraînant l’accumulation du glycogène dans le muscle et le foie. Il a été observé que le cortisol, glucocorticoïde sécrété par les glandes surrénales, est corrélé avec la régulation du glycogène (Fowden *et al.*, 1985). Les réserves de glycogène dans le muscle sont très importantes chez les

mammifères à la naissance. Le rôle du glycogène est celui de réserve énergétique, réserve immédiatement utilisable afin d'assurer le bon fonctionnement de la thermorégulation durant les premières heures de vie (Mellor & Cockburn, 1986). Après la naissance, la teneur en glycogène va rapidement diminuer au niveau des deux tissus (muscle et foie), cette diminution allant jusqu'à 82% de la réserve en glycogène au niveau musculaire (Herpin *et al.*, 2002b). Le glycogène du foie est décrit comme étant déterminant pour le maintien de l'homéostasie du glucose durant la parturition, la période post-partum immédiate et les situations d'ingestions faibles ou tardives de colostrum, tandis que les réserves de glycogène du muscle ont plutôt pour fonction post-natale d'assurer la thermorégulation. Le glycogène au niveau du coeur favorise quant à lui la résistance à l'anoxie pendant la mise-bas.

La thermorégulation, permettant le maintien du corps à la bonne température, implique une compensation entre production et perte de chaleur (Herpin *et al.*, 2002a). A la naissance, indépendamment des challenges immunologique, digestif, respiratoire et nutritionnel, le porcelet doit immédiatement pouvoir réguler sa propre température corporelle suite à une chute de 15 à 20°C de la température ambiante alors que précédemment, sa température dépendait uniquement de l'environnement intra-utérin. La capacité à conserver la chaleur est d'autant plus limitée que le porcelet est sans poils et encore relativement maigre (Herpin *et al.*, 2002a). Deux mécanismes de thermorégulation sont considérés : le frissonnement, impliquant une contraction musculaire, et le non-frissonnement, celui-ci regroupant les mécanismes de production de chaleur n'engageant pas de contractions musculaires. Il est connu depuis des années que les porcelets frissonnent fortement à la naissance afin de fournir de la chaleur, contrairement aux autres mammifères possédant du tissu adipeux brun (Herpin *et al.*, 2002a). Lors d'une acclimation au froid, chez les espèces dépourvues de tissu adipeux brun, la thermorégulation par non-frissonnement permet également d'augmenter la production de chaleur. Contrairement à de nombreux autres mammifères (ayant un tissu adipeux plus abondant), le porcelet à la naissance est décrit comme en ayant très peu (Herpin *et al.*, 2002a). De petites quantités de tissu adipeux brun sont observées à partir de 3 mois. Celui-ci permet, grâce à la lipolyse, la thermogénèse chez les nouveau-nés (par exemple, chez l'homme). Ainsi, chez le porcelet à la naissance, c'est le muscle squelettique, impliqué dans des fonctions de maintien de la posture, de motricité et de stockage d'énergie, qui joue un rôle central et essentiel dans le maintien de la thermorégulation (Lefaucheur *et al.*,

2001). La contraction des muscles entraîne également la production de chaleur et dépend du volume des fibres, celui-ci étant relativement faible à la naissance (Herpin *et al.*, 2002a). Ainsi, la maturité est couplée à l'efficacité de fonctions physiologiques comme la thermorégulation via le stockage de glycogène.

Par ailleurs, durant la dernière phase de gestation, la demande en nutriments du fœtus ne cesse de s'accroître, et le fœtus peut lui-même être responsable d'une augmentation des échanges nutritionnels mère-fœtus (Leenhouwers *et al.*, 2002). Ces besoins vont donc activer des mécanismes spécifiques à la lignée, permettant d'augmenter ces échanges nutritionnels (Biensen *et al.*, 1998). Par exemple, le placenta (sa taille, son poids et sa longueur) et des facteurs d'angiogénèse sont connus pour être modulés par des facteurs maternels et fœtaux (Biensen *et al.*, 1998, 1999). Ainsi, pour augmenter les échanges mère-fœtus, il existe plusieurs stratégies selon les races porcines (Wilson *et al.*, 1998). Les Yorkshire (YS) ou les Large White (LW) augmentent la taille de leur placenta pour pouvoir augmenter la capacité d'échange mère-fœtus. Les Meishan (MS) quant à eux augmentent la vascularisation du placenta sans augmenter sa taille. Ces échanges nutritionnels sont primordiaux pour la survie du porcelet à la naissance.

Les données de Leenhouwers *et al.* (2001, 2002) et Canario (2006) indiquent aussi que les porcelets ayant une valeur génétique élevée pour la survie sont plus matures à la naissance et par la suite mieux préparés pour l'adaptation à la vie extra-utérine. Par exemple, les fœtus MS, pourtant de plus faible poids, sont dits plus matures que les fœtus LW de poids plus élevé. En effet, il a notamment été démontré que les MS sont plus robustes au froid et à la faim (Herpin *et al.*, 1993). Il semblerait que les MS aient une meilleure capacité des métabolismes énergétiques, en particulier liés aux glucocorticoïdes, que les LW à la naissance (Herpin *et al.*, 1993). La sélection pour plus de tissu musculaire, comme chez les LW, aurait pu affecter les facteurs structuraux et métaboliques qui permettent une meilleure survie à la naissance. Par conséquent, l'environnement prénatal utérin, les échanges mère-fœtus, le poids à la naissance et la fin du développement fœtal (maturité) sont des facteurs vitaux pour la survie du porcelet et peuvent entraîner des retards de croissance.

1.1.3 Le rôle du muscle à la naissance

La maturité métabolique du muscle squelettique est donc indispensable pour une bonne survie à la naissance. En effet, comme nous l'avons souligné précédemment, le muscle joue un rôle prépondérant dans la thermorégulation via le frissonnement ou non, la motricité et le poids à la naissance. La myosine est l'une des composantes principales du muscle et son interaction avec l'actine convertit l'énergie chimique en énergie mécanique, libérant ainsi de la chaleur par hydrolyse de l'ATP (Lefaucheur *et al.*, 2001). Chez de nombreux mammifères, l'ontogenèse des myofibres est un phénomène biphasique. Chez le porc, une première génération de fibres, située entre 35 et 55 jours de gestation, est suivie d'une seconde génération située entre 55 et 90 jours de gestation (Figure 1.3) (Picard *et al.*, 2002). Le nombre total de fibres est établi à environ 90 jours de gestation. A 90 - 95 jours, la maturation par hypertrophie des fibres musculaires (augmentation du diamètre et de la longueur des fibres) commence dans le but d'être en état de plein développement à la naissance. A la naissance (après 114 jours de gestation), le porcelet doit donc avoir un muscle fonctionnel pour marcher et accéder seul à la mamelle afin de s'alimenter. Durant la première semaine de vie, ses muscles subiront quelques changements : (i) une diminution du taux des chaînes lourdes de myosine (MyHC) embryonnaire et périnatale ; (ii) une augmentation des MyHC de type I et II ; (iii) une expression transitoire des MyHC α -cardiaque ; et (iv) un remaniement du métabolisme énergétique (Picard *et al.*, 2002). Les MyHC de type I sont des fibres lentes rouges oxydatives, alors que les MyHC de type II (IIa, IIx et IIb) sont des fibres plus ou moins rapides (IIb étant des fibres plus rapides que IIx et IIa), blanches et plutôt glycolytiques (IIb étant des fibres plus glycolytiques que IIx et IIa). La composition en fibres du muscle dépend de son type et de son activité.

Le muscle squelettique joue ainsi un rôle déterminant à la naissance du porcelet et constitue un pourcentage considérable du poids après la naissance (jusqu'à 60% du poids de la carcasse à l'abattage (Lebret *et al.*, 1999)). Décrire les mécanismes moléculaires de la mise en place de la maturité des porcelets à la naissance pour le muscle paraît donc important. Il s'agit également d'identifier des leviers génétiques et mieux comprendre le processus de maturation chez les lignées hautement sélectionnées pour plus de croissance musculaire. La compréhension des processus de mise en place de la maturité pourrait également avoir un impact plus général

pour d'autres mammifères (d'intérêt agronomique comme pour le mouton ou de santé publique pour l'Homme). Ainsi, dans cette thèse, une approche systémique est proposée avec génération de plusieurs données omiques à divers stades du processus de maturation. Des approches statistiques spécifiques à l'intégration de données omiques hétérogènes sont alors requises pour analyser ces différents jeux de données.

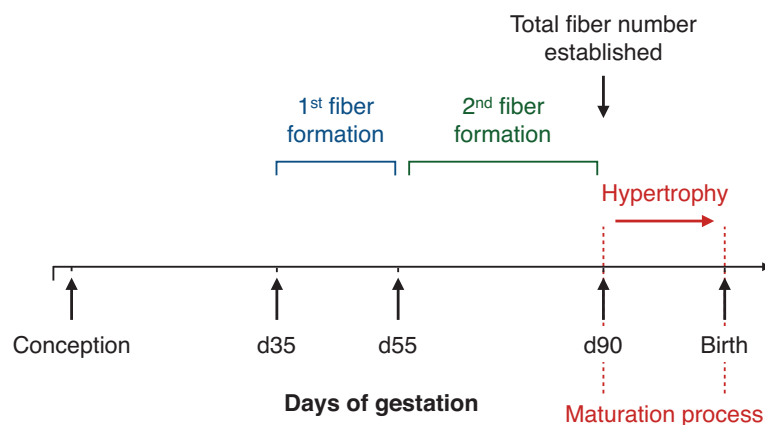


Figure 1.3 – Schéma du développement du muscle squelettique chez le porc (inspiré de Foxcroft *et al.* (2006)).

1.2 Biologie des systèmes et intégration de données omiques

1.2.1 Les différents types de données omiques

Les récentes avancées biotechnologiques permettent de mesurer une grande variété de données provenant de différentes sources cellulaires, telles que l'ADN (génomique) (Metzker, 2010), ses modifications épigénétiques comme la méthylation (épigénomique) (Bonneta, 2008; Flintoft, 2010), l'ARN (transcriptomique) (Yauk & Berndt, 2007; Wang *et al.*, 2009), les protéines (protéomique) (Blackstock & Weir, 1999; Breker & Schuldiner, 2014) ou encore les molécules de faible poids moléculaire (métabolomique) (Griffiths & Wang, 2008; Rubakhin *et al.*, 2011; Zenobi, 2013). Le nombre d'études utilisant ces technologies a constamment augmenté au cours des dernières décennies. Ces études tout génome offrent aux scientifiques une vision globale de tous les mécanismes biologiques des différents types cellulaires pour de nombreux systèmes biologiques. Ces données biologiques sont souvent hétérogènes et complexes (Figure 1.4). L'intégration de ces données a pour objectif de fournir une vision plus large, détaillée et complète des systèmes moléculaires et cellulaires, il s'agit donc d'un véritable challenge d'un point de vue de l'analyse statistique et computationnelle, autant que biologique.

Chaque type de données omiques est un domaine à part entière avec sa propre méthodologie et ses propres caractéristiques. La génomique est définie comme étant l'étude du fonctionnement d'un organisme à l'échelle du génome au lieu de se limiter à l'échelle d'un seul gène. La génomique se divise en deux branches principales : la génomique structurale, qui s'occupe du séquençage du génome entier, et pour laquelle le séquençage des génomes (bactéries, archées, eucaryotes, etc.) s'est développé ces dernières années (Liolios *et al.*, 2006) ; et la génomique fonctionnelle, qui vise à étudier la fonction et l'expression des gènes, notamment en caractérisant le transcriptome et le protéome.

L'épigénomique caractérise l'étude des facteurs épigénétiques influençant l'expression de certains gènes. La méthylation de l'ADN ou les modifications des histones sont deux processus résultant de modifications épigénétiques au niveau du génome. Ces mécanismes épigénétiques interviennent, par exemple, dans la

régulation de l'expression des gènes, notamment ceux responsables de la division, de la différenciation cellulaire et du développement, et peuvent aussi être directement impliqués dans le développement de maladies, comme le cancer ou le diabète. Le développement du séquençage à haut-débit a fourni de nombreuses applications, notamment pour l'analyse de marqueurs épigénétiques et l'analyse de la structure chromatinienne au niveau de l'ADN (ChIP-seq, methyl-seq et DNA-seq) (Wold & Myers, 2008).

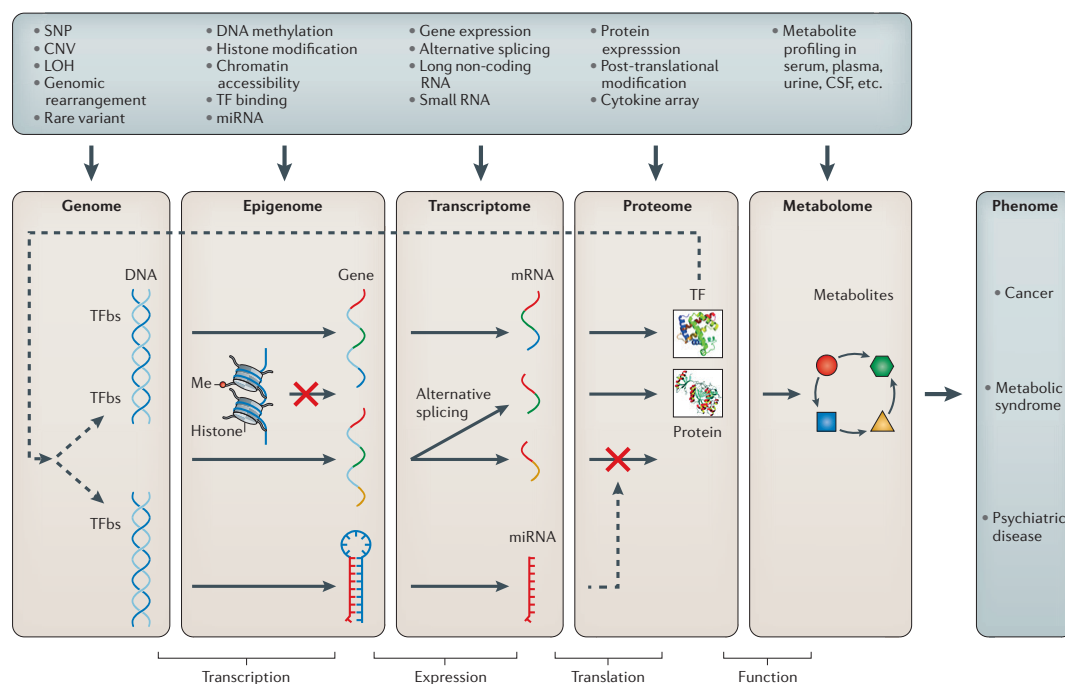


Figure 1.4 – **Vue globale des systèmes biologiques : du génome, épigénome, transcriptome, protéome et métabolome au phénotype (de Ritchie *et al.* (2015)).** Des méthodes et stratégies d'obtention de données pour chaque domaine mènent à un grand nombre de données hétérogènes. L'intégration de ces données permet d'améliorer la compréhension globale du système biologique et de décrire les attributs génétiques, moléculaires ou cellulaires influençant un phénotype.

L'analyse de l'expression de tous les transcrits, notamment les ARN messagers (ARNm), les ARN non-codants et les petits ARN, délimite le domaine de la transcriptomique. Ces transcrits peuvent être considérés comme des phénotypes intermédiaires. En effet, les variations au niveau de l'ADN contribuent, par la perturbation de l'expression des gènes, des protéines ou des métabolites, au phénotype/caractère final. Les puces à ADN (ou biopuce, *microarray* en anglais)

sont utilisées pour l'analyse des données d'expression de milliers de gènes. Le développement des méthodes de séquençage à haut-débit (RNA-seq) a également permis l'étude du transcriptome aux niveaux cellulaire, tissulaire et de l'organisme (Wang *et al.*, 2009).

La protéomique est l'étude de l'ensemble des protéines d'une cellule, d'un organe, d'un tissu, d'un organe ou d'un organisme à un moment donné et pour des conditions données, alors que la métabolomique est l'étude de l'ensemble des petites molécules (appelées métabolites - comme les sucres, les acides aminés, les acides gras, etc.). Bien qu'il existe des stratégies plus anciennes (par exemple, l'analyse par gel-2D pour l'étude de l'expression protéique), des méthodes récentes à haut-débit couplées avec de la spectrométrie de masse ont aussi été développées en protéomique ou en métabolomique afin d'identifier les protéines ou les métabolites, mais aussi d'identifier des interactions entre protéines (Altelaar *et al.*, 2013) ou entre métabolites (Shulaev, 2006). Ainsi, tous ces domaines -omiques ont donc leur propre méthode de génération de données et d'analyse, mais sont également intimement liés entre eux (Figure 1.4). Ils permettent une vision de l'expression de tout le génome à différents niveaux.

1.2.2 Pourquoi intégrer des données omiques ?

Le terme d'intégration de données se réfère à une situation où, pour un système donné, de multiples sources et types de données (ici, des données omiques) sont disponibles, le but étant de les étudier en les intégrant ensemble afin d'améliorer la compréhension globale du système (ici, des systèmes biologiques) par rapport à l'analyse d'un seul niveau d'information. De nos jours, les stratégies d'intégration de données biologiques sont prometteuses pour identifier les attributs génétiques, moléculaires ou cellulaires influençant un phénotype (Figure 1.4). De nombreuses études ont été publiées au cours des dernières années. Par exemple, des données d'expression génique ont été combinées, en utilisant des réseaux biologiques, avec des données physiologiques comme des profils d'acide gras et d'autres données cliniques afin d'analyser le rôle du tissu adipeux dans le contrôle du poids chez l'homme (Montastier *et al.*, 2015). L'objectif de cette étude était de lier des gènes avec des facteurs biologiques associés à la régulation du poids pour trouver de nouvelles caractéristiques du tissu adipeux qui permettraient d'améliorer la compréhension

du contrôle du poids en particulier chez les obèses. Comme autre illustration, des données transcriptomique et protéomique ont été intégrées afin d'observer l'effet de la faim sur une période de 72 heures chez les souris (van Iersel *et al.*, 2014). Les corrélations entre ces deux niveaux d'expression se sont révélées relativement faibles. Les auteurs ont également développé et utilisé un logiciel pour analyser l'enrichissement des voies métaboliques avec la présence de gènes et/ou de protéines. Cette étude intégrative a permis de souligner l'importance de combiner des données entre elles plutôt que d'analyser un seul niveau d'information. En effet, des effets de régulation majeurs de réponse à un mécanisme de privation alimentaire dans l'intestin chez les souris n'auraient pas été identifiés en utilisant uniquement les données protéiques.

Des projets de génomique fonctionnelle et d'intégration de données omiques ont été développés pour envisager une meilleure compréhension et peut-être permettre un meilleur contrôle des phénotypes ou caractères complexes d'intérêts socio-économiques ou médicaux. Ces caractères sont généralement contrôlés par de multiples facteurs : génétiques, environnementaux, etc. Ainsi, l'intégration de données permet de mieux comprendre les étapes se déroulant entre l'expression des gènes jusqu'au phénotype d'intérêt. Une large gamme de méthodes expérimentales et statistiques a été développée pour quantifier et intégrer les phénotypes intermédiaires, tels que les transcripts, les protéines ou les métabolites, dans les populations variant pour un caractère d'intérêt (Civelek & Lusis, 2014). En fonction du dispositif expérimental, une approche de biologie systémique consiste donc à identifier des liens entre les multiples niveaux d'information pouvant être, par exemple, l'étude de la corrélation entre différents types de données à différents niveaux, ou alors l'étude des informations génomiques par le calcul de QTL (*Quantitative Trait Loci*) ou d'association génétique. Ainsi la génétique des systèmes a pour objet de comprendre les flux d'informations biologiques qui sous-tendent les traits complexes. Ce terme (génétique des systèmes) a été proposé pour la première fois dans le contexte agronomique par Kadarmideen *et al.* (2006). Cependant, l'intégration de données ne fait pas toujours référence à de la génétique. Durant cette thèse, la composante génétique n'est que peu présente. Nous ne nous sommes pas intéressés à l'analyse de données génomiques, mais plutôt à la comparaison entre génotypes extrêmes et croisés réciproques pour la mortalité à la naissance.

1.2.3 Quelle stratégie statistique choisir ? Un cadre conceptuel pour l'intégration des données omiques

L'intégration des données omiques est un outil essentiel. Un nombre important de revues scientifiques couvrant ce sujet a été écrit au cours des dix dernières années (Reif *et al.*, 2004; Sieberts & Schadt, 2007; Hamid *et al.*, 2009; Hawkins *et al.*, 2010; Holzinger & Ritchie, 2012; Ritchie *et al.*, 2015). Bien que l'intégration de données omiques soit un terme bien défini dans la littérature, on associe à cette dénomination de nombreuses acceptions, statistiques et/ou computationnelles, voire même biologiques. Ici, nous définissons **l'intégration de données omiques comme étant la combinaison statistique et/ou computationnelle de données permettant une modélisation plus compréhensible d'un système biologique, d'un caractère complexe ou d'un phénotype, mais également permettant une identification de mécanismes, gènes, protéines, métabolites qui joueraient des rôles clés dans le système étudié.**

En premier lieu, afin d'éviter des problèmes en amont de l'intégration, il est important de contrôler et analyser chaque donnée omique de façon indépendante. De nombreux efforts sont effectués sur le nettoyage des données pour qu'elles soient prêtes à être analysées, comme la manipulation des données brutes, le contrôle qualité, la normalisation ou encore le filtrage des données (Wickham, 2014). A chaque étape, beaucoup de méthodes ont été développées, comme la normalisation de données biopuces afin de corriger d'éventuels effets techniques non-souhaités (Bolstad *et al.*, 2003), ou encore vérifier la bonne qualité de l'alignement lors d'analyses RNA-seq (Wang *et al.*, 2009). Bien que ce soit discutable sur certains points (perte d'information), Ritchie *et al.* (2015) conseillent également de procéder à une réduction de taille des données avant d'effectuer l'intégration. En effet, certaines données omiques (comme les analyses de type RNA-seq) fournissent des données avec un très grand nombre de variables. Avoir des données pré-filtrées, c'est-à-dire éviter qu'il y ait un trop grand nombre de variables, permet notamment de réduire le temps de calcul des analyses statistiques et computationnelles. La réduction du nombre de variables peut se faire par un filtrage associé à la question biologique posée : il peut être intrinsèque ou extrinsèque aux données. En revanche, il est important de noter que la réduction de données peut engendrer une perte d'information biologique. Bien que généralement un filtrage approprié permet de garder un grand nombre de variables, il est pos-

sible que le filtrage supprime aussi des variables potentiellement intéressantes. Il semble donc important que le scientifique réfléchisse à des méthodes appropriées de pré-filtrage et sur les conséquences possibles de sa réduction de données sur sa conclusion biologique.

Une fois ces contrôles effectués, deux stratégies principales peuvent être utilisées lors de l'intégration de différentes données omiques (Ritchie *et al.*, 2015) :

- **Stratégie hiérarchique** divisée en plusieurs étapes permettant de trouver des associations entre les différentes données omiques et le caractère ou phénotype d'intérêt ;
- **Stratégie méta-dimensionnelle** combinant toutes les données simultanément.

Ces deux stratégies sont décrites dans les deux parties suivantes (section 1.2.3.1 et 1.2.3.2).

1.2.3.1 Intégration de type hiérarchique

L'objectif de la stratégie d'intégration hiérarchique est de diviser l'analyse en plusieurs étapes qui se suivent de façon à déterminer les différents liens entre les données omiques, et entre les données omiques et le phénotype d'intérêt. Cette stratégie débute généralement des données génomiques jusqu'aux données phénotypiques, elle est souvent utilisée dans un contexte génétique où le scientifique espère identifier des leviers génétiques (marqueurs ou polymorphismes responsables) (Figure 1.5).

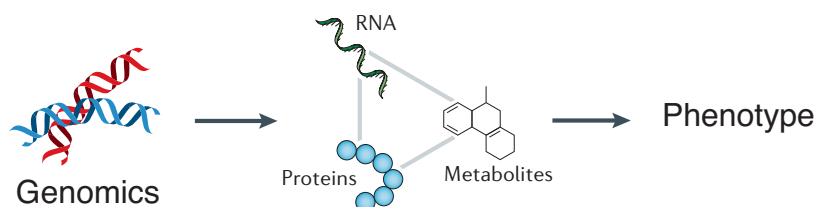


Figure 1.5 – Vue schématique de l'intégration de type hiérarchique.

Une des stratégies hiérarchiques utilisée pour l'analyse des variations génomiques est composée de trois étapes. Tout d'abord, des SNPs (*Single Nucleotide Poly-*

morphism) sont associés à un phénotype donné et sont filtrés selon un seuil de significativité. Ces SNPs sont ensuite associés avec un niveau d'expression de données omiques. Par exemple, les SNPs peuvent être combinés avec des données d'expression génique (eQTLs), des données de méthylation (metQTLs) ou encore des données d'abondance protéique (pQTLs). Pour finir, les corrélations entre le phénotype d'intérêt et les données omiques choisies sont analysées afin d'observer les liens entre ces niveaux d'expression.

Dans un organisme diploïde, pour certains gènes, un des deux allèles sera préférentiellement exprimé. Ces expressions allèles spécifiques sont associées à des modifications épigénétiques. Une autre approche de type hiérarchique pour relier les variations génomiques avec le niveau des transcripts est l'analyse d'expression allèle spécifique. En premier lieu, des données génomiques sont filtrées en fonction de leurs expressions allèles parentales (paternelle ou maternelle), puis elles sont associées avec d'autres types de données omiques, comme des données transcriptomiques ou de méthylation, afin de comparer l'expression des deux allèles. Enfin, les allèles résultantes peuvent être testées pour la corrélation avec le phénotype d'intérêt.

Cependant, les méthodes de stratégie hiérarchisée rencontrent quelques limites. En effet, cette façon d'intégrer les données ne permet pas de retour d'information d'une couche à l'autre, ce qui n'est pas toujours le cas biologiquement. Dans un système biologique, les différentes strates d'expression (transcriptome, protéome, métabolome, etc.) interagissent toutes entre-elles (Figure 1.4). Par exemple, les données transcriptomiques et protéomiques sont rarement corrélées entre elles, car des événements post-traductionnels peuvent être impliqués (Haider & Pal, 2013). A titre d'illustration, Gygi *et al.* (1999) ont trouvé que la corrélation entre les expressions génique et protéique était insuffisante et ne permettait pas de prédire les niveaux d'expression protéique à partir des niveaux d'expression génique chez la levure. Il est toutefois important de noter que la corrélation entre ces données dépend également de la qualité des données accessibles et des méthodes statistiques utilisées. Dans cette thèse, nous avons par exemple trouvé une corrélation relativement forte entre les données transcriptomique et protéomique (voir section 3.3).

1.2.3.2 Intégration de type méta-dimensionnelle

Les analyses d'intégration de données omiques méta-dimensionnelles combinent tous les différents types de données dans une analyse simultanée. Trois niveaux d'intégration existent : (i) l'intégration par la concaténation des données, (ii) l'intégration par la transformation des données et (iii) l'intégration de modèles provenant des données (Hamid *et al.*, 2009; Ritchie *et al.*, 2015). La Figure 1.6 représente ces trois niveaux spécifiques.

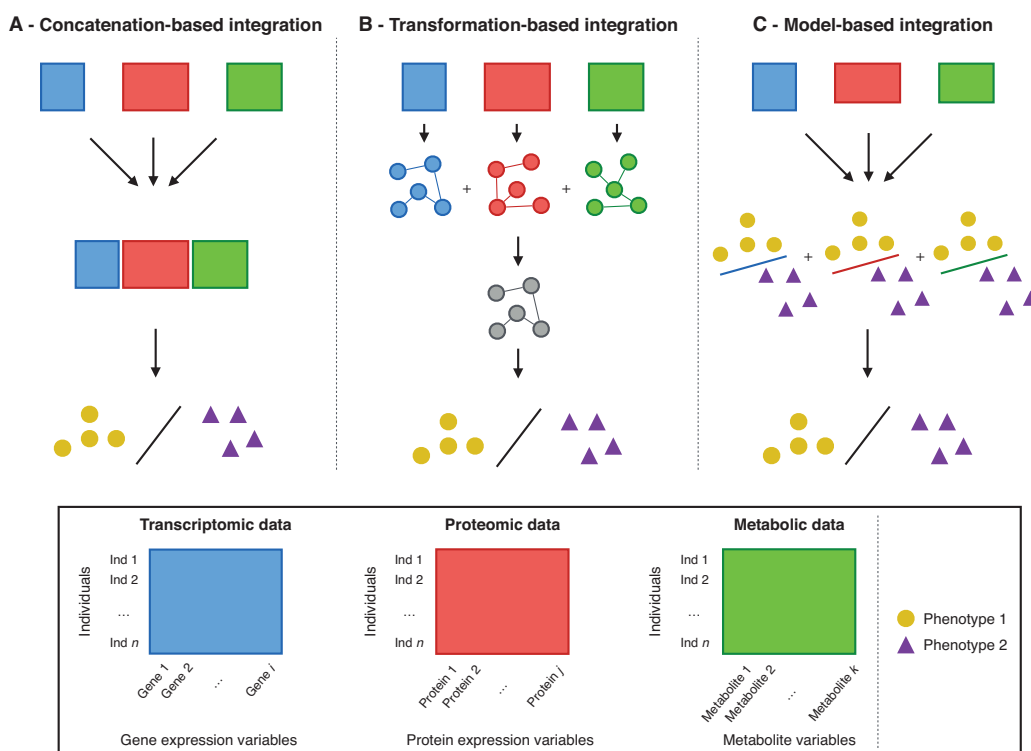


Figure 1.6 – Méthodes d'intégration multi-dimensionnelles (de Ritchie *et al.* (2015)). L'analyse d'intégration multi-dimensionnelle peut être divisée en trois catégories : (A) Intégration par la concaténation des données. (B) Intégration par la transformation des données. (C) Intégration de modèles provenant des données.

Comme son nom l'indique, l'intégration par concaténation est la simple concaténation des différentes données, afin d'effectuer des analyses statistiques et computationnelles sur les données concaténées (Figure 1.6A). Pour fusionner les données, il est primordial d'avoir les mêmes individus parmi toutes les tables de données. C'est d'ailleurs l'une des limites de cette stratégie, car il est possible que le plan expérimental ne permette pas d'avoir toutes les informations sur exactement les

mêmes individus parmi toutes les tables de données. Par ailleurs, comme souligné précédemment, une sélection de variables par filtrage est souvent conseillée avant la concaténation, afin d'être en capacité d'effectuer tout type d'analyses statistiques et computationnelles par la suite. Cette stratégie de concaténation des données est en principe assez simple et permet une observation aisée de toutes les interactions entre les différents jeux de données omiques. Cependant, une des limites importantes de cette méthode est la difficulté de combinaison de tables de données entre elles à cause de problèmes d'échelle. En effet, il peut être difficile, voire impossible, d'intégrer des données par concaténation si le type de données n'est pas équivalent ou proches (des données pouvant être qualitatives ou quantitatives). Il est également possible que, lors d'intégration de données multi-omiques par concaténation, d'être confronté à une sur-représentation des variables ayant un signal très fort (comme les transcrits (Singh *et al.*, 2016)).

La transformation des données en un intermédiaire avant l'intégration est probablement l'une des méthodes les plus utilisées pour associer les données omiques entre elles. Chaque donnée est d'abord transformée en un intermédiaire, puis ces intermédiaires sont ensuite combinés afin de construire le modèle (Figure 1.6B). Ces intermédiaires peuvent être des réseaux biologiques (Montastier *et al.*, 2015), ou encore des matrices de noyaux (Lanckriet *et al.*, 2004; Ben Hur & Noble, 2005). Cette transformation en un intermédiaire approprié permet la conservation des spécificités de chaque donnée omique. Contrairement à la concaténation des données, cette stratégie présente l'avantage de ne présenter aucun problème d'échelle entre les données, et ainsi d'intégrer des types de données différents. En revanche, le désavantage de cette méthode est qu'il peut être difficile d'identifier ces interactions entre les différents types de données omiques (comme les interactions entre SNP et expression génique), surtout si elles sont hétérogènes de nature (quantitatives, qualitatives), car la transformation séparée des données peut modifier la capacité à bien détecter les interactions, chaque donnée étant transformée indépendamment des autres. L'objectif de la transformation est donc de conserver au maximum les propriétés spécifiques des données afin de pouvoir observer ces interactions. Le choix du type d'intermédiaire (réseaux, etc.) pour l'intégration est donc déterminant et doit être pertinent selon les données à analyser.

Enfin, il est aussi possible d'intégrer des modèles statistiques provenant des données

analysées indépendamment (Figure 1.6C). Pour ce faire, un modèle statistique sera développé pour chaque type de donnée à combiner, puis un modèle, dit final, provenant de l'intégration des multiples sous-modèles, sera obtenu, préservant ainsi les caractéristiques individuelles et spécifiques. Pour utiliser cette stratégie, il est important d'avoir une hypothèse biologique pertinente et une bonne stratégie pour combiner les différents modèles entre eux. Cette stratégie peut être utilisée lorsque les données omiques sont très différentes (par exemple, des données ne portant pas sur les mêmes individus) et quand les deux stratégies précédentes sont non réalisables (par exemple, pour des raisons computationnelles ou statistiques).

1.2.4 Limites et considérations

Dans cette thèse, nous allons discuter et décrire différentes stratégies pour intégrer des données multi-omiques afin d'élucider et comprendre les systèmes biologiques sous-jacents à un caractère d'intérêt (ici, la maturité). Cependant, quelle que soit la stratégie choisie, certaines considérations sont importantes à prendre en compte avant d'effectuer l'intégration. Tout d'abord, il est indispensable d'avoir des répliquats biologiques quand l'étude porte sur des individus non consanguins, afin de prendre en compte la variabilité inter-individuelle, et permette une inférence statistique valide pour l'espèce ou la population, et non sur quelques individus particuliers.

Par ailleurs, de fréquents cas de multi-colinéarité surviennent, qui posent des problèmes d'instabilité de l'inférence statistique (Johnstone & Titterton, 2009; Verzelen, 2010). Ces cas peuvent être dus à des variables biologiques très corrélées entre elles, suite à des mécanismes biologiques intrinsèques, ou bien sont inhérents à la grande dimension, quand le nombre de variables est très supérieur au nombre d'échantillon. Quelques méthodes existent afin de pré-traiter les données, par exemple en décorrélant au préalable les variables. Il est aussi fréquent de rencontrer des problèmes de sur-apprentissage, correspondant au fait que la modélisation obtenue fonctionne bien uniquement sur les données utilisées, et non sur d'autres données indépendantes. Cela est particulièrement commun avec des données de grande dimension comme les données omiques où le nombre de variables est beaucoup plus grand que le nombre d'individus. Pour contourner ces problèmes de sur-apprentissage, il existe quelques stratégies comme des méthodes de validation croisée ou bootstrap. Toutefois, même en utilisant ces stratégies, les résultats restent

souvent biaisés. L'utilisation de jeux de données tests indépendants à l'étude permet une évaluation objective des résultats, bien qu'il est toujours possible de se heurter à des problèmes de *batch effect* (différence au niveau des plateformes, des protocoles, des lots, etc.). Il peut d'ailleurs également être intéressant de valider les résultats par l'utilisation d'outils de validation fonctionnelle, comme la fouille bibliographique ou la modélisation *in silico* avec l'utilisation des mathématiques afin de modéliser et prédire les aboutissants d'un système biologique. Par exemple, Villa-Vialaneix *et al.* (2013) se sont appuyés sur des méthodes de fouille bibliographique afin de valider ou non des méthodes de clustering.

Une autre limite importante est l'immaturation des outils statistiques et computationnels disponibles, notamment en regard de la production de données. A l'heure actuelle, les données sont générées par des techniques toujours plus performantes, mais les méthodes d'analyse statistique sont souvent encore en développement et s'avèrent non-optimales. Il faut ensuite également optimiser les stratégies d'intégration de données. Dans cette thèse, nous allons donc nous intéresser à l'intégration de données omiques afin de décrire le processus de maturation chez le porc en fin de gestation.

1.3 Intégration de données omiques pour mieux décrire le dernier tiers de gestation : contribution de la thèse

L'objectif de ma thèse est d'intégrer des données multi-omiques afin de décrire les bases moléculaires et cellulaires intervenant lors du dernier tiers de gestation et de définir de possibles marqueurs de maturité (Figure 1.7). Les données proviennent du projet ANR Porcinet qui a produit un grand nombre de données hétérogènes (transcriptome, protéome, métabolome et phénotypes) sur différents tissus (muscle, foie, glandes surrénales, tissu adipeux, intestin, urine et sang) associées à quelques phénotypes.

Au niveau du design expérimental, les tissus utilisés pour les analyses proviennent de fœtus prélevés à 90 et 110 jours de gestation, issus de deux lignées extrêmes pour la mortalité à la naissance, Large White (LW) et Meishan (MS), les fœtus MS survivant mieux à la naissance que les fœtus LW. Les truies ayant été inséminées avec un mélange de semence LW et MS, les génotypes des fœtus sont soit des génotypes purs (LW, MS), soit des génotypes croisés (MSLW, LWMS). Il y a donc plusieurs types génétiques pour les fœtus d'une même portée. La comparaison des fœtus des deux génotypes extrêmes (LW et MS) est intéressante pour mieux comprendre et décrire les mécanismes moléculaires et cellulaires de la mise en place de la maturité entre lignées, mais aussi pour identifier des marqueurs biologiques de la maturité périnatale. L'influence des génomes parentaux LW et MS, selon qu'ils sont apportés par le père ou la mère, est possible grâce à la présence de fœtus dits croisés réciproques MSLW (une mère LW et un père MS) et LWMS (une mère MS et un père LW). Ceci permet de fournir des informations importantes concernant les effets maternels et paternels sur le développement du porcelet.

Le premier objectif de cette thèse est d'analyser les données musculaires (transcriptome, protéome et phénotype) afin de décrire et comprendre les bases moléculaires de la mise en place de la maturité chez les MS et LW. Le muscle squelettique joue un rôle déterminant à la naissance pour le stockage d'énergie (surtout sous forme de glycogène et un peu de lipides) et pour assurer la locomotion. En premier lieu, le transcriptome musculaire a été analysé seul, l'étude ayant

donné lieu à la publication d'un article, mentionné dans le chapitre 2. Ensuite, ces données d'expression génique ont été combinées avec des données protéomiques et phénotypiques afin d'obtenir une vision plus intégrée du système de maturation musculaire et de mieux détailler les mécanismes biologiques intervenant dans la mise en place de la maturité à la naissance. Ce travail est présenté dans le chapitre 3, avec un article qui sera prochainement soumis.

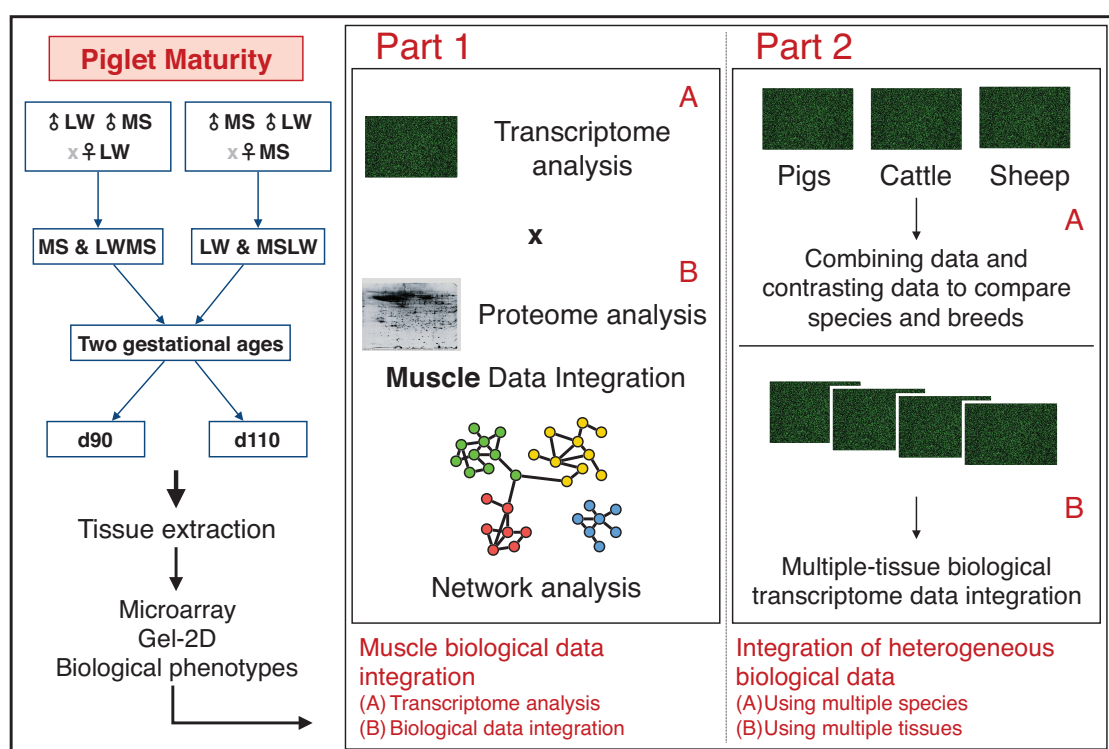


Figure 1.7 – Les différents niveaux d'intégration abordés dans cette thèse.

Le second objectif de cette thèse est l'intégration de données omiques provenant de différentes espèces sur un même tissu, et l'intégration de données omiques provenant de différents tissus d'une seule et même espèce. Lors d'une mobilité de 4 mois au CSIRO à Brisbane en Australie, des données de transcriptomes musculaires provenant de différentes espèces (cochon, mouton et bovin) ont été combinées afin d'observer des processus biologiques communs et différents entre ces trois espèces. Un article est en préparation et est présenté dans la première partie du chapitre 4. Par ailleurs, des méthodes statistiques multivariées ont été utilisées pour intégrer des données issues de différents tissus d'une même espèce. Des problèmes dus à la

présence de valeurs manquantes ont été rencontrés. Pour outrepasser ces problèmes, une méthode d'imputation de valeurs manquantes dans le cadre de l'analyse factorielle multiple (AFM) (ou *multiple factor analysis* (MFA)) a été développée et publiée, puis illustrée avec nos données transcriptomiques. Cette méthode est présentée dans la seconde partie du chapitre 4.

*

* *

Chapitre 2

Analyse du transcriptome musculaire

2.1 Introduction

Dans ce chapitre, une analyse transcriptomique (via une biopuce 60K) du tissu musculaire a été proposée afin d'identifier des processus biologiques et des gènes intervenant dans la mise en place de la maturité durant la fin de gestation chez le porc. Comme décrit au niveau de l'introduction (section 1.3), les fœtus provenant de deux races extrêmes pour la mortalité à la naissance, Large White (LW) et Meishan (MS), ont été utilisés. L'impact des génomes parentaux sur l'expression des gènes a pu être analysé grâce à la présence de fœtus croisés et purs au sein d'une même portée. Tous ces fœtus ont été prélevés à deux âges de gestation (90 et 110 jours de gestation) correspondant à la fin de gestation (naissance à environ 114 jours de gestation).

L'utilisation d'un modèle linéaire suivi d'un critère de sélection de modèle a été proposée, et 12 326 sondes ont été identifiées comme étant différentiellement exprimées selon le génotype fœtal et/ou l'âge gestationnel (Bonferroni à 1%). Ce nombre très important semble refléter un changement déterminant («*switch*») d'expression des gènes entre le début et la fin du processus de maturation. Parmi ces sondes différentielles, 2 000 sondes correspondant à 1 120 gènes annotés uniques (annotation de Mai 2014) étaient différentielles pour l'interaction entre l'âge gestationnel et le génotype fœtal. Afin de décrire le processus de maturation de façon globale,

des analyses d'enrichissement fonctionnel et d'inférence de graphe ont été effectuées et ont démontré que les gènes sur-exprimés à 90 jours de gestation étaient principalement impliqués dans le développement musculaire, alors que les gènes sur-exprimés à 110 jours de gestation étaient, quant à eux, impliqués dans des fonctions métaboliques comme la gluconéogénèse, le métabolisme du glucose, le métabolisme lipidique ou des protéines. Des gènes clés, impliqués dans ces métabolismes, comme *PCK2*, *LDHA* ou *PGK1*, ont été détectés comme ayant une expression différente à la naissance chez les MS par rapport aux LW. Grâce au dispositif de l'étude, nous avons également pu identifier 472 gènes ayant une expression préférentiellement régulée par l'un des deux génomes parentaux. Parmi ces gènes, 366 gènes étaient préférentiellement régulés par le génome paternel, notamment des gènes déjà décrits comme étant soumis à une empreinte paternelle, par exemple *MAGEL2* ou *IGF2*.

Tous ces résultats ont été publiés dans *BMC Genomics* en 2014 et sont présentés dans la première partie de ce chapitre. Ils montrent l'existence de mécanismes biologiques régulant la maturité musculaire chez les porcelets. Les porcelets ayant une immaturité au niveau du métabolisme musculaire seraient donc sujets à un plus fort risque de mortalité à la naissance. La maturité semble aussi être sous la régulation conflictuelle des génomes parentaux. Des gènes, pouvant expliquer les différences de maturité entre les LW et les MS car étant différenciellement exprimés entre ces deux génotypes, ont été identifiés dans des métabolismes déterminants pour la survie à la naissance. Ces gènes pourraient donc être de possibles marqueurs de la maturité musculaire.

2.2 Article 1 : Voillet et al., *BMC Genomics*, 2014

Cette section correspond à l'article suivant publié en septembre 2014 dans le journal *BMC Genomics* :

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

RESEARCH ARTICLE

Open Access

Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity

Valentin Voillet^{1,2,3}, Magali SanCristobal^{1,2,3,4,5}, Yannick Lippi⁶, Pascal GP Martin⁶, Nathalie Iannuccelli^{1,2,3}, Christine Lascor^{1,2,3}, Florence Vignoles^{1,2,3}, Yvon Billon⁷, Laurianne Canario^{1,2,3} and Laurence Liaubet^{1,2,3*}

Abstract

Background: In pigs, the perinatal period is the most critical time for survival. Piglet maturation, which occurs at the end of gestation, leads to a state of full development after birth. Therefore, maturity is an important determinant of early survival. Skeletal muscle plays a key role in adaptation to extra-uterine life, e.g. glycogen storage and thermoregulation. In this study, we performed microarray analysis to identify the genes and biological processes involved in piglet muscle maturity. Progeny from two breeds with extreme muscle maturity phenotypes were analyzed at two time points during gestation (gestational days 90 and 110). The Large White (LW) breed is a selected breed with an increased rate of mortality at birth, whereas the Meishan (MS) breed produces piglets with extremely low mortality at birth. The impact of the parental genome was analyzed with reciprocal crossed fetuses.

Results: Microarray analysis identified 12,326 differentially expressed probes for gestational age and genotype. Such a high number reflects an important transcriptomic change that occurs between 90 and 110 days of gestation. 2,000 probes, corresponding to 1,120 unique annotated genes, involved more particularly in the maturation process were further studied. Functional enrichment and graph inference studies underlined genes involved in muscular development around 90 days of gestation, and genes involved in metabolic functions, such as gluconeogenesis, around 110 days of gestation. Moreover, a difference in the expression of key genes, e.g. *PCK2*, *LDHA* or *PGK1*, was detected between MS and LW just before birth. Reciprocal crossing analysis resulted in the identification of 472 genes with an expression preferentially regulated by one parental genome. Most of these genes (366) were regulated by the paternal genome. Among these paternally regulated genes, some known imprinted genes, such as *MAGEL2* or *IGF2*, were identified and could have a key role in the maturation process.

Conclusion: These results reveal the biological mechanisms that regulate muscle maturity in piglets. Maturity is also under the conflicting regulation of the parental genomes. Crucial genes, which could explain the biological differences in maturity observed between LW and MS breeds, were identified. These genes could be excellent candidates for a key role in the maturity.

Keywords: Maturity, Survival, Birth, Muscle, Microarray, Systems biology, Pig

*Correspondence: laurence.liaubet@toulouse.inra.fr

¹INRA, UMR1388 Génétique, Physiologie et Systèmes d'Élevage, F-31326 Castanet-Tolosan, France

²Université de Toulouse INPT ENSAT, UMR1388 Génétique, Physiologie et Systèmes d'Élevage, F-31326 Castanet-Tolosan, France

Full list of author information is available at the end of the article

Background

Over the last decades, genetic progress has been associated with a rise in perinatal mortality in the domestic pig (*Sus scrofa*) [1]. In 2013, Strange et al. [2] noted that piglet mortality mostly occurs in the first 96 hours after birth. Because the pig is one of the most important meat-producing livestock species world-wide, this high piglet mortality at birth is a source of both economic [3] and ethical problems (public perception of the pig industry is affected by this young mortality [4]). Postnatal mortality is not an issue in pigs alone but also affects other mammals like sheep or humans [5,6]. In humans, for example, out of 4 million cases of infant death during the first four weeks of life, 28% are due to prematurity issues [6]. Adaptation to extra-uterine life is therefore a major factor for survival.

In pigs, various factors contributing to survival at birth have already been identified. They depend on maternal traits (e.g. farrowing duration, sow health), piglet characteristics (e.g. body weight at birth, genotypes) or environment [7,8]. As suggested by van der Lende [8], one of these factors is maturity. Maturation process was described to occur at the end of gestation, from 90 days of gestation to birth (114 days) [9]. Leenhouwers et al. [9,10] showed that a greater physiological maturity at birth is responsible for a higher survival. Thus, a successful maturation process leads to a state of full development and promotes early survival after birth [9,10].

Piglet maturity involves characteristics such as body size, body weight, organ characteristics and availability of body energy reserves such as glycogen or lipids [8,9]. Maturity is also coupled with the efficiency of physiological functions like thermoregulation, which is the balance between heat loss and heat production [8,11]. Although body weight has an influence on survival at birth, several studies suggest that it is not the only indicator of maturity [7-9,12]. The biological background imputable to the piglet's genetics has also been shown to impact survival [9,10], for example, Meishan piglets have a better survival rate than Large White piglets although they are lighter at birth. Moreover, Herpin et al. [11,13] highlighted that glucose homeostasis and body energy-glycogen storage are essential for survival.

Glycogen is the main source of polysaccharide stored in cells [11]. Glycogen storage is used to promote piglet thermoregulation at birth which takes place mainly in skeletal muscle (89% of the total glycogen) because of the absence of functional brown adipose tissue in piglets [8,11]. After birth, glycogen levels decrease by as much as 82% in muscle to provide the energy required [14]. Proficient thermoregulation, via glycogen storage in muscle, is thus an essential prerequisite for survival after birth. Thereby, the maturation of skeletal muscle metabolism

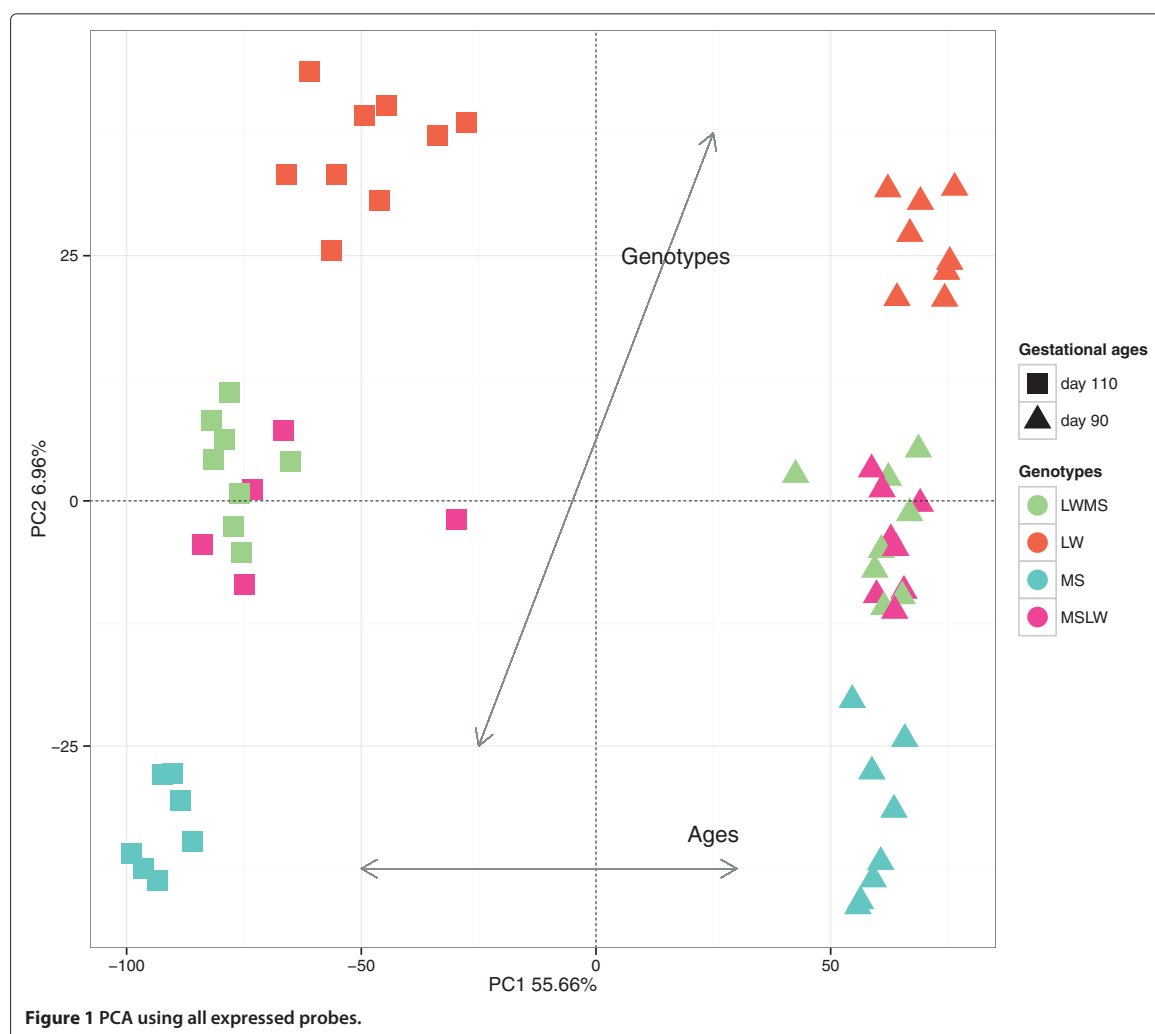
is indicative of metabolic maturity of piglet at the time of birth [11]. The muscle maturity could be defined by an immediate efficient motor function but also by an effective thermogenesis.

Some transcriptomic studies have already been performed to compare different stages of fetal muscle development in the pig. Cagnazzo et al. [15] compared seven prenatal stages (14, 21, 35, 49, 63, 77, and 91 days of gestation) in two breeds (Duroc and Pietrain) to highlight the differences in muscle development in these two breeds, while Xu et al. [16] compared a prenatal stage (65 days of gestation) with postnatal stages (3, 60 and 120 days after birth) to bring to light the mechanisms underlying muscle growth in Meishan pigs. To our knowledge, no transcriptomic studies have yet been carried out on the last phases of fetal development in connection with maturity. Here, we performed microarray analysis to describe the biological processes underlying muscle piglet maturity and identify candidate genes. The objective was to identify the genes and biological processes that are specifically involved in the differences in muscle development observed between two extreme breeds: Large White and Meishan. The Large White (LW) breed is a highly selected breed with a high rate of mortality at birth, whereas the Chinese Meishan (MS) breed produces piglets with extremely low mortality [12,17]. The high selection in LW has led to a lower maturity of these piglets at birth [12]. MS and LW sows were inseminated with mixed semen (LW and MS). Hence each litter was composed of purebred fetuses (LW or MS) and crossbred fetuses (LWMS from MS sows and MSLW from LW sows). In the present study, we highlight key genes and biological functions involved in piglet maturity. This analysis will help to improve our knowledge of maturity in the pig.

Results

Power of experimental design

Microarray analysis was performed to study the last step of fetal development in the pig. Muscle samples (Longissimus dorsi) were collected from 61 fetuses in 8 different conditions (four genotypes (LW, MS, LWMS, MSLW) at two gestational ages (90 and 110 days)). After normalization, the signal intensity was found to be above background noise for 44,368 spots. Principal component analysis (PCA) was carried out to evaluate microarray quality and observe fetus dispersion (Figure 1). Principal component (PC) 1 (56% of the total dispersion) segregated fetuses according to gestational age (day 90 and day 110) while PC 2 (7%) separated fetuses according to genotype. The two groups of crossbred fetuses were mixed even with PC 3 (4%). The results of this initial descriptive study, performed without preselecting spots,



showing a clear separation between gestational ages and purebreds, demonstrate that the experimental design was very powerful.

Identification of differentially expressed genes

A mixed linear model was applied to each spot. This model involved two factors, gestational age and fetal genotype (fixed effects) as well as their interaction, and the sow as a random effect. The number of differentially expressed probes (DEP) was high even with a stringent correction for multiple tests (Bonferroni or False Discovery Rate (FDR)). Indeed, a total of 12,326 DEPs (corresponding to 5,634 unique annotated genes) were identified with a significance threshold of 1% with Bonferroni correction. This large number could be

explained by a large effect of fetal gestational age and the high power of the experimental design.

The list of 12,326 DEPs was then partitioned into 4 sub-models using the Bayesian Information Criterion (BIC). Sub-model 2 (additive model for age and genotype effects) including 41% of the DEPs accounted for the largest proportion of DEPs, followed by sub-model 3 (age effect only) with 40% of the DEPs. Sub-model 4, which included the genotype effect only, contained only 3% of the DEPs. Sub-model 1 contained 2,000 DEPs (16%) (Additional file 1) and was particularly interesting because it combined the two factors of interest (gestational age and fetal genotype) and their interaction. This suggests that the last phase of the developmental process between 90 and 110 days of gestation is different across genotypes. It was therefore deemed that further analysis of this probe list

would help to identify the biological processes involved in maturity.

Ontological and functional biological analysis of differentially expressed genes

Biological processes enriched during the maturation process

Gene Ontology (GO) is a standard system of classification of gene product attributes in terms of their associated biological processes, cellular components and molecular functions. Sub-model 1 identified 2,000 DEPs corresponding to 1,120 unique annotated genes. GO functional enrichment analysis was performed on two lists of genes from sub-model 1 using an absolute log₂-fold change $> \frac{1}{2}$ (corresponding to an absolute fold change 1.4) between fetal gestational ages averaged over all genotypes. The first list contained 394 unique up-regulated genes at gestational day 110, and the second list contained 441 unique up-regulated genes at day 90 (Additional file 1). The top significant GO annotations indicated that the enriched biological processes at 90 days of gestation were related to muscle development (Figure 2A). Cell adhesion or signal transduction as biological processes, extracellular matrix as cellular component and extracellular matrix structural constituent as molecular function were enriched at 90 days of gestation (Figure 2A). At 110 days of gestation, the top significant enriched biological processes and molecular functions were generally involved in energy metabolism, e.g. gluconeogenesis, glucose metabolic process, cellular lipid metabolic process or oxidoreductase activity (Figure 2B). Enriched cellular components were linked, inter alia, to mitochondrion (Figure 2B). All enriched GO terms (125 GO terms at 90 days of gestation and 75 GO terms at 110 days of gestation) are presented in Additional files 2 and 3 (with unadjusted and adjusted p-values, descriptions and genes).

Differences of gene expression between gestational ages for each extreme fetal genotype (LW and MS)

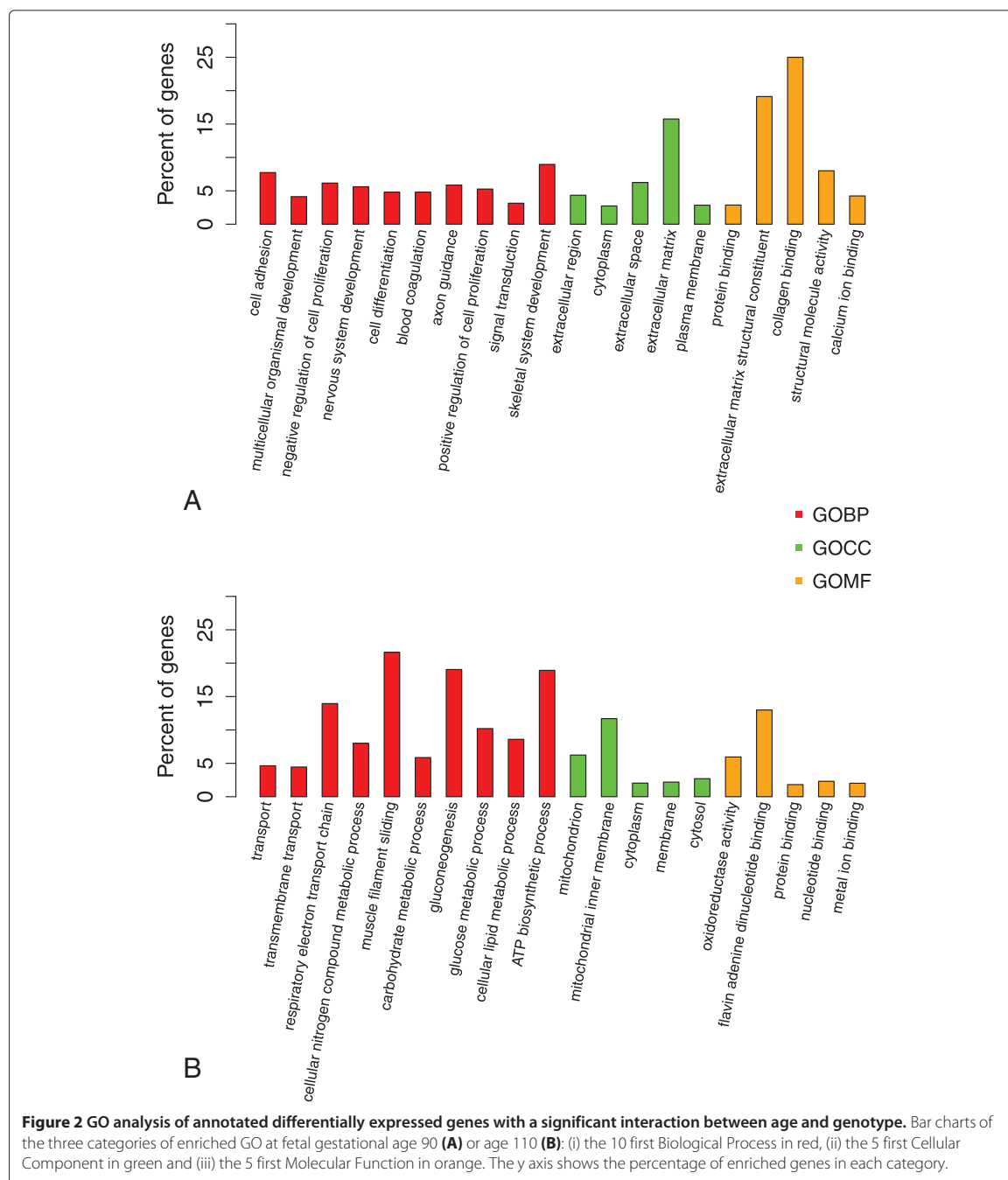
The aim of this part of the study was to highlight the differences of maturation process observed between purebreds (LW and MS) and more particularly to identify genes that may explain the impaired maturity of LW piglets at birth. As above, GO functional enrichment analysis was performed on the lists of genes from sub-model 1 using an absolute log₂-fold change $> \frac{1}{2}$ between fetal gestational ages separately for fetuses of each breed (LW and MS). These genes were differentially expressed between the both gestational ages but not necessarily for every genotype. Interestingly, more up-regulated genes were identified in MS than in LW for several enriched biological processes, e.g. various metabolic processes at 110 days of gestation or muscle development at 90 days of gestation (Table 1, Additional file 4). Another example was the

glycolysis and gluconeogenesis KEGG pathway (Figure 3). Among seven enriched genes in this KEGG pathway, *PGK1* (Phosphoglycerate Kinase 1), *PCK2* (Phosphoenolpyruvate Carboxykinase 2, mitochondrial also known as *PEPCK*) or *LDHA* (Lactate dehydrogenase A) were only up-regulated in MS at day 110 (see box-plots on Figure 3). These genes may illustrate the differences in the muscle maturation process between purebred MS and LW piglets.

Differences between extreme fetal genotypes (LW and MS) with a relevance network approach

In contrast with bibliometric networks (e.g. with Ingenuity), a relevance network approach can use the transcriptomic information of both annotated and unannotated genes. Influential genes can be found by looking at degrees and betweenness centrality. The degree is the number of edges a gene, as a node, has to other genes. A gene with the highest degree is usually considered as a hub. Betweenness centrality quantifies the number of times a gene acts as a bridge along the shortest path between two other genes. A relevance network between the 1,516 unique genes (annotated or not) from sub-model 1 was built on Pearson correlation r using a threshold of $|r| > 0.98$. It should be noted that with a lower threshold the resulting graph was too large and too highly connected to be interpreted (Additional file 5). The largest connected component obtained was composed of 96 nodes and 381 edges (Figure 4, Additional file 6). *NUSAPI* (Nucleolar and spindle associated protein 1) was the gene with the highest degree, while *CDK6* (Cyclin-dependent kinase 6) was the gene with the highest betweenness centrality (Additional file 6). By maximizing the modularity criterion, four communities were identified in this graph (Figure 4). According to functional enrichment analysis, each community was related to a particular biological function: (i) community 1 was involved in cell division and nucleus, (ii) community 2 in cell adhesion and extracellular matrix, (iii) community 3 in collagen, while (iv) community 4 was involved in regulation of the fatty acid metabolism and oxidation-reduction process. Detailed results (unadjusted and adjusted p-values, descriptions and genes) of enriched GO analysis are presented in Additional file 7. Genes belonging to communities 1, 2 and 3, e.g. *NUSAPI*, *STMN1* (Stathmin 1) and *COL5A2* (Collagen alpha-2(V) chain), were mainly up-regulated at day 90, with a higher expression in LW than in MS at day 110 (Figure 5). On the contrary, the genes of community 4, e.g. *DCI* (or *ECI1*) (Enoyl-CoA Delta Isomerase 1), were mainly up-regulated at day 110 with a higher expression in MS (Figure 5). The expression profiles of the relevance network highlighted a delay of gene expression in LW fetuses at 110 days of gestation.

From a methodological point of view, it should be noted that a bias could have been introduced by the choice



of the Pearson correlation to represent the relationships between genes [18]. Partial correlation, which discriminates between direct and indirect relationships, may have led to more relevant measurements of the direct dependence between variables [18,19]. However, in our study, the genes of interest were too numerous (1,516 unique

genes) and too highly correlated to compute partial correlations correctly.

Taken together, the up- or down-regulation of the genes involved in these four communities was delayed at 110 days of gestation in LW fetuses compared with MS fetuses. These results are consistent with a lower maturity (and a

Table 1 Table of six enriched GOBP at day 90 or 110 in the two extreme breeds

Day	Items	GOBP Terms	Genes
90	GO:0007275	Multicellular organismal development	<i>TCF12 MGP KLF3 FRZB DIAPH2 IGF2 SEMA4D GPSM1 MESP1 CCBE1 VEGFC CREM RYBP JAG1 KDR CSPG4</i> CECR1
	GO:0030154	Cell differentiation	<i>TCF12 MGP FRZB DIAPH2 SEMA4D GPSM1 VEGFC CREM CSPG4</i> SH2B3
	GO:0001501	Skeletal system development	<i>FRZB IGF2 GDF11 IGF1</i>
110	GO:0006094	Gluconeogenesis	<i>PCK2 GPD1 PGK1</i>
	GO:0006006	Glucose metabolic process	<i>PCK2 UPG2 PGK1 PYGL SORD</i>
	GO:0044255	Cellular lipid metabolic process	<i>GPD1 OXCT1 SLC25A20</i>

In italic, genes are up-regulated in MS only, and in bold, genes are up-regulated in LW only. Genes up-regulated in MS and LW are not represented. The complete list of genes up-regulated in MS and/or LW is given in Additional file 4.

higher mortality) of LW piglets at birth compared to MS piglets.

Influence of the paternal or maternal genome on gene expression

This experiment used a reciprocal design to independently evaluate the effect of each parental genome on the maturation process (MS and LW). In other words, if both genomes contributed to the same extent, as expected in the Mendelian context, gene expression would be identical in the two sets of crossbred fetuses. However, the reality is that some genes are not regulated in the same manner depending on the origin of the allele. Eight hundred and five probes were identified to be impacted by the parental genotype in interaction with gestational age (FDR < 1%). One hundred and six unique annotated genes (164 probes identified) were influenced by the maternal genotype and 366 unique annotated genes (641 probes identified) were influenced by the paternal genotype (Additional files 8 and 9). It should be noted that 19 probes (12% of the 164) influenced by the maternal genome were located on chromosome X versus 19 probes (3% of the 641) influenced by the paternal genome. Because only male fetuses were studied, all genes from the X chromosome were of maternal origin. Moreover, 4 probes out of the 164 probes influenced by the maternal genome were located on mitochondrial chromosome.

Several previously identified genes (of sub-model 1) were also influenced by a parental effect (602 probes). For example, *PCK2* and *LDHA* were impacted by the paternal genome as can be seen in the box-plots (Figure 3). Other identified genes, such as *SORD* (sorbitol dehydrogenase) and *CREM* (cAMP responsive element modulator), showed both a parental effect and a difference between purebred fetuses (Table 1, Figure 6). *CREM* expression was influenced by the maternal genotype, whereas *SORD* expression was influenced by the paternal genotype (Figure 6). These genes illustrate the possible impact of the parental genotype on gene expression, and its effect on maturity.

Validation of differential expression by quantitative real time PCR

To validate the microarray results, the expression profiles of 10 genes of interest were monitored using qRT-PCR. The 10 selected genes showed differential expression for the two fixed effects and their interaction in the microarray (sub-model 1). The similarity between the results obtained with the microarray and qRT-PCR confirmed the accuracy of gene expression measurements (illustrated in Additional file 10). Indeed, the Pearson correlation between the differences in expression measured by qRT-PCR and microarray was greater than 0.70 for all genes, except *ILIRAPL2* and *SPG7* (Table 2). The high variability obtained by qPCR suggests that all genes were not highly correlated, especially in the LW sample at 110 days of gestation. Nevertheless, the correlation obtained confirms our previous results and all the expression profiles are similar.

Discussion

Important transcriptomic changes between 90 and 110 days of gestation

The statistics chosen here to detect DEPs made no use of a minimum fold change filter. This choice was based on the complexity of the experimental design (8 conditions in a 2 by 4 factorial design) making a fold change calculation difficult. Moreover, no associated values indicate the level of confidence in the designation of genes as differentially expressed or not differentially expressed with such a fold change filter. A fold change filter was applied only at the point of gene function enrichment. The counterpart of this choice may be a potential for over-interpretation of data, by extracting genes with very small changes in expression. With a commonly-used threshold in FDR of 1%, as high as 28,833 probes were found DE, corresponding to 65% of expressed probes of the microarray. We chose instead a conservative cutoff of 1% with a Bonferroni correction for multiple testing. A high number of DEPs (12,326 probes corresponding to 28% of the expressed probes) was obtained even with

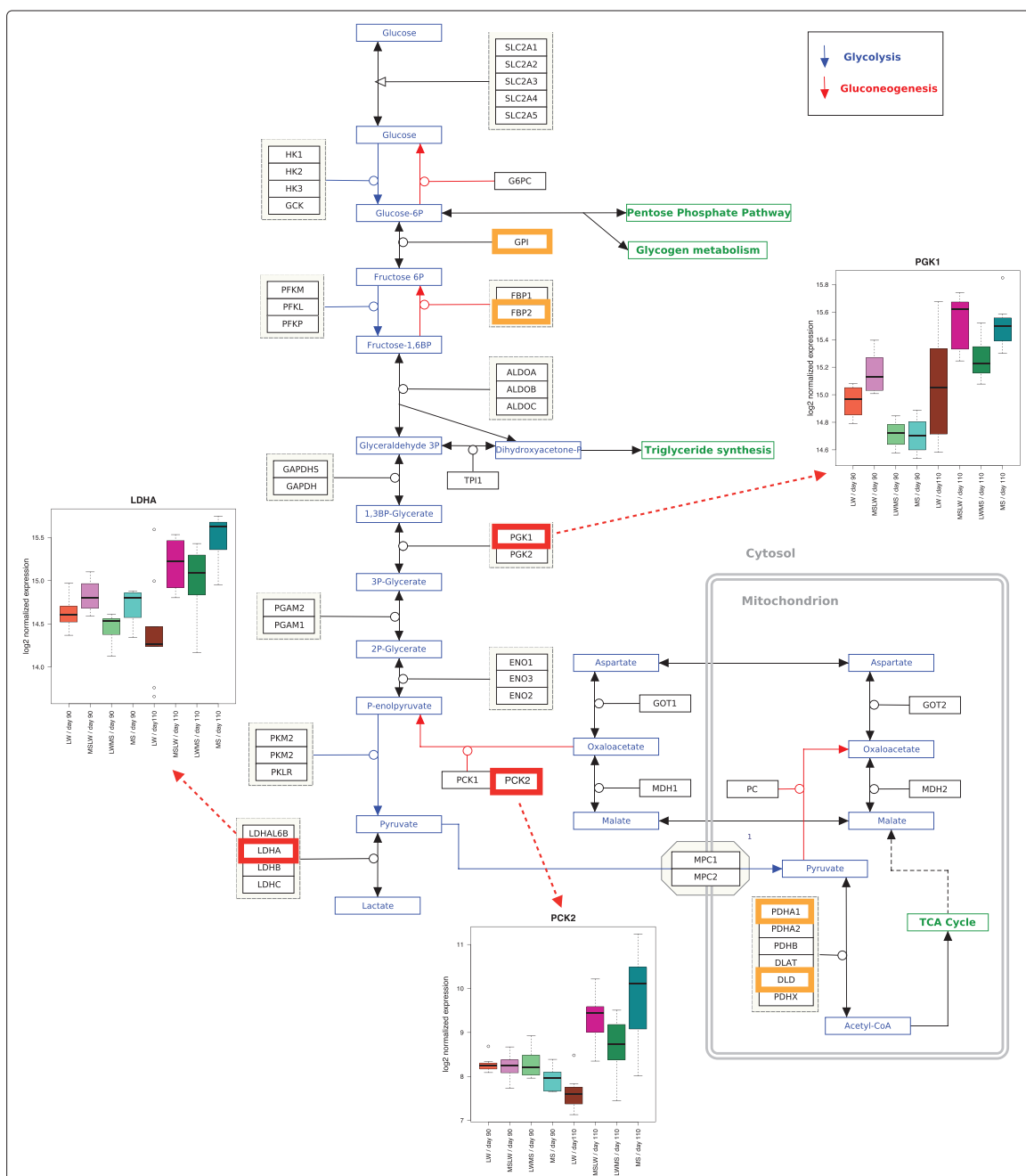


Figure 3 Example of difference between LW and MS at fetal gestational age 110: pathway of Gluconeogenesis. This figure represents the metabolic pathway of glycolysis and gluconeogenesis. Genes circled in orange (*GPI, G6PC, PDHA1, DLD*) are up-regulated at day 110 in KEGG pathway gluconeogenesis in both extreme breeds. Genes circled in red (*PKG1, PCK2, LDHA*) are up-regulated at day 110 in KEGG pathway gluconeogenesis in MS only. Box-plots of *PCK2, PKG1* and *LDHA* are added and allow to observe up-regulated genes at day 110 in MS only. The gene expression were log2 transformed. The red-circled genes illustrate the difference of maturity between MS and LW.

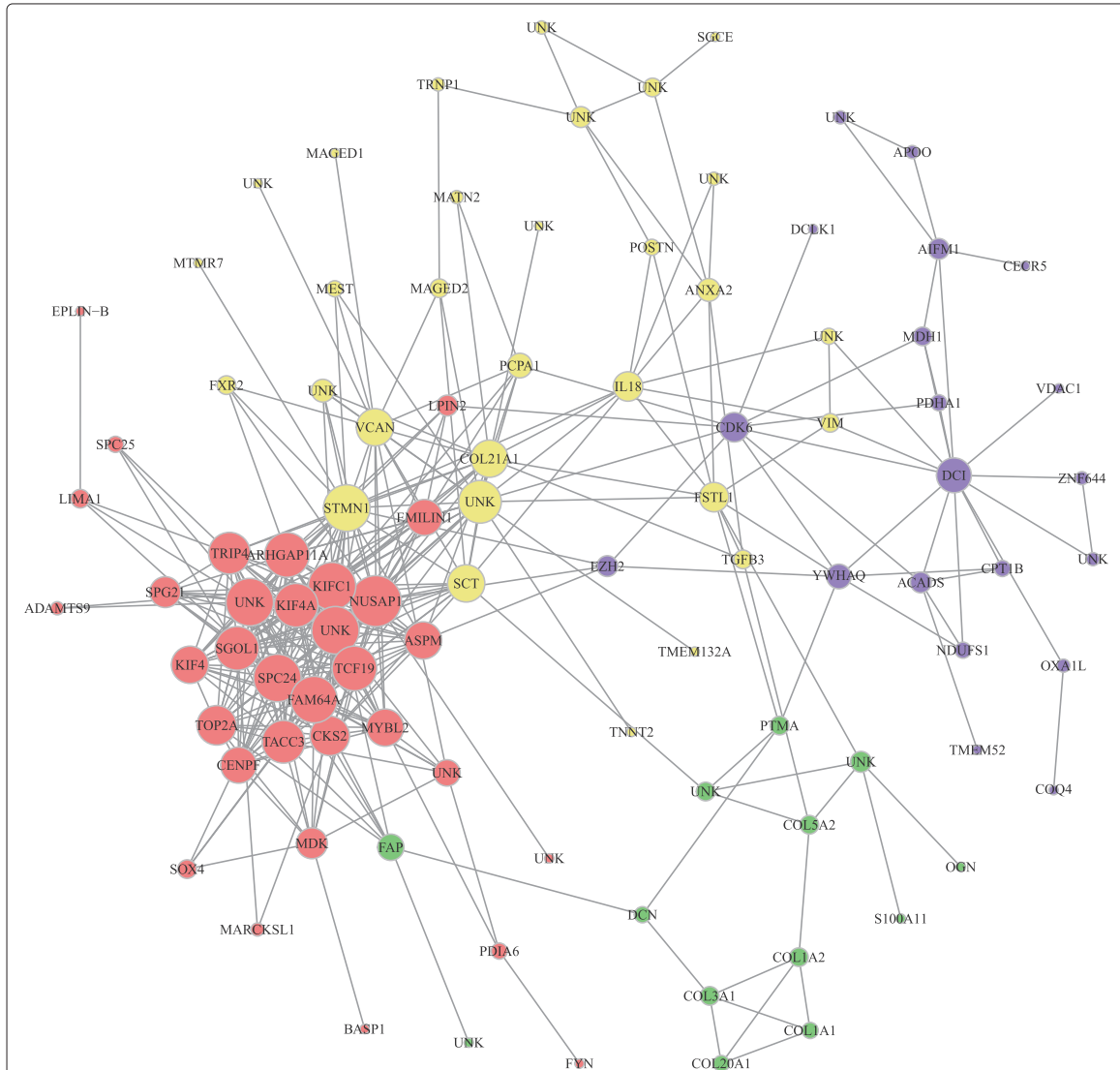
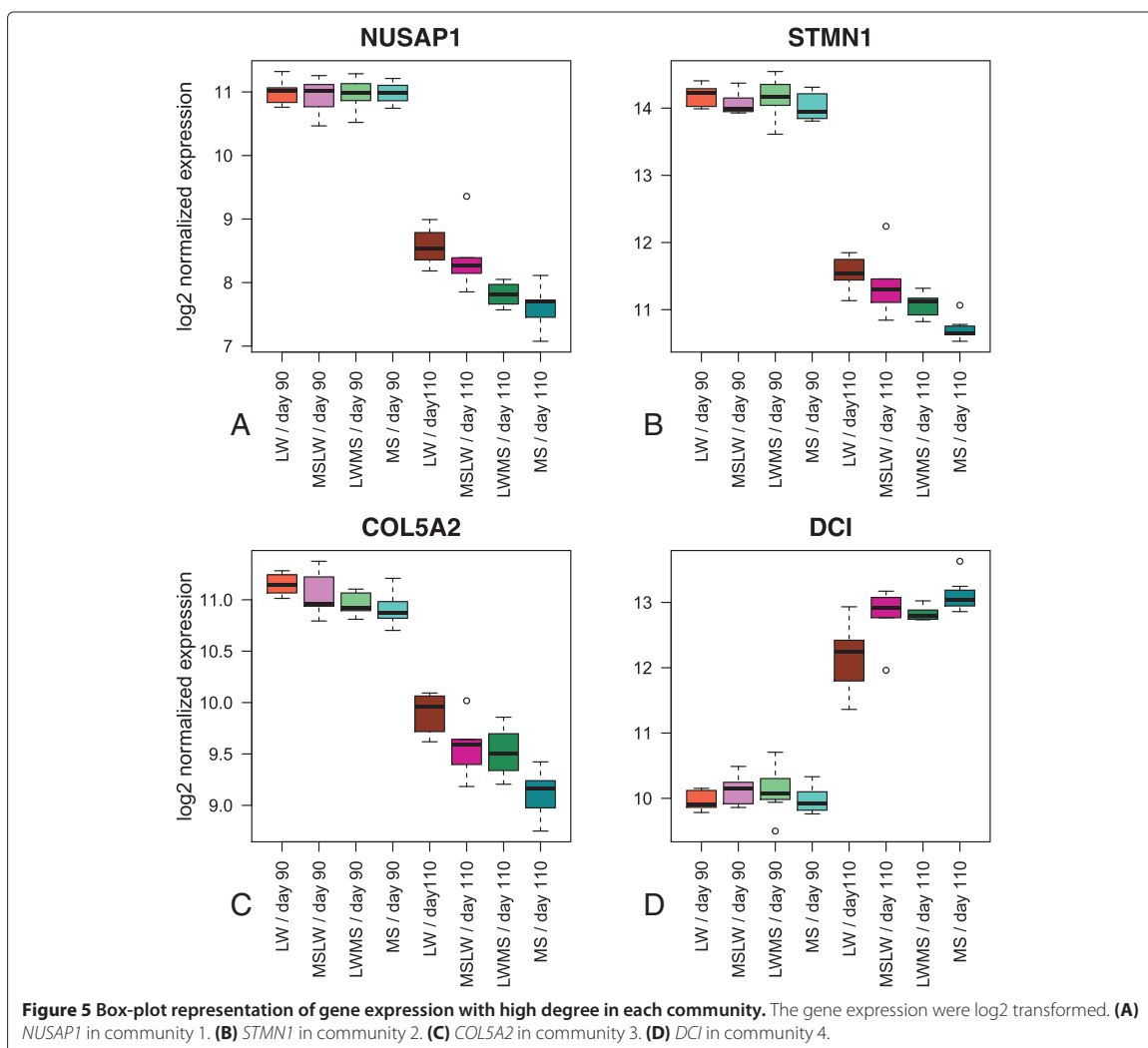


Figure 4 Largest connected component of the relevance network. Each node in the graph represents a gene and each edge corresponds to a Pearson correlation between two genes above the defined threshold ($|r| > 0.98$). The size of each node is proportional to its degree. The graph included 96 nodes and 381 edges. Modularity was maximized to find communities in the graph. Four communities were found and labelled with color (red for community 1, yellow for community 2, green for community 3 and blue for community 4). The percentages represent the number of genes in each community. For each community, we studied biological processes with GO functional enrichment analysis. Communities 1, 2 and 3 were mainly involved in muscle development, such as cell division, cell adhesion and collagen. Community 4 was mainly involved in metabolism like the fatty acid one. All enriched GO Terms are detailed in the Results section and Additional file 7. UNK stands for 'unknown'.

this stringent correction. This high number demonstrates, as does PCA (Figure 1), that the choice of breeds and the two gestational ages were highly relevant in order to study contrasted situations linked to the maturation process. The important impact of genotype was expected in agreement with Hazard et al. [20], where 82% of the differentially expressed genes were impacted by the

genotype, comparing LW and MS in another experimental context.

Among these DEPs, 11,952 probes (97%) were influenced by the gestational age of fetus (genes in sub-models 1, 2 and 3). PCA of all expressed genes also showed the prime importance of fetal gestational age (see the first axis on Figure 1). This high number of DEPs related to



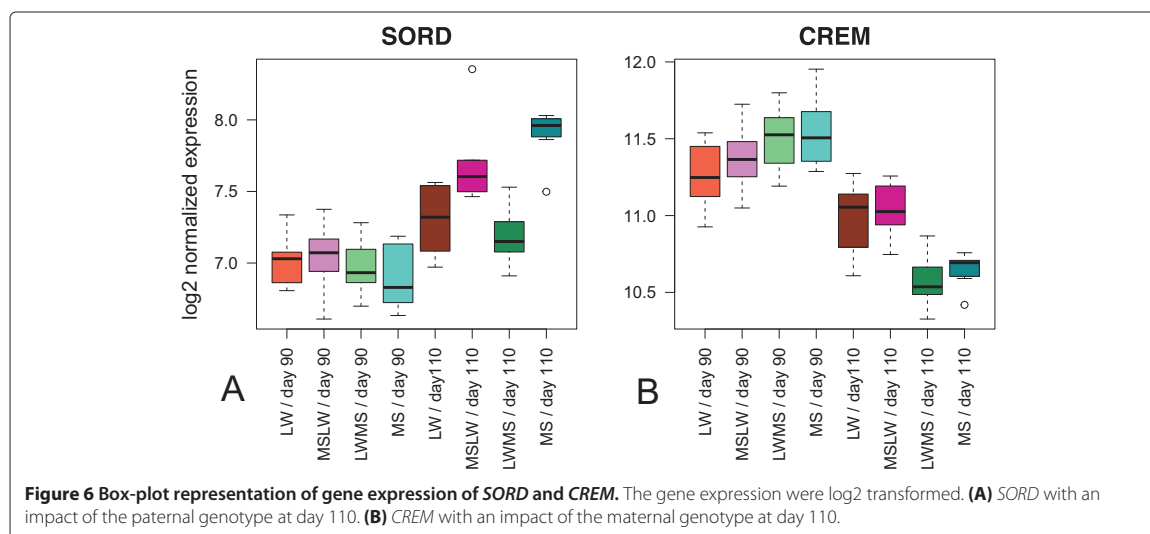
gestational age, about 27% of the probes in the microarray, could be explained by a switch of gene expression between 90 days and 110 days of gestation. Muscle development is described to end around 90 days of gestation [21], and gives way to the maturation process in order for the organs and tissues to be functional at birth. In other words, for the fetus to be able to adapt to the extra-uterine environment, fetal tissues must have acquired complete functionality at birth.

Using embryo transfers between two breeds (MS and Yorkshire), of the results published by Wilson et al. [22] and Biensen et al. [23,24] suggested that fetal development is determined by the uterine environment until 90 days of gestation, regardless of the fetal genotype. After 90 days of gestation, the last phase of fetal development is preferentially modulated by the fetal genotype with mechanisms

specific to each genotype [23]. Our experiment explored this final step of development in utero. We revealed the importance of the transition between fetal development and metabolism in muscle tissue for survival (i.e. energy storage and function: gluconeogenesis, glycolysis and fatty acid metabolisms).

Main biological mechanisms of maturity in pigs

To identify the biological processes underlying muscular maturity, functional enrichment analysis was performed on two gene lists from sub-model 1 (model which combined two factors, gestational age and fetal genotype (fixed effects), as well as their interaction, and the sow as random effect). The first list consisted of 441 up-regulated genes at day 90, and the second consisted of 394 up-regulated genes at day 110. These genes were chosen because of



their interaction between the gestational age and the fetal genotype. We wanted to observe differences in the muscle maturation process between the both extreme breeds.

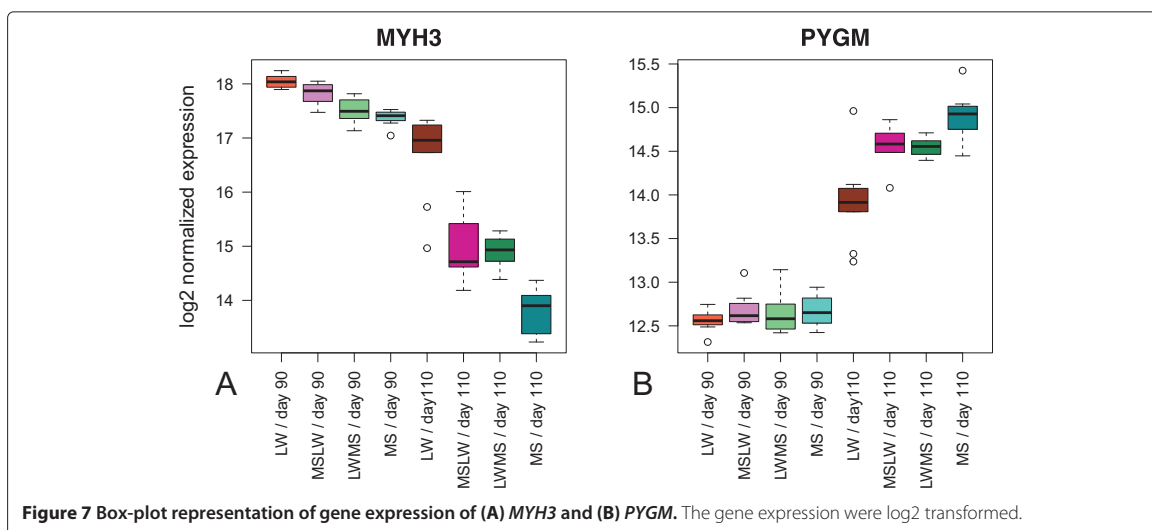
Enriched biological functions at 90 days of gestation were involved in muscle development and reflected processes such as cell adhesion, signal transduction or skeletal muscle development (Figure 2A). These results are consistent with the second phase of muscular development known to occur in pigs between 55 and 90 days of gestation [21,25]. In the pig, ontogenesis of muscle fibers is a biphasic phenomenon [21,25,26]. A first generation of myofibers develops between 35 and 55 days of gestation, followed by a second generation between 55 and 90 days of gestation. The second generation develops around each primary myotube using it as a scaffold and determines the number of myofibers [21,25]. The total number

of myofibers is definitively fixed at approximately 90 days of gestation [27]. Moreover, it has been reported that the number of myofibers is lower in MS than LW which may explain the small postnatal muscle growth capacity of MS pigs, in particular of large glycolytic muscle, e.g. longissimus muscle [27].

Among genes of sub-model 1 (Additional file 1, Figure 7A), the embryonic heavy chain isoform of myosin (corresponding to *MYH3* gene) is described to be only expressed between 50 days of gestation and birth [28]. In this study, while *MYH3* expression decreased in MS just before birth, its expression in LW stayed at a high level even at 110 days of gestation. It may correspond to the delay of variation of expression mainly observed between LW and MS at 110 days of gestation, suggesting a possible state of immaturity at birth as expected in LW. In another way, a variation of copy number of *MYH3* has been recently identified in cattle, that was positively correlated with its transcript expression and body traits (body height, body weight and body size) [29]. It would be possible to imagine that selection on growth traits impacted polymorphisms in this gene, such as CNV (Copy number variations). In our study, *MYH3* expression is eight fold higher in LW than in MS just before birth. Further studies would be needed to explore the relationship between this possible delay of maturity in LW with the differential proportion of glycolytic myofiber observed between LW and MS (larger proportion in LW [27]). It was observed that comparison between LW and MS may suggest that intensive selection for lean muscle growth induced a shift in muscle metabolism toward a more glycolytic and less oxidative myofiber type [27]. However, it is not clear if a functional link may be proposed to explain an effect of the genetic difference of a higher expression of *MYH3* at birth

Table 2 Correlation between qPCR expression with microarray expression for selected genes (n = 43)

Genes	Pearson's correlation	P-value
<i>ARG2</i>	0.95	< 0.001
<i>PHKA1</i>	0.86	< 0.001
<i>SLC38A4</i>	0.84	< 0.001
<i>DLK1</i>	0.83	< 0.001
<i>RASGRP3</i>	0.81	< 0.001
<i>GPD1</i>	0.76	< 0.001
<i>DUT</i>	0.74	< 0.001
<i>GBP1</i>	0.71	< 0.001
<i>IL1RAPL2</i>	0.65	< 0.001
<i>SPG7</i>	0.40	0.008



with the metabolic status in longissimus dorsi muscle. In another way, the expression of *PYGM* gene (glycogen phosphorylase) was up-regulated at 110 days of gestation with a higher expression in MS (Figure 7B) meaning a delayed expression in LW. A higher activity of the glycogen phosphorylase at protein level would illustrate a higher capacity to degrade the glycogen store at birth to produce energy.

The pattern of muscle development was confirmed by the results of our relevance network in which the genes involved in development were up-regulated at 90 days of gestation. For example, *NUSAPI*, the gene with the highest number of connections to other nodes (or degree) in our network, is a key gene for spindle microtubule organization and has been identified as playing a central role in regulating mitosis depending on its phosphorylation state [30,31]. The gene with the highest degree of relevance in community 2, *STMNI*, is also involved in the regulation of the microtubule filament system [32]. Moreover, part of the collagen family was present and connected in community 3 (Figure 4). *CDK6* is an important gene for two reasons: it has the highest betweenness centrality, indicative of its prime role in the structure of the network, and it links community 4 (muscle metabolism, with genes up-regulated at 110 days) with the remaining three communities (involved in muscle development, genes up-regulated at 90 days). Interestingly, *CDK6* is up-regulated at 90 days, unlike genes of community 4. This member of the cyclin-dependent protein kinase family regulates cell cycle progression and is involved in the regulation of skeletal muscle regeneration [33,34]. Because these genes are down-regulated between 90 and 110 days of gestation, our results are in accordance with the previous hypothesis that muscle development “switches off” at around 90 days.

At 110 days of gestation, the enriched biological functions detected were involved in energy metabolism, especially in gluconeogenesis and cellular lipid processes (Figure 2B). In contrast with intra-uterine life where fetal body temperature depends on the sow, autonomous thermoregulation must occur immediately upon birth for the piglet to survive [11]. Energy reserves, i.e. glycogen and fat, must therefore be maximal in the neonatal period because the piglet cannot oxidize protein efficiently before 5-7 days of life [11]. In pig fetuses, glycogen is stored in skeletal muscle (89% of all glycogen reserves [11]) and liver, because pigs lack brown adipose tissue [8,35]. The initial role of muscle glycogen, in addition to motor function, is postnatal thermogenesis, especially prior to colostrum intake. Later, if energy intake is insufficient, the piglet draws down on its muscle glycogen reserves [11]. Large amounts of glycogen are therefore stored in muscle before birth (around 114 days of gestation) [11]. The genes of community 4 that were up-regulated at 110 days of gestation were involved in fatty acid metabolism which is also important for forming the body energy reserves required at birth [11]. *DCI*, which was the gene with the highest degree of relevance in this community and the second betweenness centrality in our network, encodes a key mitochondrial enzyme involved in beta-oxidation of unsaturated fatty acids [36].

Globally, the process of muscle maturation was the same in each studied genotype: muscle fiber proliferation is switched off at around day 90, and the enzymes coding genes for glycogen and lipid metabolism are up-regulated at around day 110 to ensure regulation of mechanisms essential for survival at birth. This is also in line with the changes in gene expression that have been reported at the end of gestation.

Contrasted maturation process between extreme breeds

Even if the overall muscle maturation process is the same for each genotype, some important differences were found between the extreme breeds that affected crucial biological processes (Table 1 and Additional file 4). Genes such as *PGK1*, *PCK2* or *LDHA*, encoding key enzymes involved in gluconeogenesis and glycolysis KEGG pathway, were up-regulated at day 110 in MS only (Figure 3). It should be noted that *PCK2* and *LDHA* were down-regulated at day 110 compared to day 90 in LW (see box-plots on Figure 3). Of special interest is *PCK2* that encodes an enzyme that catalyzes the irreversible conversion of oxaloacetate (OAA) to phosphoenolpyruvate (PEP), the rate-limiting step in the metabolic pathway that produces glucose from lactate and other precursors derived from the citric acid cycle [37]. This result may be surprising as gluconeogenesis is more often described to be a liver function than a property of muscle [38].

Body energy reserves, i.e. glycogen and fat, are important predisposing factors involved in the maturation process [8,9]. The fact that a piglet is unable to produce heat, may be the result of an immature metabolic capacity. A weaker expression of the genes (e.g. *PCK2*, *PGK1* or *LDHA*) involved in these metabolic pathways between the two gestational ages may cause the lower maturity observed in LW, and therefore be responsible for the larger proportion of deaths at birth in this breed. Moreover, muscle glycogen content has been already studied just before birth [9,10]. The animal's requirement for energy is maximum in the neonatal period to promote thermoregulation and growth [11]. It strongly suggested that piglets with high value for survival, like MS, have a higher ability to maintain glucose levels during and after farrowing and are better able to maintain body temperature. For a long time, the storage and mobilization of the glycogen in muscle was known to be essential for survival at birth [39].

Genetic selection has been shown to alter genes which may be associated with marked differences in maturity between MS and LW [40]. Canario et al. [12] have already shown that selection for leanness in LW resulted in a lower maturity of piglets at birth, for example, with an effect on the body protein content and liver glycogen stores. Genetic polymorphisms could affect the genes involved in these processes. For example, *PGK1* showed a greater variability of expression in LW suggesting a possible underlying polymorphism in purebred LW fetuses (see box-plot on Figure 3). *PCK2* and *LDHA* showed distinct expression profiles in MS and LW (down-regulated at day 110 compared to day 90 in LW) (see box-plot on Figure 3).

In addition to the genes up-regulated at 110 days of gestation in MS pigs only, the lesser maturity of LW piglets could also be explained by a delay in gene expression at the end of the intra-uterine developmental period. As found with the relevance network approach (Figure 4),

most genes in the communities involved in development were down-regulated just before birth, such as *NUSAPI*, *STMNI* or *COL5A2*. These genes had a higher expression in LW than in MS (Figure 5). Most genes of community 4 (metabolic processes), such as *DCI*, were up-regulated at 110 days of gestation and the related genes were mainly expressed at a higher level in MS than in LW. Our relevance network approach is therefore in accordance with the assumption that MS piglets are more mature than LW piglets. Furthermore, the network may also provide information on unannotated genes by guilt-by-association. These results suggest that the genetic selection may affect genes involved in the muscle metabolic capacity confirming the initial assumption that MS newborns are more mature than the LW ones [17] and enables us to suggest candidate genes for piglet maturity.

Impact of the parental genotypes

We identified a great number of annotated genes (472) that were impacted by one of the two parental genotypes during maturation process: 106 genes were influenced by the maternal genotype and 366 genes by the paternal genotype (Figure 6 and Additional files 8 and 9). The influence of the parental genotype was found for several genes, such as *CREM* and *SORD* (Figure 6), which also showed difference between extreme breeds (Table 1). Some of these genes are known to be related to imprinting in pigs. Many studies of genome scanning for QTL (Quantitative Trait Loci) in pigs revealed that many of them are maternally or paternally imprinted, which significantly affect growth, backfat thickness, carcass composition and reproduction [41]. Among these genes, we identified *IGF2* (Insulin-like growth factor 2) [42] and *MAGEL2* (MAGE-like 2) [43]. The selection pressure for enhancing lean meat content was described to be related to increase *IGF2* transcript expression in muscle [44]. The imprinting of the *MAGEL2* gene is highly conserved among species [43]. These results are consistent with the theory explaining parental conflict during gestation [45]. The parental genomes have a direct impact on fetal gene expression (2.8% of expressed genes in our microarray). The genetic component of piglet survival consists of a maternal genetic component (genotype of the mother) and a direct genetic component (genotype of the piglet) [9]. In this context, at the end of gestation, the paternal expression genes are up-regulated to allow the fetus to express its growth potential [23]. It has already been shown that paternally expressed genes are not essential for the initiation of fetal development [46], but their role becomes more critical at the end of gestation.

Moreover, these genes could play a key role in the maturation process. A large number of imprinted genes in humans are known to affect metabolic parameters such as

glycogen metabolism [47]. For example, the expression of *MAGEL2*, known to have an effect on metabolic parameters and fetal growth [47], is impacted by the paternal genotype in our study (Additional file 9). This gene was up-regulated at 90 days of gestation and further up-regulated in paternal genotype MS. The MS paternal genotype therefore represents a strong genetic component for this gene's expression. Thus, the delay in the maturity of LW piglets may also be explained by differential impacts of the parental imprinting of some genes involved in metabolic processes. However, in our study, the relative involvement of each parental genome was studied only by comparison between the reciprocal crossed fetuses then no affirmation could be done on the imprinting status of the identified genes. Allele specific expression could be investigated in the future.

Conclusions

Our experimental design was very powerful in order to unravel the biological processes underlying the last phase of muscle development, and identify key muscle differences between LW and MS pigs that could explain lesser maturity of LW piglets at birth. Biological functions and genes involved in maturity have been identified in each breed. This study shows that a considerable transcriptomic change occurs between 90 and 110 days of gestation, and corresponds to a switch between muscle development and muscle metabolism. This study also highlighted genes with differences of expression between extreme breeds LW and MS, such as *PCK2*, *PGK1* and *LDHA*. These genes play roles in the implementation of the muscle metabolic processes necessary for thermoregulation at birth. These processes are also under the conflicting regulation of the two parental genomes with a predominance of the paternal genotype which affects genes such as *MAGEL2* and *IGF2*. These results are a first step in understanding the global system biology of piglet maturity. Some genes described in this report could be candidates to explore the genetic control of maturity. Further functional and genetical studies may be focused on the LW breed with its increased mortality at birth, and then be continued to identify the genetic mechanisms underlying the differences in maturity.

Methods

Ethical statement

All animal use was performed under European Union legislation (directive 86/609/EEC) and French legislation of région Midi-Pyrénées in France (Décret no:2001-464 29/05/01; <http://ethique.ipbs.fr/sdv/charteexpeanimale.pdf>; accreditation for animal housing number C-35-275-32). The technical and scientific staff obtained individual accreditation (Ref: MP/01/01/01/11) from the ethics committee (région Midi-Pyrénées - France;

<http://comethmp.ipbs.fr/>) to experiment on living animals. All pigs used in this study were males and were obtained by caesarean.

Experimental design and RNA preparation

To assay for changes in gene expression during piglet maturity, mRNA was isolated from 64 fetal muscle samples (longissimus dorsi) in 8 different conditions: two fetal gestational ages (day 90 and day 110) associated with four fetal genotypes. The four fetal genotypes consisted of two extreme breeds for mortality at birth (LW and MS) and two crosses (MSLW and LWMS). MS and LW sows were inseminated with mixed semen (LW and MS) so that each litter was composed of purebred fetuses (LW or MS) and crossbred fetuses (LWMS from MS sows and MSLW from LW sows). Total RNA was isolated from each of the 64 muscle samples. Briefly, muscle samples were disrupted, homogenized and ground to a fine powder by rapid agitation for 1 min in a liquid-nitrogen cooled grinder with stainless steel beads. An aliquot of 100 mg of the fine powder was then processed for total RNA isolation and purification using Trizol (Invitrogen, France) and the Nucleospin RNA II kit (Macherey-Nagel, France) according to the manufacturer's instructions. The method included a DNase digestion step to remove contaminating DNA. The extracted total RNA was eluted in 300 μ l of RNase-free water and stored at -80°C. RNA quality and concentration were verified using an Agilent 2100 bioanalyzer (RNA solutions and RNA 6000 Nano Lab- Chip Kit, Agilent Technologies France, Massy, France).

Microarray description

The microarray GPL16524 (Agilent technology, 8 × 60K) used in this experiment consisted in 43,603 spots derived from the 44K (V2:026440 design) Agilent porcine specific microarray, 9,532 genes from adipose tissue, 3,776 genes from the immune system and 3,768 genes from skeletal muscle (Liaubet et al. (personal communication), unpublished data).

After quality control and a quantile normalization step as described in [48], the data of fluorescence signal from 61 microarrays containing 44,368 spots were kept for further analysis and log₂ transformed. These spots correspond to 34,945 annotated genes, i.e. 16,712 unique annotated genes. It is important to consider that the annotations are constantly being improved due to annotation issues in pig. However, all cited genes were checked. Raw data and information are available in NCBI (GEO accession number GSE56301).

Statistical analysis

Statistical analyzes were performed with R 3.0.2 software [49]. To analyze jointly differences between breeds and

gestational ages, the following mixed linear model was fitted to each probe (R nlme package, lme function [50]):

$$y_{ijk} = \mu + A_i + FG_j + A.FG_{ij} + S_k + \epsilon_{ijk} \quad (1)$$

with $i \in \{d90, d110\}$, $j \in \{LW, MS, LWMS, MSLW\}$, $k = 1, \dots, 18$, $S_k \sim N(0, \sigma_S^2)$ independent and identically distributed (iid) and $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$ iid. S_k and ϵ_{ijk} are mutually independent. y_{ijk} is the expression of the probe (gene) being studied, μ a general mean of the considered gene expression and ϵ_{ijk} is a residual. This model includes two fixed effects and their interaction: A_i is the effect of fetal gestational age i , FG_j the effect of fetal genotype j and $A.FG_{ij}$ the interaction effect between gestational age i and genotype j . S_k represents the random sow effect.

To identify differentially expressed probes (DEPs) for gestational age and/or genotype, the mixed linear model (1) was fitted to the microarray data. A F-type test was performed by comparing the complete model (1) and the reduced model $y_{ijk} = \mu + S_k + \epsilon_{ijk}$. A correction for multiple testing was then implemented using Bonferroni [51] or False Discovery Rate (FDR) [51,52] using the multtest R package [53].

The list of DEPs was then partitioned into 4 sub-models. Sub-model 1 combined the two fixed effects and their interaction. Sub-model 2 involved the two fixed effects in an additive manner. Sub-model 3 included only the fetal gestational age effect whereas sub-model 4 included only the fetal genotype effect. All models included the random sow effect. In summary:

$$\begin{cases} \text{Sub - model1} : y_{ijk} = \mu + A_i + FG_j + A.FG_{ij} + S_k + \epsilon_{ijk} \\ \text{Sub - model2} : y_{ijk} = \mu + A_i + FG_j + S_k + \epsilon_{ijk} \\ \text{Sub - model3} : y_{ijk} = \mu + A_i + S_k + \epsilon_{ijk} \\ \text{Sub - model4} : y_{ijk} = \mu + FG_j + S_k + \epsilon_{ijk} \end{cases} \quad (2)$$

The Bayesian Information Criterion (BIC) was used to associate each DEP with one of these four sub-models.

To analyze the parental impact, the following mixed linear models involving the two parental genotypes were fitted to each probe:

$$y_{ijkl} = \mu + A_i + MG_j + PG_k + A.MG_{ij} + A.PG_{ik} + MG.PG_{jk} + S_l + \epsilon_{ijkl} \quad (3)$$

with $i \in \{d90, d110\}$, j and $k \in \{LW, MS, LWMS, MSLW\}$, $l = 1, \dots, 18$, $S_l \sim N(0, \sigma_S^2)$ independent and identically distributed (iid) and $\epsilon_{ijkl} \sim N(0, \sigma_\epsilon^2)$ iid. S_l and ϵ_{ijkl} are mutually independent. y_{ijkl} is the expression of the probe (gene) being studied, μ a general mean of the considered gene expression and ϵ_{ijkl} is a residual. This model includes two fixed effects and their interaction: A_i is the effect of fetal gestational age i . MG_j is the effect of maternal genotype j and $A.MG_{ij}$ the interaction effect between gestational age i and maternal genotype j . PG_k is the effect of paternal genotype k and $A.PG_{ik}$ the interaction

effect between gestational age i and paternal genotype k . $MG.PG_{jk}$ is the interaction between parental genotypes.

To identify DEPs, these mixed linear models were fitted to the microarray data. F-type tests were performed by comparing the complete model (3) and two reduced models: one without the interaction between gestational ages and maternal genotype to identify genes influenced by the maternal genotype and the other without the interaction between gestational ages and paternal genotype to identify genes influenced by the paternal genotype. A correction for multiple tests was then implemented using Bonferroni [51] or FDR [51,52] using the multtest R package [53] as previously.

Gene Ontology functional enrichment analysis

Functional annotation of genes from sub-model 1 based on Gene Ontology (GO) was provided by GeneCoDis 3.0 software [54]. Enrichment analysis was applied to lists of genes selected for an absolute log₂-fold change greater than $\frac{1}{2}$ between both gestational ages. This threshold of a $\frac{1}{2}$ -log₂-fold change was used to obtain two values: up-regulated genes at day 90 or up-regulated genes at day 110. This threshold was applied to ensure that only the genes with a minimal change between gestational ages were retained for the GO functional enrichment analysis. The two lists contained up-regulated genes at 90 days and up-regulated at 110 days respectively. To set the statistical enrichment of a particular biological function, a hypergeometric test was used. Resulting p-values were adjusted for multiple tests using the FDR approach (FDR < 1%).

Relevance network

A relevance network is a graphical model displaying genes as edges and relationships (correlations here) between genes as vertexes [55]. In our study, a network building pipeline, using the igraph R package [56], was decomposed into three steps. In a first step, a similarity matrix S was calculated using the Pearson correlation coefficient between pairs of genes. In a second step, S was transformed into a binary adjacency matrix A using hard thresholding. This matrix A was composed of 0 and 1 depending on whether the correlation coefficient was lower or greater than 0.98 (in absolute value), respectively. The final step consisted in a graph representation of A . An edge was present between two nodes (genes) i and j if the value a_{ij} in A was 1.

To determine communities in the graph, a fast-greedy algorithm was used to optimize the modularity of a partition of the network [57]. The modularity is a measure of the quality of communities in the network: highly connected genes within each community, and lowly connected genes between communities. Finally, GO functional enrichment analysis was performed to determine enriched biological processes in each community. In

addition to dividing the network structure into sub-networks, influential genes were highlighted as described in Villa et al. [19] based on other criteria, i.e. degree and betweenness centrality.

Quantitative real time RT-PCR analysis for gene expression

Gene primers were designed from pig genes taking into account intron-exon organization using Primer3 software (<http://frodo.wi.mit.edu/primer3/>). Sequences are available in Additional file 11. RNA samples were reverse transcribed from 1 μ g as previously described in [58]. The resulting cDNA samples were completed to 50 μ l. The assay for each gene consisted of four replicates per genotype and development stage (from the 61 used in the microarray experiment) and negative controls.

The expression of 10 genes was analyzed using 48.48 Dynamic Array™ IFCs and the BioMark™ HD System from Fluidigm. Two specific target amplifications (STA) were performed on cDNA muscle samples according to the manufacturer's recommendations. As previously described [59], a 14 cycle STA treated with Exonuclease I was performed, diluted and transferred to the BioMark™ HD for final STA. The efficiency of PCR amplification was determined specifically for each gene, by serially diluting (1, 1:2; 1:2; 1:2) the muscle cDNA pool.

Data was then analyzed using Fluidigm Digital PCR Analysis software with the Linear (Derivative) Baseline Correction Method.

After determination of the threshold cycle (Ct), the Pfaffl method [60] was applied as described in [59] to calculate the relative expression of each gene. *HPRT*, which was not regulated during the maturation process, was used as the reference gene. Pearson's correlations were calculated between microarray expression and qPCR values.

Availability of supporting data

Microarray data are MIAME compliant and available in Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) through the accession number GSE56301.

Additional files

Additional file 1: Complete list of differentially expressed probes of sub-model 1 '.xlsx' file. Features of 2000 genes of sub-model 1 (Gene symbol, gene name, probe name, p-value and log₂-fold change (d110/d90)). P-value, Bonferroni and FDR were obtained by a F-type test comparing the complete model (1) and the reduced model $y = \mu + S + \epsilon$.

Additional file 2: Complete list of enriched GO (Biological Process (BP), Molecular Function (MF) and Cellular Component (CC)) at day 90 '.xlsx' file. The file gives the GO items, the corresponding functions, the classes of ontology, the lists of genes, the numbers of genes in the input lists and the reference lists and p-values (unadjusted and FDR).

Additional file 3: Complete list of enriched GO (BP, MF and CC) at day 110 '.xlsx' file. The file gives the GO items, the corresponding functions,

the class of ontology, the list of genes, the number of genes in the input list and the reference list and p-values (unadjusted and FDR).

Additional file 4: Complete list of the first twelve enriched GOBP at day 90 or at day 110 in LW or MS '.xlsx' file. The first twelve enriched GOBP in LW and MS at day 90 and day 110 (Items, functions, gene lists and p-value (unadjusted and FDR)). In red, genes are up-regulated in MS only and in blue, genes are up-regulated in LW only. In black, genes are up-regulated in LW and MS.

Additional file 5: Frequency distribution of Pearson's correlation network '.pdf' file. Frequency distribution of Pearson's correlation between the entire set of 1516 genes (annotated or not) used to build our network.

Additional file 6: Features of genes in the relevance network '.xlsx' file. Gene names, probe name, degree, betweenness centrality and community of genes in the relevance network.

Additional file 7: Complete list of enriched GO (BP, MF and CC) in the four communities of the relevance network '.xlsx' file. The file gives the GO items, the corresponding functions, the classes of ontology, the lists of genes, the numbers of genes in the input lists and the reference lists and p-values (unadjusted or FDR).

Additional file 8: Complete list of differentially expressed probes impacted by maternal genome '.xlsx' file. Features of 164 DEPs using model (3) (Gene symbol, gene name, probe name, p-value and *Sus scrofa* chromosome localization). P-value, Bonferroni and FDR were obtained by a F-type test comparing the complete model (3) and the reduced model without interaction between gestational age and maternal genotype.

Additional file 9: Complete list of differentially expressed probes impacted by paternal genome '.xlsx' file. Features of 641 DEPs using model (3) (Gene symbol, gene name, probe name, p-value and *Sus scrofa* chromosome localization). P-value, Bonferroni and FDR were obtained by a F-type test comparing the complete model (3) and the reduced model without interaction between gestational age and paternal genotype.

Additional file 10: Box-plot representation of the 10 tested genes in qPCR compared to their microarray expression '.pdf' file. (A) ARG2. (B) PHKA1. (C) SLC38A4. (D) DLK1. (E) RASGRP3. (F) GPD1. (G) DUT. (H) GBP1. (I) IL1RAPL2. (J) SPG7. All box-plots are normalized in log₂.

Additional file 11: Features of genes tested by real time RT-PCR '.xlsx' file. Gene names, description and primer (up and down) of the 10 genes tested by real time RT-PCR.

Abbreviations

BH: Benjamini-Hochberg; BIC: Bayesian information criterion; BP: Biological process; CC: Cellular component; CDK6: Cyclin-dependent kinase 6; CNV: Copy number variations; COL5A2: Collagen alpha-2(V) chain; CREM: cAMP responsive element modulator; DCI or ECI1: Enoyl-CoA delta Isomerase 1; DEP: Differentially expressed probes; FDR: False discovery rate; iid: independent and identically distributed; GO: Gene ontology; IGF2: Insulin-like growth factor 2; LDHA: Lactate dehydrogenase A; LW: Large white; MAGEL2: MAGE-like 2; MF: Molecular function; MS: Meishan; MYH3: Embryonic myosin heavy chain; NUSAP1: Nucleolar and spindle associated protein 1; OAA: Oxaloacetate; PCA: Principal component analysis; PC: Principal component; PGK1: Phosphoglycerate kinase 2; PCK2 or PEPCK: Phosphoenolpyruvate carboxykinase 2; PEP: Phosphoenolpyruvate; PYGM: Glycogen phosphorylase; qRT-PCR: Quantitative real time polymerase chain reaction; QTL: Quantitative trait locus; SGOL1: Shugoshin-like 1; SORD: Sorbitol dehydrogenase; STA: Specific target amplifications; STMN1: Stathmin 1; UNK: Unknown.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LL and LC conceived and designed the study. YB supervised the performance testing, from animal production to biological sampling. NI was the project data manager. CL extracted RNA and performed control quality. YL and PM provided the transcriptomic data set. FV performed qPCR. VV analyzed the expression data set supervised by MSC and LL. VV carried out the enrichment and network analyzes supervised by MSC and LL. VV drafted the manuscript

with help of MSC and LL. LL supervised the project. All authors read and approved the final manuscript.

Acknowledgments

This project received financial support from French National Agency of Research (PORCINET project, ANR-09-GENM005). VV is a PhD fellow supported by the INRA GA (Génétique Animale), the INRA PHASE (Physiologie Animale et Systèmes d'Élevage) and the région Midi-Pyrénées. Microarray annotations were managed by SIGENAE (Système d'Information des GENomes des Animaux d'Élevage, <http://www.sigena.org>). The authors would like to thank Jasha Leenhouders, Agnès Bonnet for help, and also Helen Mundutéguy-Hutchings for the English revision. The authors thank the reviewers and the editors for their useful comments and suggestions that helped improve the quality of this paper.

Author details

¹INRA, UMR1388 Génétique, Physiologie et Systèmes d'Élevage, F-31326 Castanet-Tolosan, France. ²Université de Toulouse INPT ENSAT, UMR1388 Génétique, Physiologie et Systèmes d'Élevage, F-31326 Castanet-Tolosan, France. ³Université de Toulouse INPT ENVT, UMR1388 Génétique, Physiologie et Systèmes d'Élevage, F-31076 Toulouse, France. ⁴INSA, Département de Génie Mathématiques, F-31077 Toulouse, France. ⁵Université de Toulouse, UMR 5219 Institut de Mathématiques, F-31077 Toulouse, France. ⁶INRA, UMR1331 ToxAlim, F-31027 Toulouse, France. ⁷INRA, UE1372 GenESI, F-17700 Surgères, France.

Received: 12 May 2014 Accepted: 11 September 2014

Published: 17 September 2014

References

- Canario L: **Aspects génétiques de la mortalité des porcelets à la naissance et en allaitement précoce: relations avec les aptitudes maternelles des truies et la vitalité des porcelets.** PhD thesis. Institut National Agronomique Paris-Grignon; 2006.
- Strange T, Ask B, Nielsen B: **Genetic parameters of the piglet mortality traits stillborn, weak at birth, starvation, crushing, and miscellaneous in crossbred pigs.** *J Anim Sci* 2013, **91**(4):1562–1569.
- Serenius T, Muhonen P, Stalder K: **Economic values of pork production related traits in finland.** *Agric Food Sci* 2007, **16**:79–88.
- Tuchscherer M, Puppe B, Tuchscherer A, Tiemann U: **Early identification of neonates at risk: traits of newborn piglets with respect to survival.** *Theriogenology* 2000, **54**:371–388.
- Miller DR, Blache D, Jackson RB, Downie EF, Roche JR: **Metabolic maturity at birth and neonate lamb survival: Association among maternal factors, litter size, lamb birth weight, and plasma metabolic and endocrine factors on survival and behavior.** *J Anim Sci* 2010, **88**(2):581–593.
- Lawn JE, Cousens S, Zupan J: **Team LNSS: 4 million neonatal deaths: when? where? why?** *The Lancet* 2005, **365**(9462):891–900.
- Baxter EM, Jarvis S, D'Eath RB, Ross DW, Robson SK, Farish M, Nevison IM, Lawrence AB, Edwards SA: **Investigating the behavioural and physiological indicators of neonatal survival in pigs.** *Theriogenology* 2008, **69**:773–783.
- van der Lende T, Knol EF, Leenhouders J: **Prenatal development as a predisposing factor for perinatal losses in pigs.** *Reprod Suppl* 2001, **58**:247–261.
- Leenhouders J, Knol EF, de Groot PN, Vos H, van der Lende T: **Fetal development in the pig in relation to genetic merit for piglet survival.** *J Anim Sci* 2002, **80**(7):1759–1770.
- Leenhouders J, Knol EF, van der Lende T: **Differences in late prenatal development as an explanation for genetic differences in piglet survival.** *J Anim Sci* 2002, **78**:57–62.
- Herpin P, Damon M, LeDividich J: **Development of thermoregulation and neonatal survival in pigs.** *Livestock Production Sci* 2002, **78**(1):25–45.
- Canario L, Pèrè MC, Tribut T, Thomas F, David C, Gogué J, Herpin P, Bidanel JP, Le Dividich J: **Estimation of genetic trends from 1977 to 1998 of body composition and physiological state of large white pigs at birth.** *Animal* 2007, **1**:1409–1413.
- Herpin P, Lossec G, Schmidt I, Cohen-Adad F, Duchamp C, Lefaucheur L, Goglia F, Lanni A: **Effect of age and cold exposure on morphofunctional characteristics of skeletal muscle in neonatal pigs.** *Pflugers Archiv* 2002, **444**(5):610–618.
- Bielanska-Osuchowska Z: **Ultrastructure and stereological studies of hepatocytes in prenatal development of swine.** *Folia Morphol* 1996, **55**:1–193.
- Cagnazzo M, te Pas MF, Priem J, de Wit AA, Pool MH, Davoli R, Russo V: **Comparison of prenatal muscle tissue expression profiles of two pig breeds differing in muscle characteristics.** *J Anim Sci* 2006, **84**(1):1–10.
- Xu Y, Qian H, Feng X, Xiong Y, Lei M, Ren Z, Zuo B, Xu D, Ma Y, Yuan H: **Differential proteome and transcriptome analysis of porcine skeletal muscle during development.** *J Proteomics* 2012, **75**(7):2093–2108.
- Canario L, Cantoni E, Le Bihan E, Caritez JC, Billon Y, Bidanel JP, Foulley JL: **Between-breed variability of stillbirth and its relationship with sow and piglet characteristics.** *J Anim Sci* 2006, **84**(12):3185–3196.
- Markowetz F, Spang R: **Inferring cellular networks - a review.** *BMC Bioinformatics* 2007, **8**(Suppl 6):5.
- Villa-Vialaneix N, Liaubet L, Laurent T, Cheral P, Gamot A, SanCristobal M: **The structure of a gene co-expression network reveals biological functions underlying eqtls.** *PLoS ONE* 2013, **8**(4):60045.
- Hazard D, Liaubet L, SanCristobal M, Mormede P: **Gene array and real time pcr analysis of the adrenal sensitivity to adrenocorticotrophic hormone in pig.** *BMC Genomics* 2008, **9**(1):101.
- Foxcroft GR, Dixon WT, Novak S, Putman CT, Town SC, Vinsky MDA: **The biological basis for prenatal programming of postnatal performance in pigs.** *J Anim Sci* 2006, **84**(13 suppl):105–112.
- Wilson ME, Biensen NJ, Youngs CR, Ford SP: **Development of meishan and yorkshire littermate conceptuses in either a meishan or yorkshire uterine environment to day 90 of gestation and to term.** *Biol Reprod* 1998, **58**(4):905–910.
- Biensen NJ, Wilson ME, Ford SP: **The impact of either a meishan or yorkshire uterus on meishan or yorkshire fetal and placental development to days 70, 90, and 110 of gestation.** *J Anim Sci* 1998, **76**(8):2169–76.
- Biensen NJ, Wilson ME, Ford SP: **The impacts of uterine environment and fetal genotype on conceptus size and placental vascularity during late gestation in pigs.** *J Anim Sci* 1999, **77**(4):954–9.
- Lefaucheur L, Edom F, Ecolan P, Butler-Browne GS: **Pattern of muscle fiber type formation in the pig.** *Dev Dyn* 1995, **203**(1):27–41.
- Picard B, Lefaucheur L, Berri C, Duclos MJ: **Muscle fibre ontogenesis in farm animal species.** *Reprod Nutr Dev* 2002, **42**:415–431.
- Lefaucheur L, Milan D, Ecolan P, Le Callennec C: **Myosin heavy chain composition of different skeletal muscles in large white and meishan pigs.** *J Anim Sci* 2004, **82**(7):1931–1941.
- Lefaucheur L, Ecolan P, Losse G, Gabillard JC, Butler-Browne GS, Herpin P: **Influence of early cold exposure on myofiber maturation in pig skeletal muscle.** *J Muscle Res Cell Motil* 2001, **22**:439–452.
- Xu Y, Shi T, Cai H, Zhou Y, Lan X, Zhang C, Lei C, Qi X, Chen H: **Associations of myh3 gene copy number variations with transcriptional expression and growth traits in chinese cattle.** *Gene* 2014, **535**(2):106–111.
- Raemaekers T, Ribbeck K, Beaudouin J, Annaert W, Van Camp M, Stockmans I, Smets N, Bouillon R, Ellenberg J, Carmeliet G: **Nusap, a novel microtubule-associated protein involved in mitotic spindle organization.** *J Cell Biol* 2003, **162**(6):1017–1029.
- Chou AY, Wang TH, Lee SC, Hsu PH, Tsai MD, Chang CN, Jeng YM: **Phosphorylation of nusap by cdk1 regulates its interaction with microtubules in mitosis.** *Cell Cycle* 2011, **10**:4083–4089.
- Belmont LD, Mitchison TJ: **Identification of a protein that interacts with tubulin dimers and increases the catastrophe rate of microtubules.** *Cell Press* 1996, **84**(4):632–631.
- Harbour JW, Luo RX, Dei Santi A, Postigo AA, Dean DC: **Cdk phosphorylation triggers sequential intramolecular interactions that progressively block rb functions as cells move through g1.** *Cell Press* 1999, **98**(6):859–869.
- Drummond MJ, McCarthy JJ, Sinha M, Spratt HM, Volpi E, Esser KA, Rasmussen BB: **Aging and microRNA expression in human skeletal muscle: a microarray and bioinformatics analysis.** *Physiol Genomics* 2011, **43**(10):595–603.
- Trayhurn P, Temple NJ, Van Aerde J: **Evidence from immunoblotting studies on uncoupling protein that brown adipose tissue is not present in the domestic pig.** *Can J Physiol Pharmacol* 1989, **67**:1480–1485.

36. van Weeghel M, te Brinke H, van Lenthe H, Kulik W, Minkler PE, Stoll MSK, Sass JO, Janssen U, Stoffel W, Schwab KO, Wanders RJA, Hoppel CL, Houten SM: **Functional redundancy of mitochondrial enoyl-coa isomerases in the oxidation of unsaturated fatty acids.** *FASEB J* 2012, **26**(10):4316–4326.
37. Inoue E, Yamauchi J: **Amp-activated protein kinase regulates [PEPCK] gene expression by direct phosphorylation of a novel zinc finger transcription factor.** *Biochem Biophys Res Commun* 2006, **351**(4):793–799.
38. Cotter DG, Ercal B, d'Avignon DA, Dietzen DJ, Crawford PA: **Impact of peripheral ketolytic deficiency on hepatic ketogenesis and gluconeogenesis during the transition to birth.** *J Biol Chem* 2013, **288**(27):19739–19749.
39. Mellor DJ, Cockburn F: **A comparison of energy metabolism in the new-born infant, piglet and lamb.** *Exp Physiol* 1986, **71**(3):361–379.
40. Herpin P, Le Dividich J, Amaral N: **Effect of selection for lean tissue growth on body composition and physiological state of the pig at birth.** *J Anim Sci* 1993, **71**(10):2645–53.
41. de Koning D-J, Rattink AP, Harlizius B, van Arendonk JAM, Brascamp EW, Groenen MAM: **Genome-wide scan for body composition in pigs reveals important role of imprinting.** *Proc Natl Acad Sci U S A* 2000, **97**(14):7947–7950.
42. Nezer C, Moreau L, Brouwers B, Coppieters W, Detilleux J, Hanset R, Karim L, Kvasz A, Leroy P, Georges M: **An imprinted qtl with major effect on muscle mass and fat deposition maps to the igf2 locus in pigs.** *Nat Genet* 1999, **21**:155–156.
43. Zhang FW, Han ZB, Deng CY, He HJ, Wu Q: **Conservation of genomic imprinting at the ndn, magel2 and mest loci in pigs.** *Genes Genet Syst* 2012, **87**(1):53–58.
44. Van Laere S, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M, Adersson G, Georges M, Adersson L: **A regulatory mutation in igf2 causes a major qtl effect on muscle growth in the pig.** *Nature* 2003, **425**:832–836.
45. Haig D: **The kinship theory of genomic imprinting.** *Annual Rev Ecol Syst* 2000, **31**:9–32.
46. Bischoff SR, Tsai S, Hardison N, Motsinger-Reif AA, Freking BA, Nonneman D, Rohrer G, Piedrahita JA: **Characterization of conserved and nonconserved imprinted genes in swine.** *Biol Reprod* 2009, **81**(5):906–920.
47. Piedrahita JA: **The role of imprinted genes in fetal growth abnormalities.** *Birth Defects Res A Clin Mol Teratol* 2011, **91**(8):682–692.
48. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19**(2):185–193.
49. R Core Team: *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2013.
50. Lindstrom MJ, Bates DM: **Nonlinear mixed effects models for repeated measures data.** *Biometrics* 1990, **46**:673–687.
51. Shaffer JP: **Multiple hypothesis testing.** *Annu Rev Psychol* 1995, **46**:561–584.
52. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Statist Soc B* 1995, **57**:289–300.
53. Pollard KS, Dudoit S, van der Laan MJ: **Multiple testing procedures: R multtest package and applications to Genomics.** In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Edited by Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S. New York: Springer (Statistics for Biology and Health Series); 2005:251–272.
54. Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A: **Genecodis3: a non-redundant and modular enrichment analysis tool for functional genomics.** *Nucleic Acids Res* 2012, **40**:478–483.
55. Butte AJ, Kohane IS: **Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.** In *Proceedings of the Pacific Symposium on Biocomputing.* Edited by Lauderdale K. New York: World Scientific Publishing Company; 2000:418–429.
56. Csardi G, Nepusz T: **The igraph software package for complex network research.** *Inter J Complex Syst* 2006, **1695**:1695–1708.
57. Clauset A, Newman MEJ, Moore C: **Finding community structure in very large networks.** *Phys Rev* 2004, **70**:066111.
58. Bonnet A, Bevilacqua C, Benne F, Bodin L, Cotinot C, Liaubet L, Sancristobal M, Sarry J, Terenina E, Martin P, Tosser-Klopp G, Mandon-Pepin B: **Transcriptome profiling of sheep granulosa cells and oocytes during early follicular development obtained by laser capture microdissection.** *BMC Genomics* 2011, **12**(1):417.
59. Bonnet A, Cabau C, Bouchez O, Sarry J, Marsaud N, Foissac S, Woloszyn F, Mulsant P, Mandon-Pepin B: **An overview of gene expression dynamics during early ovarian folliculogenesis: specificity of follicular compartments and bi-directional dialog.** *BMC Genomics* 2013, **14**(1):904.
60. Pfaffl MW: **A new mathematical model for relative quantification in real-time rt-pcr.** *Nucleic Acids Res* 2001, **29**(9):45.

doi:10.1186/1471-2164-15-797

Cite this article as: Voillet et al.: Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity. *BMC Genomics* 2014 **15**:797.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



2.2.1 Quelques mises à jour et commentaires

En premier lieu, dans cet article, nous avons souligné la puissance du dispositif expérimental, notamment l'âge gestationnel étant une des sources majeures de variation (Figure 1 de l'Article 1). Afin de fournir plus de précisions concernant le dispositif du projet, la Table 2.1 ci-dessous détaille le nombre d'échantillons utilisés par condition pour le transcriptome musculaire. Par ailleurs, le nombre de truies (comme indiqué dans l'article) est de 18, avec 2 à 5 fœtus par truie.

	MS	LW	MSLW	LWMS
d90	7	8	8	8
d110	8	9	5	8

Table 2.1 – **Nombre d'échantillons par conditions du transcriptome musculaire.**

Concernant l'analyse biologique des sondes déclarées différentielles, bien que tous les sous-modèles proposés soient intéressants, nous avons choisi de nous limiter uniquement à l'étude des sondes obtenues à partir du sous-modèle 1 (correspondant aux sondes différentielles pour l'interaction entre l'âge gestationnel et le génotype fœtal). Ces sondes sont à priori celles qui concernent plus les mécanismes de maturité comme étant différentielles à la fin du développement fœtal (effet stade) mais aussi influencées par l'effet génotype et donc présentant des expressions différentes entre LW et MS. Nous aurions pu aussi nous intéresser au modèle additif comme pouvant apporter des informations complémentaires sur le phénotype d'intérêt. Les sondes obtenues à partir des autres sous-modèles ont néanmoins été conservées pour de possibles futures analyses.

Au cours de l'écriture de cet article, nous avons aussi choisi d'inférer un réseau de co-expression utilisant la corrélation de Pearson. Il existe d'autres méthodes d'inférence, et elles seront décrites plus en détails dans la section 3.2.2.2 du chapitre suivant. Brièvement, il existe notamment des méthodes d'inférence de type modèle graphique gaussien (GGM), utilisant la corrélation partielle (déterminée à partir de la matrice variance-covariance). Une première approche développée par Schäfer & Strimmer (2005a) consiste en l'utilisation d'une méthode «d'estimateur à retrécisseur» (ou *shrinkage*) afin d'estimer la matrice de variance-covariance. En

outre, une approche alternative a également été proposée avec l'utilisation de la régression régularisée où l'estimation et la sélection sont effectuées simultanément en utilisant la pénalité *sparse*. Toutefois, comme souligné dans l'article (page 5), ces méthodes sont difficilement applicables lorsque (i) le nombre de variables p est beaucoup plus grand que le nombre d'individus n , ou lorsque (ii) les variables sont fortement corrélées entre elles, ce qui était le cas dans l'Article 1. De plus, les méthodes GGM sont généralement adaptées aux variables gaussiennes ($Y \sim \mathcal{N}_p(\mu, \Sigma)$), ce qui n'était pas le cas dans cet article (très fort effet de l'âge gestationnel, distribution bimodale). D'autre part, le nombre de variables (environ 1 500 gènes) était relativement grand pour utiliser ces stratégies d'un point de vue temps de calcul, en particulier pour les stratégies de pénalisation.

Par ailleurs, comme discuté dans l'introduction, un pré-filtrage est souvent nécessaire avant de pouvoir effectuer les analyses. Ici, avant l'inférence de réseau, nous avons choisi d'utiliser uniquement les gènes annotés uniques et les gènes non-annotés du sous-modèle 1. Ce pré-filtrage nous semblait correct afin d'obtenir un nombre adéquat de gènes. Cependant, il est possible que ce filtrage nous ait limité. En effet, nous avons mis en relation uniquement les gènes modulés selon l'interaction entre l'âge gestationnel et le génotype fœtal. Nous aurions pu utiliser un autre type de filtrage (comme l'utilisation d'un autre modèle) afin d'obtenir une autre liste de gènes, ces gènes n'auraient probablement pas tous été différentiels selon le sous-modèle 1. Toutefois, dans l'article, comme souligné précédemment, nous avons choisi de nous focaliser uniquement sur les sondes du sous-modèle 1 qui nous semble plus proche de la question de la maturité.

Pour finir, je souhaite également apporter quelques détails sur la validation des résultats par qPCR. Les p-valeurs présentées dans la Table 2 de l'article correspondent à la significativité de la corrélation entre les valeurs issues de la biopuce et les valeurs issues de la qPCR pour un gène. Une p-valeur significative ne veut pas automatiquement souligner une forte corrélation (bien que ce soit le cas dans notre étude), mais une corrélation significativement différente de 0. La qPCR et les biopuces sont deux techniques différentes. Des corrélations significatives comme présentées dans l'article peuvent quelquefois être biaisées. Par exemple, au niveau de la qPCR, le design des amorces est une étape très limitante et fastidieuse. Il est possible, entre autres, de tomber sur des régions contenant un polymorphisme ou un site d'épissage alternatif, ce qui peut affecter la corrélation. Une faible corrélation

peut être observée due à la différence entre ces techniques, même si ces dernières mesurent le même type de variables (les gènes).

Il existe d'autres stratégies permettant de comparer des données issues de deux plateformes différentes. Par exemple, le sous-modèle 1 aurait pu être développé sur les 10 gènes d'intérêt de la biopuce, puis un autre sous-modèle 1 sur les gènes de la qPCR, afin d'observer et de comparer si les mêmes conclusions sont obtenues à partir de ces deux techniques.

Chapitre 3

Intégration de données omiques musculaires

3.1 Introduction

Dans la première partie ce chapitre, différentes méthodes existantes pour l'inférence de réseaux en biologie seront introduites. Une stratégie d'intégration, utilisant des réseaux, est ensuite décrite dans la deuxième partie, correspondant à un article en préparation. L'objectif fut de combiner des données protéomiques, transcriptomiques et phénotypiques afin de décrire le processus de maturation musculaire chez le porc. Les mêmes individus sont présents dans ces trois jeux de données, où, comme décrit précédemment, les fœtus (purs et croisés) provenant de deux races extrêmes pour la mortalité à la naissance, Large White (LW) et Meishan (MS), ont été utilisés.

Une méthode d'intégration originale a été proposée. Cette stratégie correspond à une analyse de réseaux intégrés afin d'explorer les relations entre des réseaux de co-expressions protéiques, construits à partir des données protéiques, et des phénotypes d'intérêts (glycogène et myosines). La corrélation entre le protéome et le transcriptome, par l'utilisation de tests de corrélation, a ensuite été examinée pour observer de possibles régulations transcriptionnelles. Une p-valeur significative pour une variable indiquait une corrélation non-nulle entre les données protéiques et transcriptomiques. Avec l'utilisation de données d'abondance de 113 spots protéiques (dont 89 protéines uniques), l'inférence de réseau, les corrélations

avec les phénotypes et l'enrichissement fonctionnel ont souligné que le métabolisme énergétique oxydatif semble être un déterminant clé de la maturité musculaire néonatale. Quelques protéines, comme ATP5A1 et CKMT2, ont été identifiées comme étant des nœuds importants dans les réseaux et étant fortement liées à la mise en place du métabolisme énergétique musculaire. En outre, 31 protéines avaient une corrélation positive et significative entre leurs expressions protéiques et leurs expressions géniques, suggérant une possible régulation transcriptionnelle dans les deux génotypes extrêmes. Parallèlement, grâce à une étude d'enrichissement bibliographique avec le logiciel Ingenuity (*upstream regulators*), des facteurs de transcription, comme *PPARGC1A* et *ESR1*, ont également été proposés, avec de possibles effets sur la fin de développement musculaire fœtal.

Tous ces résultats ont été décrits dans un article (en préparation, il sera soumis très prochainement) et sont présentés dans la seconde partie de ce chapitre.

3.2 Les réseaux en biologie

3.2.1 Introduction

Au cours des dernières années, l'intérêt pour la construction de réseaux en biologie s'est intensifié avec l'arrivée massive des données de grande dimension, au point de devenir l'un des domaines de recherche très actif de la biologie des systèmes. L'utilisation des réseaux (ou graphes en termes mathématiques) en biologie est très utilisé, notamment en transcriptomique (Stuart *et al.*, 2003), en protéomique (Sabido *et al.*, 2012) et en métabolique (Thiele & Palsson, 2010). Un graphe est défini comme étant un ensemble d'objets représentés par des nœuds et des arêtes modélisant les relations entre ceux-ci (Figure 3.1). Les nœuds dépendent des données analysées ; ils peuvent être des gènes, des protéines ou des métabolites. En outre, en fonction des données utilisées pour la construction du graphe, la nature des arêtes peut représenter différents types de relations entre variables (nœuds). On trouve par exemple des liens de co-expression, d'interaction physique ou encore de co-localisation. Les réseaux permettent ainsi d'étudier et d'élucider les différentes interactions possibles entre variables, l'étude de ces relations permettant de faciliter l'interprétation des données (Aittokallio & Schwikowski, 2006). Des analyses utilisant des réseaux biologiques ont, par exemple, permis de décrire et détailler les bases moléculaires et cellulaires de certaines maladies complexes (Barabási *et al.*, 2011). La comparaison des interactions moléculaires et cellulaires d'un individu sain à celles d'un individu malade peut quelques fois mener à l'identification de gènes (protéines ou métabolites) fortement impliqués dans l'établissement de la maladie, offrant ainsi de possibles meilleures cibles thérapeutiques pour le développement de drogues médicinales. Suite à l'inférence, des modules (aussi appelés clusters ou communautés) peuvent être mis en évidence et attirer l'attention sur des groupes de variables partageant des propriétés communes et/ou possiblement impliquées dans une même action ou processus biologique (Figure 3.1).

De manière globale, indépendamment de la nature des interactions, deux principaux types de réseaux existent : orienté ou non-orienté. Les réseaux dits non-orientés sont composés d'arêtes n'ayant aucune direction entre variables (Figure 3.2A). Par exemple, un réseau d'interactions protéiques est de type non-orienté : un lien y représente une relation physique entre deux protéines. Les réseaux orientés sont,

quant à eux, constitués d'arêtes (appelés arcs) ayant une direction entre les variables (Figure 3.2B). Ainsi, une arête allant du nœud v_i au nœud v_j est différente d'une arête allant du nœud v_j au nœud v_i . Comme illustration, sachant que de nombreuses réactions métaboliques sont irréversibles, les réseaux métaboliques peuvent être des réseaux de type orienté. Les données disponibles et utilisées dans cette thèse ne permettant pas d'obtenir des réseaux de type orienté ; dans ce chapitre, nous nous limiterons à la description de l'inférence des réseaux de type non-orienté.

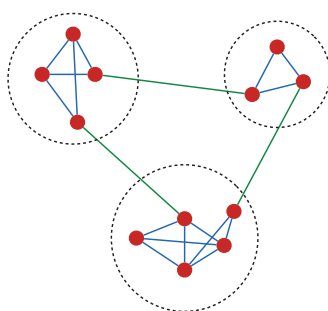


Figure 3.1 – Exemple d'un réseau composé de trois communautés (entourées par les cercles pointillés noirs) inspiré de Fortunato (2010).

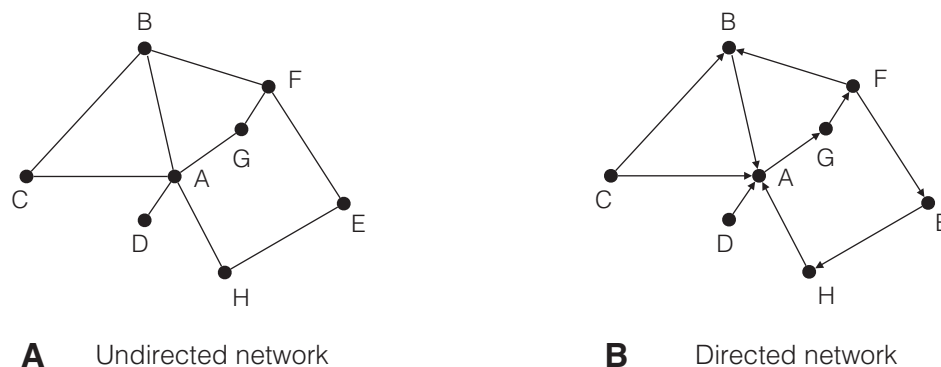


Figure 3.2 – Exemple de réseaux, (A) non-orienté et (B) orienté (inspiré de Barabási & Oltvai (2004)).

3.2.2 Méthodes d'inférence des réseaux biologiques

Le principe de la construction d'un réseau est d'inférer les interactions entre variables en utilisant directement les données biologiques (par exemple des données omiques d'expression de gènes), avec ou sans informations extrinsèques aux données (comme

des fonctions biologiques partagées entre groupes de variables). Par ailleurs, l'une des limites rencontrées lors de l'inférence de réseaux à partir de données de grande dimension, comme les données omiques, est le trop grand nombre de variables p par rapport au nombre d'individus n ($p \gg n$). Des difficultés peuvent alors survenir lors d'inversion de matrices (utilisées dans certaines méthodes d'inférences), d'où découlent notamment des problèmes d'estimation. Pour l'inférence de réseaux de type non-orienté, deux approches sont souvent utilisées : les relevance networks et les modèles graphiques gaussiens (ou *Gaussian Graphical Models* (GGM)).

- **Relevance network** (Butte *et al.*, 2000) : la dépendance entre deux variables est déterminée par une simple corrélation (comme la corrélation de Pearson). Si la corrélation entre les deux variables est supérieure à un seuil donné ou significative selon un test statistique donné, une arête sera inférée entre ces deux variables. Cette approche est détaillée dans la partie 3.2.2.1.
- **Modèle graphique gaussien** (Schäfer & Strimmer, 2005a,b; Meinshausen & Bühlmann, 2006; Friedman *et al.*, 2008; Peng *et al.*, 2009) : cette approche fait l'hypothèse de la distribution normale multivariée des données. La dépendance entre deux variables est déterminée grâce à l'utilisation de la corrélation partielle, où la corrélation entre deux variables est conditionnée à l'expression de toutes les autres variables. Plusieurs stratégies existent afin de déterminer la significativité (ou non) des arêtes entre deux variables. Cette approche est détaillée dans la partie 3.2.2.2.

La totalité des méthodes d'inférence de réseaux n'est pas exhaustivement présentée ici. Nous n'avons par exemple pas détaillé les réseaux de type bayésien (Friedman *et al.*, 2000; Scutari, 2010), exprimant les dépendances entre variables selon des *a priori*, ou encore les réseaux de type bipartite, exprimant des relations entre deux types différents de variables à partir de méthodes statistiques multivariées comme la sCCA (sparse Canonical Correspondence Analysis) ou la sPLS (sparse Partial Least Square) (González *et al.*, 2012).

3.2.2.1 Réseaux de type relevance network

Cette approche d'inférence de réseaux est considérée comme étant l'une des plus simples. Trois étapes sont présentes (Figure 3.3) :

- Détermination d'une matrice de similarité S . Un choix de mesure de dépendance entre variables est à effectuer, comme la corrélation de Pearson (Figure 3.3A) ;
- Détermination d'une matrice d'adjacence A . La matrice de similarité S est transformée en matrice d'adjacence A par l'utilisation d'un seuil ou d'un test statistique afin de conserver ou non les relations entre variables (Figure 3.3B) ;
- Construction du réseau à partir de la matrice d'adjacence A . Les variables sont connectées par une arête si l'élément correspondant de la matrice d'adjacence est non nul (Figure 3.3C).

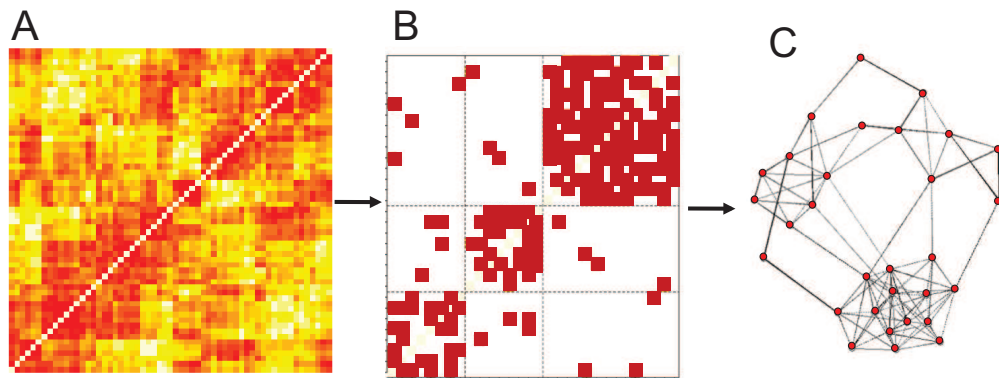


Figure 3.3 – **Etapes pour la construction d'un réseau de type relevance network.** (A) Calcul d'une matrice de similarité S . (B) Transformation de la matrice de similarité S en une matrice d'adjacence A . (C) Inférence du réseau à partir de la matrice d'adjacence A . Cette figure est inspirée du document *An introduction to network inference and mining* de Villa-Vialaneix (2013).

La corrélation de Pearson peut être utilisée afin de déterminer les liens de dépendance entre deux variables y_i et y_j :

$$\text{corr}(y_i, y_j) = \frac{\text{Cov}(y_i, y_j)}{\sqrt{\text{Var}(y_i)\text{Var}(y_j)}}$$

La matrice d'adjacence A est composée des valeurs 0 ou 1. Deux types de seuillage existent afin de déterminer cette matrice d'adjacence A : le seuillage dur (*hard-thresholding*) ou le seuillage doux (*soft-thresholding*). Pour transformer la matrice de similarité S en matrice d'adjacence A , un seuil dur est généralement utilisé. Le seuillage de type dur (Butte *et al.*, 2000) correspond au choix d'un seuil τ selon :

$$a_{ij} = \begin{cases} 0 & \text{si } s_{ij} < \tau \\ 1 & \text{si } s_{ij} \geq \tau \end{cases}$$

Le principal problème du seuillage de type dur est la perte d'information, due à l'obtention d'un réseau trop parcimonieux à cause d'un choix de seuil τ non adéquat. Le choix du seuil τ est assez arbitraire. Plusieurs auteurs ont ainsi proposé des méthodes de seuillage de type doux afin de déterminer les corrélations significatives ou non. Par exemple, la transformation Z de Fisher des données (Davidson *et al.*, 2001) et l'utilisation de tests de permutation (Carter *et al.*, 2004) ont été développées. Zhang & Horvath (2005) ont également proposé l'utilisation de règles de décision basées sur la loi de puissance de la matrice d'adjacence pour déterminer les arêtes significatives. Ils ont dénommé leur stratégie WGCNA (pour *Weighted Gene Co-expression Network Analysis*) et ont implémenté un package R **WGCNA** (Langfelder & Horvath, 2008). *In fine*, une arête est construite (ou non) entre deux variables i et j lorsque que la valeur A_{ij} est égale à 1 (ou à 0).

3.2.2.2 Réseaux de type modèle graphique gaussien

Bien que l'interprétation d'un réseau de type relevance network soit relativement simple, cette approche ne permet pas de distinguer les relations directes des relations indirectes entre variables. Par exemple, lorsque nous avons un triplet de variables (y_i , y_j et y_k), il est possible qu'une arête entre deux variables y_j et y_k soit présente à cause d'une trop forte corrélation de y_j et y_k avec une autre variable y_i (Figure 3.4). Ainsi, même s'il n'y a pas de lien biologique direct entre y_j et y_k , ces deux variables seront très corrélées entre elles (car peut-être régulées par y_i , l'exemple le plus simple, mais pas unique, étant la régulation de deux gènes par un même facteur de transcription) et une arête sera donc présente. Si l'on souhaite uniquement la représentation des liens directs, la corrélation partielle sera préférée lors de l'inférence de réseaux en biologie, notamment avec des données d'expression génique.

Introduit par Dempster (1972), les réseaux de type modèle graphique gaussien se basent sur la normalité multivariée des données. En premier lieu, définissons une matrice Y composée de p colonnes (correspondant au nombre de variables) et de n lignes (correspondant au nombre d'individus). Cette matrice d'expression Y est

supposée suivre une distribution normale : $Y \sim \mathcal{N}_p(\mu, \Sigma)$ avec $\mu = (\mu_1, \dots, \mu_p)$ étant un vecteur des moyennes des variables, et $\Sigma = (\sigma_{ij})$ la matrice de variance-covariance avec $i, j \in [1, p]$.

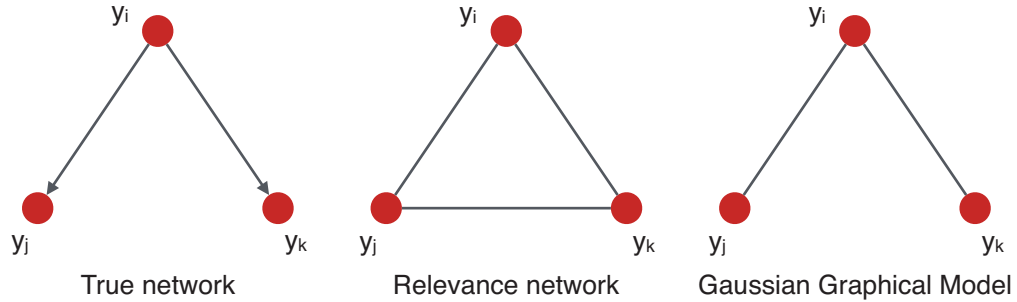


Figure 3.4 – **Illustration des différents types de dépendance entre variables.** Les modèles graphiques gaussiens permettent de distinguer les liens directs des liens indirects.

La corrélation partielle π_{ij} entre y_i et y_j est déterminée par :

$$\pi_{ij} = \text{corr}(y_i, y_j | y_{\setminus ij})$$

Ainsi, la corrélation partielle est la corrélation entre deux variables avec prise en compte de l'expression de toutes les autres variables. Elle permet donc de différencier les liens directs des liens indirects (Figure 3.4), ce qui peut être très intéressant lors de l'étude de réseaux de co-expression génique ou protéique. Afin de déterminer cette matrice de corrélation partielle, la matrice de variance-covariance doit être inversée en $\Sigma^{-1} = (\omega_{i,j})$ pour $i, j \in [1, p]$ afin de donner :

$$\pi_{ij} = \frac{-\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}}$$

Cependant, avec des données de grande dimension ($p \gg n$), la matrice de covariance n'est pas définie positive et sera donc impossible à inverser (Dykstra, 1970). Pour contourner ce problème, plusieurs méthodes ont été proposées afin d'estimer cette matrice Σ^{-1} .

Une première approche, développée par Schäfer & Strimmer (2005a), consiste en l'utilisation d'une méthode « d'estimateur à retrécisseur » (ou *shrinkage*) afin d'estimer la matrice de variance-covariance Σ :

$$\Sigma_{shrink} = \lambda T + (1 - \lambda)S$$

avec $\lambda \in [0, 1]$ le coefficient de *shrinkage*, S la matrice empirique de variance-covariance et T un estimateur sous-dimensionné. Il existe plusieurs façons de choisir T , par exemple la matrice diagonale constituée des variances empiriques. L'estimation peut aussi être combinée avec une approche de type bootstrap. Cette stratégie nous permet d'obtenir la matrice Σ_{shrink} , matrice inversible, et donc d'estimer une matrice de corrélation partielle dans laquelle nous devons, ensuite, déterminer les arêtes significatives ou non. Alors qu'un simple seuil, comme précédemment, peut être utilisé, Schäfer & Strimmer (2005a) ont proposé un test statistique, basé sur un modèle bayésien, où la distribution de la corrélation partielle observée est supposée suivre un modèle de mélange (ou *mixture*) :

$$n_0 f_0 + (1 - n_0) f_A$$

où n_0 est la proportion (non-connue) de «vraies arêtes» ($\pi \neq 0$), f_0 la distribution sous l'hypothèse nulle ($\pi = 0$) et f_A la distribution observée des corrélations partielles pour les vraies arêtes. n_0 est estimé par un algorithme espérance-maximisation (EM) (Schäfer & Strimmer, 2005a). Cette approche est implémentée dans le package R **GeneNet** (Schäfer & Strimmer, 2005a).

Une approche alternative est exposée par l'utilisation de la régression régularisée. La stratégie précédente proposait en premier lieu d'estimer la corrélation partielle, puis de sélectionner les arêtes significatives. Ici, l'estimation et la sélection sont effectuées simultanément en utilisant une pénalité sparse, connue sous le nom de glasso (pour *Graphical Least Absolute Shrinkage and Selection Operator*). Dans le cadre des propriétés gaussiennes, y_i peut être exprimé comme étant une combinaison linéaire des $y_{j \neq i}$:

$$y_i = \sum_{j \neq i} \beta_{i,j} y_j + \epsilon_i$$

Les corrélations partielles sont directement liées au coefficient de régression $\pi_{i,j} = \text{sign}(\beta_{i,j}) \sqrt{\beta_{i,j} \beta_{j,i}}$. Plusieurs auteurs (Meinshausen & Bühlmann, 2006; Friedman *et al.*, 2008; Peng *et al.*, 2009) ont proposé d'intégrer une pénalisation P sparse de type LASSO (L_1) dans l'estimation, afin de réduire certaines valeurs à 0 (et donc le nombre d'arêtes) :

$$P(\beta) = \alpha \|\beta\|_{L_1} = \alpha \sum_i |\beta_i|$$

où $\alpha > 0$ est le paramètre de régularisation contrôlant la parcimonie de β_i . Plus α est grand, plus le nombre de non-entrées dans β_i est grand. Cette valeur α peut varier au cours de l'estimation et dépend du nombre d'arêtes souhaité. Plusieurs stratégies existent pour le choix du paramètre α . Il est, par exemple, possible de sélectionner ce paramètre en fonction du nombre d'arêtes souhaitées afin d'obtenir un réseau de densité voulue et donc facilement analysable. Toutefois, les risques de sur-apprentissage ou de sous-apprentissage sont possibles, il faut donc être prudent. Il est également possible d'effectuer une validation croisée (comme souligné dans l'introduction, section 1.2.4) afin d'obtenir une meilleure estimation de la valeur α . Par ailleurs, Peng *et al.* (2009) ont aussi proposé l'utilisation du critère *BIC-type* (Bayesian Information Criterion) de par sa simplicité computationnelle et de calcul. Ces méthodes ont été implémentées dans les packages R **glasso** (Meinshausen & Bühlmann, 2006) ou **space** (Peng *et al.*, 2009).

3.2.3 L'algorithme PCIT (Partial Correlation with Information Theory)

J'ai choisi d'utiliser l'algorithme PCIT (pour *Partial Correlation with Information Theory*), présenté par Reverter & Chan (2008), pour inférer des réseaux de co-expression. Bien que cette méthode soit en premier lieu utilisée pour inférer des réseaux de co-expression génique (Hudson *et al.*, 2009; Pérez-Montarelo *et al.*, 2012), elle est utilisée ici afin d'obtenir des réseaux de co-expression protéique. Ainsi, un lien entre deux protéines indique que ces deux protéines ont un profil d'expression corrélé/similaire.

L'algorithme PCIT combine le concept de la corrélation partielle avec la théorie de l'information afin d'identifier les arêtes significatives lors de la construction de réseaux biologiques. Cet algorithme est composé de deux étapes distinctes :

- Calcul de la corrélation partielle de premier ordre pour chaque trio de variables ;
- Utilisation du théorème de l'information afin d'obtenir un niveau de tolérance utilisé pour la détermination des arêtes significatives.

Ainsi, en premier lieu, pour chaque trio de variables x , y et z , la corrélation partielle est estimée selon :

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}$$

où, comme expliqué en amont, la corrélation partielle ($r_{xy.z}$) entre x et y indique la corrélation linéaire entre x et y en prenant en compte l'expression de la troisième variable z . Il en est de même pour $r_{xz.y}$ et $r_{yz.x}$. Si les données ne sont pas normales (ce qui est souvent le cas lorsqu'il y a plusieurs conditions expérimentales) ou de distribution asymétrique (ce qui est fréquent avec les données protéomiques), des corrélations partielles égales à 0 ne sous-entendent pas nécessairement une indépendance, mais plutôt une dé-corrélation conditionnelle. L'indépendance conditionnelle est une notion clé dans la modélisation de réseaux, dans lesquels deux nœuds sont connectés par une arête si et seulement si ces variables ne sont pas conditionnellement indépendantes. Par exemple, dans le contexte des réseaux de régulation génique, Zampieri *et al.* (2008) ont démontré l'importance des corrélations conditionnelles et leur capacité à déterminer des interactions de régulation. Ils ont comparé différentes métriques de similarité et ont mis en avant que les informations concernant certains modules et autres régulations sont plus facilement capturées par l'utilisation de la corrélation partielle que par les mesures de similarité directe.

Ensuite, la théorie de l'information est utilisée afin de déterminer les arêtes significatives. Cette théorie est inspirée par le théorème de *Data Processing Inequality* (DPI) supprimant la majorité des interactions indirectes. D'après ce théorème, si deux variables v_1 et v_2 interagissent seulement via une troisième variable v_3 (si le réseau d'interaction est $v_1 - v_3 - v_2$ et qu'il n'existe pas de chemin alternatif possible entre v_1 et v_2), alors $r(v_1, v_2) \leq \min[r(v_1, v_3); r(v_2, v_3)]$ ($r(v_1, v_2) \leq r(v_1, v_3)$ et $r(v_1, v_2) \leq r(v_2, v_3)$). Ainsi, dans le but de déterminer un niveau de tolérance ϵ qui sera utilisé comme seuil local pour la détermination des arêtes significatives ou non, pour chaque trio de variables x , y et z , le ratio moyen entre la corrélation partielle (conditionnelle) et la corrélation directe est calculé selon :

$$\epsilon = \frac{1}{3} \left(\frac{r_{xy.z}}{r_{xy}} + \frac{r_{xz.y}}{r_{xz}} + \frac{r_{yz.x}}{r_{yz}} \right)$$

Ce seuil de tolérance est utilisé pour tenir compte des estimations inexactes des différences entre les deux valeurs de corrélations proches. Ainsi, une relation entre deux variables x et y est supprimée si :

$$|r_{xy}| \leq |\epsilon r_{xz}| \text{ et } |r_{xy}| \leq |\epsilon r_{yz}|.$$

Dans le cas contraire, la relation est conservée et une arête entre ces deux variables x et y sera établie lors de la construction du réseau de co-expression. En outre, afin de déterminer l'importance de l'association entre les variables x et y , les deux étapes mentionnées ci-dessus seront répétées pour chacune des $p - 2$ autres variables (correspondant à la variable z dans notre cas). Un package R **PCIT** a été implémenté (Watson-Haigh *et al.*, 2010). Cette méthode a été utilisée dans la construction des réseaux proposés dans la partie suivante 3.3.

3.2.4 Définitions et propriétés des réseaux biologiques

En amont de la description de notre stratégie d'intégration des données protéomiques, transcriptomiques et phénotypiques, nous allons définir les principaux termes qu'il est possible de rencontrer lors de l'utilisation de réseaux en biologie. Spécifions un graphe $G = (V, E)$ où :

- $V = \{v_1, \dots, v_p\}$ est un ensemble de nœuds (ou sommets) représentant les variables biologiques (gènes, protéines, etc.) ;
- E est un ensemble d'arêtes représentant les relations entre les différents nœuds.

Tout d'abord, un réseau est défini comme étant **connecté** (ou connexe) si, pour les nœuds v_i et v_j de G , il existe une chaîne de v_i vers v_j , c'est-à-dire une suite d'arêtes permettant d'atteindre v_j à partir de v_i , comme représenté dans la Figure 3.1.

La **densité** d est définie comme étant le nombre d'arêtes dans le réseau divisé par le nombre de paires de nœuds dans le réseaux : $d = \frac{|E|}{V(V-1)/2}$. Elle est utilisée afin de déterminer le taux de connexions (arêtes) présentes dans le réseau. Un réseau dit dense présente un nombre d'arêtes proche du nombre maximal, alors qu'un réseau dit parcimonieux possède au contraire très peu d'arêtes.

La **transitivité** d'un réseau est indiquée par le nombre de triangles (entre variables) présents dans le réseau divisé par le nombre de triplets de variables connectées par au moins deux arêtes. De façon simplifiée, dans un réseau social, la transitivité mesurerait la probabilité que deux de mes amis soient également amis. Lorsque la transitivité est plus grande que la densité, cela indique que les nœuds ne sont pas connectés entre eux par hasard.

L'une des caractéristiques élémentaires d'un nœud est son **degré** (ou connectivité). Celui-ci informe sur le nombre de relations que ce nœud possède avec les autres nœuds du réseau : $d_i = |\{(v_i, v_j) \in E\}|$. Les nœuds ayant les plus forts degrés sont généralement appelés *hubs*. La distribution des degrés, $\{P(k)\}_k$, correspond à l'ensemble des probabilités qu'un nœud ait exactement le nombre d'arêtes k . En général, de nombreux réseaux rencontrés en biologie suivent une *scale-free topology*, dans laquelle la distribution des degrés suit une loi de puissance $P(k) \propto k^{-\gamma}$ avec $\gamma > 0$, où des nœuds avec de faibles degrés (fréquents) coexistent avec des nœuds ayant de forts degrés (rares).

En parallèle au degré, la **centralité** (ou *betweenness* en anglais) d'un nœud est définie comme étant le nombre de plus courts chemins passant par ce nœud. Typiquement, la suppression d'un nœud avec une très forte centralité mène à une possible déconnexion du réseau, c'est-à-dire que ce nœud est important pour la structure du réseau.

La détection de **modules**, aussi appelés **clusters** ou **communautés**, est l'un des aspects importants lors de l'analyse de réseaux, c'est-à-dire la détection de groupes de nœuds (variables) étant fortement inter-connectés entre eux (Fortunato, 2010). La Figure 3.1 représente un exemple schématique de modules dans un graphe. Il est possible d'observer très facilement trois modules de nœuds dans le graphe. Les modules sont des groupes de variables pouvant partager des propriétés communes et/ou jouant des rôles biologiques similaires. La détection de modules dans un graphe possède des applications concrètes. Par exemple, dans les réseaux d'interaction protéine-protéine, les communautés regroupent souvent des protéines ayant les mêmes fonctions dans la cellule (Spirin & Mirny, 2003; Chen & Yuan, 2006), ou encore dans les réseaux métaboliques, les communautés peuvent être reliées à des voies métaboliques spécifiques (Guimerà & Amaral, 2005). Fortunato (2010) a récemment proposé une revue discutant de toutes les méthodes existantes afin de déterminer la structure en clusters d'un graphe : il existe des méthodes dites traditionnelles, comme le clustering hiérarchique ou les k-means, des méthodes d'algorithmes diviseurs où l'objectif est plutôt de supprimer les arêtes inter-clusters que les arêtes entre les paires de nœuds ayant une faible similarité, mais également des méthodes basées sur l'optimisation de la **modularité** Q . La

modularité est définie comme étant une mesure de la qualité d'un partitionnement des nœuds d'un graphe en communautés (Newman & Girvan, 2004). Elle comprend des valeurs entre -1 et 1, et elle augmente si la taille du graphe et/ou le nombre de communautés bien séparées augmente. La modularité n'est pas comparable entre graphes, mais est uniquement utilisée afin de déterminer et quantifier la structure en communautés d'un graphe, car elle est dépendante de la taille du graphe. Dans les différents articles proposés au cours de cette thèse, nous avons choisi de mettre en œuvre des algorithmes utilisant la modularité. Les algorithmes de type *greedy techniques* ont particulièrement été utilisés. Ils se basent sur des méthodes proches de la classification hiérarchique ascendante. Ils débutent par une partition triviale associant une classe à chaque sommet du graphe, puis les classes sont fusionnées de façon gloutonne (étape par étape afin d'obtenir une solution optimale globale au problème) en choisissant la meilleure fusion au sens d'un critère adapté (ici, la modularité). La procédure s'arrête quand plus aucune fusion n'est possible sans dégrader la modularité. Pour finir, ces algorithmes vont changer des sommets de classe de façon opportuniste en cherchant à augmenter la modularité. C'est ce type d'approche qui semblait donné les résultats statistiques les plus proches de la réalité biologique dans Villa-Vialaneix *et al.* (2013)

3.3 Article 2 : Voillet et al., *En préparation*, 2016

Cette section correspond à l'article suivant qui va être prochainement soumis :

- **V. Voillet**, M. San Cristobal, M.C. Père, Y. Billon, L. Canario, L. Liaubet and L. Lefaucheur. Integrated network analysis of proteomic and transcriptomic data highlights late fetal muscle maturation process. *In preparation* (2016).

Integrated Analysis of Proteomic and Transcriptomic Data Highlights Late Fetal Muscle Maturation Process

Valentin Voillet¹, Magali San Cristobal^{1,2,3}, Marie-Christine Père⁴, Yvon Billon⁵, Laurianne Canario¹,
Laurence Liaubet¹ and Louis Lefaucheur^{4,*}

¹ Université de Toulouse, INRA, INPT, INP-ENVT, UMR1388, GenPhySE, F-31326
Castanet-Tolosan, France

² INSA, Département de Génie Mathématiques, F-31077 Toulouse, France

³ Université de Toulouse, UMR5219, Institut de Mathématiques, F-31077 Toulouse,
France

⁴ INRA, UMR1348, PEGASE, F-35590 Saint-Gilles, France

⁵ INRA, UE1372, GenESI, F-17700 Surgères, France

* E-mail: louis.lefaucheur@rennes.inra.fr

Running title: Integration of omics data describes muscle maturity

Abbreviations

ACADVL: Acyl-CoA dehydrogenase very long chain, mitochondrial
AMPK: AMP-activated protein kinase
ANXA2: Annexin A2
ATP5A1: ATP synthase subunit alpha, mitochondrial
BP: Biological process
CC: Cellular component
CKM: Creatine kinase, muscle, cytoplasm
CKMT2: Creatine kinase, mitochondrial 2
DDAH1: Dimethylarginine dimethylaminohydrolase
ESR1: Estrogen receptor alpha
FDR: False discovery rate
FETUB: Fetuin-B, cystein protease inhibitor family
GO: Gene ontology
GPD1: Glycerol-3-phosphate dehydrogenase 1, cytoplasmic
GSN: Gelsolin
HPRT1: Hypoxanthine phosphoribosyltransferase 1
KCNJ11: Potassium Channel, Inwardly Rectifying Subfamily J, Member 11
LDB3: LIM domain binding 3
LM: Longissimus muscle
LW: Large White
MACROD1: MACRO domain-containing protein 1
MF: Molecular function
MS: Meishan
MyHC: Myosin heavy chain
OXCT1: Succinyl-CoA:3-ketoacid-coenzyme A transferase 1, mitochondrial
PCIT: Partial correlation information theory
PCK2: Phosphoenolpyruvate carboxykinase 2, mitochondrial
PDIA3: Protein disulfide-isomerase A3
PPARGC1A: Peroxisome proliferator-activated receptor gamma coactivator 1-alpha
PSMC5: Proteasome 26S regulatory subunit
RF: Random forest
sCCA: Sparse canonical correspondance analysis
SEPT2: Septin 2
SIRT1: Sirtuin 1
sPLS: Sparse partial least square
sPLS-DA: Sparse partial least square - discriminant analysis
TNNT3: Troponin T type 3

Summary

Background. In pigs, the perinatal period is the most critical time for survival. Piglet maturation, which occurs at the end of gestation, leads to a state of full development after birth. Maturity is thus an important determinant of early survival. Skeletal muscle plays a key role in adaptation to extra-uterine life, e.g. motor function and thermoregulation. Progeny from two breeds with extreme neonatal mortality rates were analyzed at 90 and 110 days of gestation. The Large White breed is a highly selected breed for lean growth with a high rate of mortality at birth, whereas the Chinese Meishan breed is a fatter and more robust and has a low mortality rate. The aim of our study was to identify important molecular signatures underlying late fetal muscle development.

Method. A strategy combining state-of-the-art statistical and computational methods was developed to integrate multi-omics datasets. First, integrated analysis was used to explore relationships between co-expression network models built from a proteomic dataset, and biological phenotypes of interest. Second, possible correlations with a transcriptomic dataset were investigated to combine different layers of expression with a focus on transcriptional regulation.

Results. Muscle glycogen content and myosin heavy chain polymorphisms were found to be good descriptors of muscle maturity and were used for further data integration analysis. Using 113 protein spots (89 identified unique proteins), network inference, correlation with biological phenotypes and functional enrichment revealed that mitochondrial oxidative metabolism were a key determinant of neonatal muscle maturity. Some proteins, including ATP5A1 and CKMT2, were identified as important nodes in the network related to muscle metabolism. GPD1, an enzyme involved in the mitochondrial oxidation of cytosolic NADH, was also over-expressed in the Meishan breed. Thirty-one proteins exhibited a positive correlation between their mRNA and protein levels, suggesting transcriptional regulation in both extreme fetal genotypes. Gene ontology enrichment analysis and Ingenuity analysis identified *PPARGC1A* and *ESR1* as possible transcriptional factors positively involved in late fetal metabolic maturation of skeletal muscle.

Introduction

One objective of systems biology is to investigate the regulation and interaction of various components of the cell including DNA (genomics) [1], mRNA (transcriptomics) [2], proteins (proteomics) [3] or metabolites (metabolomics) [4]. Even though transcriptomic analysis provides deep insights into cellular processes, possible conclusions are limited [5]. Indeed, mRNA expression is not always a good predictor of protein level because of low correlations between mRNA and protein expression levels are often observed. For example, a relatively small change in mRNA expression can result in a major change in the total abundance of the corresponding protein, enabling potentially different conclusions to be drawn from transcriptomic and proteomic analyses.

Biological integration of transcriptome and proteome remains challenging in omics studies due to the marked differences between the two approaches, e.g. the dynamic range of regulation, incomplete annotation or isoform differences [6]. The expression of genes may not be correlated with the abundance of proteins and provide no information on post-transcriptional events, e.g. translational efficiency, alternative splicing, folding or assembly into complexes [5, 7, 8]. An integrated multi-omics approach, with gene expression experiments and large-scale protein identification experiments, should provide a deeper understanding of the functional interactions between mRNA and protein layers as a complex biological system. Several biological data integration strategies have already been suggested, see review by Ritchie et al. (2015) [9]. In this review, the authors suggested three meta-dimensional analyses combining multiple data types in the same analysis: concatenation, transformation and model-based integration. Here we chose to develop an innovative integration strategy combining state-of-the-art statistical and computational methods using proteomic and phenotypic data, with the incorporation of transcriptomic information, to observe and identify some important proteins with possible transcriptional regulation. One of the advantages of our strategy is to avoid the scale issues that may arise when different types of data are combined.

Networks are increasingly used as tools for analysis and for the visualization of data in biology and genetics [10–13]. A complex network architecture into clusters of functionally related genes/proteins can be explored and genes/proteins with high connectivity (called hubs) can be identified. In our study, integrated network analysis was first performed to explore relationships between co-expression network models, built from a proteomic dataset, and phenotypes of interest that would enable identification of important molecular signatures underlying late fetal muscle development. Correlations with a transcriptomic dataset were then investigated to complete and combine different layers of expression.

Here we report the results of multi-omics analyses of muscle during the last three weeks of gestation in pigs. The objective was to identify proteins, with possible transcriptional regulation, and related biological mechanisms, specially those involved in differences in the muscle maturation process in late gestation between two extreme pig breeds: Large White and Meishan, and reciprocal crosses. The Large White (LW) breed is a highly selected breed for lean growth with a high rate of mortality at birth, whereas the Chinese Meishan (MS) is a fatter and more robust breed that produces piglets with an extremely low mortality rate [14, 15]. Physiological muscle maturity, which occurs at the end of gestation, has already been shown to improve early survival after birth [16–18]. Successful maturation in late gestation thus likely leads to a state of full development and promotes early survival after birth. Postnatal mortality due to immaturity is not only an issue in pigs but affects other mammals including sheep [19] and humans [20, 21]. Adaptation to extra-uterine life is therefore a major factor in the survival of all mammal species. In the present study, using network data integration analysis, we identified key proteins involved in piglet maturity during late gestation and provided an overview of the muscle maturation process in late gestation of which many aspects can be generalized to other mammals.

Materials and Methods

Source data

Ethics statement

Use of animals and procedures performed in this study was approved by the European Union legislation (directive 86/609/EEC) and French legislation in the Midi-Pyrénées Region of France (Decree 2001-464 29/05/01; accreditation for animal housing C-35-275-32). The technical and scientific staff obtained individual accreditation (MP/01/01/01/11) from the ethics committee (region Midi-Pyrénées, France) to experiment on living animals. All the fetuses used in this study were males and were obtained by caesarean.

Study design

Details regarding animal resources and experimental designs can be found in [18]. Briefly, *longissimus skeletal muscle* mRNA, protein and phenotypic data were acquired at two developmental time points (90 and 110 days of gestation (dg)) from four fetal genotypes. These genotypes consisted in two extreme breeds concerning mortality at birth (Meishan (MS) and Large White (LW)) and two crosses (MSLW from LW sows and LWMS from MS sows). MS and LW sows were inseminated with mixed semen so that each litter was composed of purebred and crossbred fetuses. The two developmental time points correspond to the end of gestation when intense maturation of muscle fibers occurs from 90 dg to birth (around 114 dg) [16, 17, 22]. Therefore, eight conditions were considered according to the two development time points and the four fetal genotypes ($n = 64$ fetuses exhibiting birth weight close to the average birth weight within litter and genotype, $n_{d90.MS} = 8$, $n_{d90.LW} = 8$, $n_{d90.MSLW} = 8$, $n_{d90.LWMS} = 8$, $n_{d110.MS} = 8$, $n_{d110.LW} = 10$, $n_{d110.MSLW} = 6$ and $n_{d110.LWMS} = 8$).

Muscle sampling and biochemical analysis

The *longissimus muscle* (LM) was collected at the last rib level within 30 min after death, cut into small pieces, snap frozen in liquid nitrogen, and stored at -75°C until further analyses. Glycogen content was determined in LM according to the method described by Good et al. [23] with minor adaptations [24], and is expressed in g/100 g tissue wet weight. Myosin heavy chain (MyHC) polymorphism was characterized both at the mRNA level by quantitative real-time PCR amplification using the TaqMan technology and at the protein level using one dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis (1D SDS-PAGE) as previously described in [25]. At the mRNA level, the expression of each MyHC ($MYH7 =$ type I MyHC, $MYH2 =$ type IIa MyHC, $MYH1 =$ type IIx MyHC, $MYH4 =$ type IIb MyHC, $MYH3 =$ embryonic MyHC, $MYH8 =$ perinatal MyHC, $MYH6 = \alpha$ -cardiac MyHC and $MYH13 =$ extraocular MyHC) was calculated based on the PCR efficiencies and a calibrator, and expressed in comparison to an invariant endogenous reference gene (hypoxanthine phosphoribosyltransferase 1, *HPRT1*) as described by Pfaffl [26]. *HPRT1* expression was not affected by the fixed factors used in the statistical analysis. Types I, IIa, IIx and IIb MyHC are the four MyHC that define muscle fiber types in adult skeletal muscle [27], whereas embryonic, perinatal, and α -cardiac MyHC are expressed transiently during the fetal and early postnatal periods in pigs [28, 29]. At the protein level, four bands were separated by 1D SDS-PAGE and corresponded to the embryonic, perinatal, fast-type II (IIa + IIx + IIb) and slow-type (I + α -cardiac) MyHC, respectively [30]. Each band is expressed as a percentage of all four bands within a lane.

Proteome analysis

Muscle total protein extraction and bi-dimensional electrophoresis were performed on the 64 muscle samples as previously described in [31]. Briefly, for the first dimension, 300 μg protein were loaded onto

immobilized pH gradient strips (pH 3-11 NL, GE Healthcare, Uppsala, Sweden) and isoelectric focusing (IEF) was performed using an Ettan IPGphorII system (GE Healthcare) at 20°C up to a total of 88,600 Vh. At the completion of IEF, equilibrated strips were transferred onto the top of a 12.5% uniform SDS-PAGE gel using a vertical Ettan DALTsix system (GE Healthcare). After migration, gels were stained with Coomassie Brilliant Blue G-250 (Bio-Rad). The gels were scanned using an UMAX ImageScanner (GE Healthcare) and spot detection and quantification were performed by image analysis (Melanie 2D gene analysis software V7.0; Swiss Institute of Bioinformatics, Lausanne, Switzerland). Artefacts and saturated spots were removed from the analysis. Spots were matched across all 64 samples and 1,025 valid spots were successfully matched between gels. For each gel, each spot volume was expressed as a percentage of the volume of all matched spots on a given gel. Data were further log₂ transformed to get approximately normally distributed data before detection of differentially expressed and discriminating spots. A representative 2D electrophoresis gel is included in Additional file 1.

Protein identification is a really *a posteriori* time-consuming work. Statistics were done in a complete blind manner and about 200 spots were expected for mass spectrometry identification. So that different statistical methods were chosen to obtain a wide spectrum of protein expression including differential analyse for depicting biological processes and discriminant analyse to identify possible markers. These methods included analyses of variance to detect spots significantly affected by gestational time points and fetal genotypes, in an additive or non-additive manner, as described in [18]. Secondly, random forest (RF) [32] and sparse partial least square discriminant analysis (sPLS-DA) [33] were performed to find supplementary spots with a predictive power for gestational stages and/or fetal genotypes. Several RFs analyses were conducted. We computed a classification RF according to the experimental design. And, we also performed some regression RFs using phenotypes of interest (such as embryonic or adult fast myosins). In each case, we selected spots with a high stability. To do that, we performed 20 RFs and kept the first twenty spots according to the importance criteria. In sPLS-DA, we chose to keep 25 spots by axis. The first axis discriminated the fetuses according to the gestational age, whereas the second one discriminated fetuses according to the fetal genotype. Finally, several statistical multivariate methods (sparse partial least square (sPLS) [34] and sparse canonical correlation analysis (sCCA) [35]) were also performed to find additional spots correlated with phenotypic biological characters of interest such as MyHC profiles and muscle glycogen content. In sPLS analysis, 20 and 30 spots were kept according to axis 1 and 2, respectively. Like in sPLS-DA, the first axis discriminated the fetuses according to the gestational age and the second one according to the fetal genotype. In total, 179 spots were selected and manually excised from preparative gels loaded with 600 µg of proteins from pooled samples for further identification by nano-LC-MS/MS, as described in [36].

In-gel tryptic digestion and protein identification by mass spectrometry were performed at the proteomic facilities in Clermont-Ferrand (PFEMcp, INRA, Clermont-Ferrand Theix, France). Tryptic peptides were analyzed by nano LC-MS/MS using nano-LC system Ultimate 3000™ RSLC (Dionex, Voisins le Bretonneux, France) coupled on-line to an LTQ VELOS mass spectrometer (ThermoFisher Scientific, Courtaboeuf, France) operated in a CID top 5 mode (*i.e.* one full scan MS and the five major peaks in the full scan were selected for MS/MS). For database searches and protein identification, Thermo Proteome Discoverer 1.4 software was used with Mascot (Mascot server v2.2, <http://www.matrixscience.com>) to submit MS/MS data to the SwissProt sequence database restricted to *Sus scrofa* (UniProt Sus scrofa, 26127 sequences). The following parameters were chosen for the searches: the mass tolerance for parent and fragment ions was set to 1.5 Da and 0.5 Da, respectively, and a maximum of two missed cleavages was allowed. Variable modifications were methionine oxidation (M) and carbamidomethylation (C) of cysteine. Results were scored using the probability-based Mowse algorithm, where the protein score is $-10 \cdot \log(P)$ and P is the probability that the observed match is a random event. Protein identifications were considered valid if at least two peptides with a statistically significant Mascot score > 36 were assigned, and at least

20% sequence coverage was required. The accuracy of the experimental to theoretical isoelectric point and molecular weight were also considered. Among the 179 selected spots, 120 spots, corresponding to 89 unique proteins, were successfully identified (Additional Files 2 and 3).

RNA preparation and gene expression data set

Total RNA was isolated from each of the 64 muscle samples as previously described in Voillet et al. [18]. After quality control and quantile normalization steps, 61 microarrays containing 44,368 probes were kept. The raw and normalized data are available in the Gene Expression Omnibus under the accession number GSE56301.

Integration of proteome network data

Network analysis was performed using a three step approach as illustrated in Figure 1 and at the two developmental time points separately (90 and 110 dg (32 individuals per condition)), resulting in two global networks. We chose to infer two networks to avoid the high gestational age effect already observed in a previous study analyzing a transcriptomic dataset with the same individuals [18]. All analyses were performed using the R computing environment [37].

Figure 1 about here.

Inference of the proteome co-expression network

The first step consisted in inferring a proteome co-expression network using the PCIT (partial correlation and information theory) algorithm [38] (Figure 1A). PCIT has already been successfully used in transcriptomic analyses [39, 40]. In the present study, we performed PCIT on a proteomic data set. PCIT belongs to the family of weighted network algorithms and is based on the combination of the concept of partial correlation coefficient and the information theory to identify meaningful associations. Briefly, the co-expression arrangements for all triplets are compared, with all triplets being exhaustively explored. PCIT estimates the correlation for each pair of proteins taking the presence of a third protein into account. Significant correlations establish an edge in the reconstruction of the network. Even though PCIT is a soft-thresholding method, the number of selected edges was also adjusted according to the network density (around 5% for each network) to obtain readable networks. This first step of the analysis was performed with the R package PCIT [41].

Proteome network clustering

The second step consisted in finding communities within networks (Figure 1B). For each network, a spin-glass model was performed to optimize the modularity Q [42] and to cluster nodes [43]. The modularity Q is a measure of the quality of communities (or clusters) in a network: highly connected genes within each community and weakly connected genes between communities [42]. As already described in [44], a permutation test (100 permutations of edges corresponding to 100 random networks with the same degree distribution) was used to declare whether or not the clustering was significant.

Proteome network community analysis and relationship with a phenotype of interest

The third step consisted in the biological analysis of each community (or sub-network). GeneMANIA networks [45] were used to explore and confirm the relevance of the proposed communities. Gene ontology (GO) enrichment analysis using the GeneCodis webservice [46] was used to identify the biological functions represented by each community (Figure 1C). For the significance of GO enrichment, multiple testing was controlled using the false discovery rate (FDR) approach [47].

The third step included highlighting some nodes in each community. As already reported in other studies [18,44], the betweenness of nodes within each community was analyzed (Figure 1C). A permutation test with 1,000 permutations was performed to check if the betweenness centrality was significant with respect to the node's degree. A significant result ($FDR < 0.05$) indicated a node was more central in the network than expected. Therefore, betweenness centrality is a good measure of the importance of the node in the network. A node with a high betweenness (or centrality) value has a marked influence on the structure of the network. All these analyses were performed using the R package *igraph* [48]. If no node with significant betweenness was found, the nodes with the highest betweenness were highlighted.

In an integrative strategy, phenotypic information was also added to the co-expression network. Links between sub-networks and biological phenotypes of interest were investigated (Figure 1C). To this end, methods coming from spatial statistics were used as described in [13,49]. First, the correlation between each protein expression community and the phenotype of interest was calculated. Second, a Moran's I was calculated to measure the spatial correlation between the sub-network structure and the phenotype of interest. Then, to check if the correlation between the biological phenotype and the sub-network was significant ($p < 0.05$), a permutation test with 1,000 node permutations was computed.

Network layout

All the graphs were laid out using the Force Atlas layout [50]. The degree (the number of adjacent edges) is indicated by the size of the node, and betweenness is indicated by the color of the node. The colors of the edges show whether the correlation between nodes is positive (in red) or negative (in blue). All these analyses were performed using Gephi software [51].

Gene-protein integration

Among the 89 unique proteins identified, 81 corresponding genes were available in the muscle transcriptome dataset [18]. When multiple probes mapped to the same gene, the highest differentially expressed probe according to the interaction between gestational time points and fetal genotypes was retained. Transcriptomic and proteomic analyses were carried out using the same 64 muscle samples, but 60 fetuses were available in both proteomic and transcriptomic datasets.

To identify proteins with possible transcriptional regulation, for each protein within each fetal genotype, a Pearson's correlation was computed between mRNA (from the gene expression dataset [18]) and protein expression levels. A test was also run to assess if the correlation was significantly non-null ($p < 0.01$). Multiple testing was controlled using the FDR approach [47].

From a biological point of view, the Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, www.qiagen.com/ingenuity) software was used to check enrichment analysis (biological functions and canonical pathways), to construct bibliographic networks and regulation networks based on the identification of potential upstream regulators. A focus was on sub-network 2 at 110 dg, because of its relevant spatial correlation with two biological phenotypes of interest and a high number of proteins with a possible transcriptional regulation. Briefly, IPA constructed two separated networks. The first was based on bibliographic data in which the edges were obtained from biological links such as receptor-ligand interactions, enzyme activity on another protein, or a transcriptional factor activating the expression of targeted genes. IPA proposed the most probable network with an associated score. IPA also identified upstream regulators with a statistical likelihood of targeting some of the genes or proteins in the network. Finally, IPA generated a regulated network with the latest information. The proposed network (described in the Results section) was reconstructed from both IPA networks by integrating direct and regulated

links. The PathDesigner function of IPA was used to draw a final graph with all previous information plus the information related to the correlation between transcripts and proteins when the information was available from the transcriptomic data.

Results

Analysis of MyHC polymorphism and muscle glycogen content

MyHC mRNA and protein levels and muscle glycogen content are listed in Table 1. All these biological phenotypes were differentially expressed according to the interaction between the two developmental time points (90 and 110 dg) and the four fetal genotypes (MS, LW, MSLW and LWMS), except slow (I + α -cardiac) MyHC at the protein level. The expression of IIB and extra-ocular MyHC were undetectable at 90 and 110 dg. In all four genotypes, embryonic and perinatal MyHC (mRNA and protein levels) decreased at the end of gestation, whereas fast (IIa + IIx) and slow (I + α -cardiac) MyHC increased during the maturation process. Embryonic MyHC was more highly expressed in LW than in MS fetuses at 90 dg, whereas a higher value of fast (IIa and IIx) MyHC was observed in MS than in LW at 110 dg. A high correlation between mRNA and protein levels was found for both embryonic and adult fast MyHC (Pearson's correlation between mRNA and protein levels equal to 0.95 and 0.93, respectively) (Figure 2).

Table 1 about here.

The profile of muscle glycogen content resembled that of adult fast MyHC between 90 and 110 dg with a 2.6 fold increase during the maturation process (Table 1). At 90 dg, MS and LWMS fetuses already exhibited higher muscle glycogen content than LW and MSLW fetuses. At 110 dg, levels reached 12.2% in MS and LWMS vs 10.5% and 9.5% in MSLW and LW, respectively ($p < 0.001$). Figure 2 shows the corresponding changes in muscle glycogen content and both embryonic and adult fast MyHC. Muscle glycogen content was particularly well correlated with embryonic MyHC at 90 dg and with adult fast (IIa + IIx) MyHC at 110 dg.

Figure 2 about here.

Because MS piglets are known to be more mature than LW piglets at birth [15], embryonic MyHC, fast (IIa and IIx) MyHC and muscle glycogen content are likely good descriptors of muscle physiological maturity in pig neonates. These biological phenotypes were consequently used in the proteomic and transcriptomic analyses to help identify proteins and genes potentially involved in the muscle maturation process in late fetal stages.

Proteomic analysis

Multiple statistical analyses to select potentially relevant protein spots

As described in Materials and Methods, several statistical methods were conducted to obtain a large spectrum of proteins related to the experimental design. One hundred seventy nine spots were selected among the 1,025 spots matched between gels. After protein identification, 120 spots, corresponding to 89 unique proteins, were successfully identified by LC-MS/MS and 18 proteins exhibited different isoforms. To facilitate data interpretation and computation, protein isoforms were filtered to exclude redundant isoforms (i.e. highly positively correlated isoforms). When multiple highly positively correlated isoforms (Pearson's correlation > 0.9) mapped to the same protein, only the most differentially expressed isoform for the interaction between developmental time points and fetal genotypes was retained. Isoforms without a high positive correlation were kept (Pearson's correlation < 0.9). Finally, a total of 113 proteins were retained and used for further analyses (Additional File 3).

The distribution of the 113 identified spots among the different statistical analyses is shown as a Venn diagram in Figure 3. Seventy-one spots were significantly differentially expressed of which 13 were affected by the interaction and 58 by the additive model between gestational time points and fetal genotypes. Using RF and sPLS-DA analyses to obtain spots that distinguished gestational stages and/or fetal genotypes, 55 spots were selected. Then, 62 spots were also selected using several multivariate methods to obtain spots correlated with the biological phenotypes of interest (MyHC polymorphism and glycogen content). Altogether, a large number of spots overlapped between the multivariate and discriminant analyses, so that 179 spots were finally selected. After spot identification by mass spectrometry and some isoform removing, 113 proteins were identified.

Figure 3 about here.

The classification of the 113 identified proteins according to the biological axes is presented as a level plot in Figure 4. Proteins involved in energy metabolism processes, such as glucose metabolism, gluconeogenesis, oxidation-reduction activity and mitochondrion, were mostly over-expressed at 110 dg in all genotypes. The figure also shows that oxidation-reduction related to mitochondria at 110 dg increased in the order LW < MSLW < LWMS < MS. On the other hand, proteins involved in muscle development, such as system development, actin skeleton, muscle filament sliding, cytoskeleton and mRNA metabolic process, were mostly over-expressed at 90 dg in all genotypes. All enriched biological functions and cellular components are not shown in this figure, as we chose to highlight only important GO terms with a high number of proteins. For that reason, some proteins do not belong to any of the biological processes shown.

Figure 4 about here.

Proteome sub-network analysis

Network inference was performed at each developmental time point according to the three-step inference method presented in Figure 1. The largest connected component of the d90 and d110-proteome networks are presented in Additional File 4 and characteristics of these networks (degree, betweenness and clustering) are summarized in Additional Files 5 and 6, respectively.

d90-proteome network analysis

The largest connected component of the d90-proteome network was composed of 94 nodes and 314 edges (density of 7.2%) and followed a scale-free topology denoting a non-random organization of the network [10]. After node clustering, five sub-networks were obtained (Table 2 and Figure 5A). All five sub-networks displayed between 16-21 nodes and 26-55 edges. Additional File 7 displays all five sub-networks, and Additional File 8 shows the enriched Gene Ontology (GO) terms for each sub-network. In several sub-networks, some isoforms for the same protein were linked because only one developmental time point was used for the network inference. Figure 5A shows that sub-networks 1, 2 and 4 had a high number of edges in common, and shared a GO term corresponding to muscle filaments. Notably, these three sub-networks were also those that were significantly and spatially correlated with the three biological phenotypes of interest (Table 2). Figure 6 shows the three sub-networks (sub-networks 1, 2 and 4) were significantly and spatially correlated with the biological phenotypes of interest. Sub-network 1 was correlated with glycogen and embryonic MyHC, sub-network 2 to glycogen and adult fast (IIa + IIb + IIx) MyHC, whereas sub-network 4 was only correlated with embryonic MyHC.

Table 2 about here.

Figure 5 about here.

Figure 6 about here.

According to GO functional enrichment analysis, each sub-network was related to several biological functions (Table 2): sub-network 1 was mainly involved in the myofibril (GO cellular component (CC)), glycolysis (GO biological process (BP)) and mitochondrion (CC), sub-network 2 in actin filament (CC), gluconeogenesis (BP) and mitochondrion (CC), and sub-network 4 in muscle filament sliding (BP), gluconeogenesis (BP) and cell cycle checkpoint (BP). Sub-network 3 was mainly related to the respiratory electron transport chain (BP), cell redox homeostasis (BP) and mitochondrion (CC), whereas sub-network 5 was related to the creatine metabolic process (BP) and sarcolemma (CC). It is important to note that some identical GO terms were enriched in several sub-networks because a large number of the proteins identified exhibited different isoforms and these isoforms could be present in different sub-networks. In addition, the proteins we identified could also be involved in several metabolic functions.

To go deeper into the biological interpretation of the clusters, we chose to highlight the three sub-networks correlated with the biological phenotypes of interest (Figure 6 and Table 2). Sub-network 1 showed a significant correlation with muscle glycogen content and embryonic MyHC. The enriched GO biological processes, such as glycolysis, glucose metabolic process and gluconeogenesis, were in agreement with the glycogen correlation. FETUB (Fetuin-B, a member of the cysteine protease inhibitor family - involved in the negative regulation of endopeptidase activity) and OXCT1 (Succinyl-CoA:3-ketoacid-coenzyme A transferase 1, mitochondrial - involved in ketone body catabolic process) exhibited the highest degree of this sub-network. In addition, FETUB exhibited the highest betweenness and was more highly expressed in LW than in MS fetuses. Muscle filament sliding was also one of the enriched GO terms describing this sub-network. In sub-network 2, one isoform (isoform 2) of PSMC5 (Proteasome 26S Regulatory Subunit, ATP-dependent degradation of ubiquitinated proteins, 5) had significantly high betweenness and the highest degree of this sub-network. This isoform of PSMC5 was more highly expressed in MS than in LW fetuses (Figure 7). Three isoforms (isoforms 1, 2 and 4) of GPD1 (Glycerol-3-phosphate dehydrogenase 1, cytoplasmic, involved in the glycerol phosphate shuttle) were also identified in this sub-network. Interestingly, these three isoforms were more highly expressed in MS than in LW fetuses, whereas GPD1 isoform 3 was less expressed in MS than in other genotypes (Figure 7). Muscle filament sliding was also identified as a significant GO term to characterize sub-network 2. Sub-network 4 was correlated with the embryonic MyHC. One isoform (isoform 2) of TNNT3 (Troponin T type 3) had significantly high betweenness and the highest degree of this sub-network. The enriched GO biological processes, e.g. muscle filament sliding and development and cell cycle checkpoint, were in agreement with the values of the phenotypic correlation. Two other TNNT3 isoforms (isoform 1 and 3) were also present in this sub-network. It is noteworthy that a common GO term applied to sub-clusters 1, 2 and 4 related to muscle filaments, which could denote a strong involvement of muscle filaments in the maturational process at 90 dg, as previously highlighted on the level plot in Figure 4.

Figure 7 about here.

d110-proteome network analysis

The largest connected component of the d110-proteome network was composed of 86 nodes and 313 edges (density of 8.6%) and had scale-free topology denoting a non-random organization as observed at 90 dg. We obtained six clusters (Table 3 and Figure 5B). Additional File 9 shows all six sub-networks and Additional File 10 shows the enriched GO terms for each sub-network. Among these six sub-networks, five sub-networks displayed more than 10 nodes and 19 edges. Four sub-networks were spatially correlated with muscle glycogen content (sub-networks 3, 4, 5 and 6), whereas three sub-networks were spatially correlated with the embryonic MyHC (sub-networks 2, 3 and 4) or the adult fast MyHC (sub-networks 2, 3 and 5).

Table 3 about here.

GO term enrichment analysis was performed for each sub-network (Table 3): (i) sub-network 1 (only five nodes) was involved in the tricarboxylic acid cycle (BP) and mitochondrion (CC), (ii) sub-network 2 consisted in 24 nodes and 65 edges and was primarily involved in the tricarboxylic acid cycle (BP), respiratory electron transport chain (BP), glucose metabolic process (BP) and muscle filament sliding (BP), (iii) sub-network 3 was involved in gluconeogenesis (BP) and glycolysis (BP), (iv) sub-network 4 in the mRNA metabolic process (BP) and in the proteasome complex (CC), (v) sub-network 5 in muscle development (BP) and also lipid metabolism (BP), and (vi) sub-network 6 was mainly involved in gluconeogenesis (BP), ATP catabolic/anabolic process activity (BP) and mitochondrion (CC).

As previously, we chose to highlight the five sub-networks correlated with a biological phenotype of interest (Figure 8). Sub-network 2 was correlated with both adult fast and embryonic MyHC and had the highest number of nodes (24) and edges (65). This sub-network was primarily characterized by GO terms corresponding to mitochondrial oxidative metabolism (Additional File 10). ATP5A1 (ATP synthase subunit alpha, mitochondrial) showed significant high betweenness and CKMT2 (Creatine kinase, mitochondrial 2) exhibited the highest degree of the sub-network. Both CKMT2 and ATP5A1 were more highly expressed in MS than in LW fetuses (Figure 7). Sub-network 3 was significantly correlated with all three phenotypes of interest and was mainly involved in glucose metabolic process, gluconeogenesis and glycolysis occurring in the cytoplasmic compartment. GSN (Gelsolin - involved in actin filament reorganization) was the protein with the highest betweenness, but also the highest degree of this sub-network with PDIA3 also known as GRP58 (protein disulfide-isomerase A3 - involved in protein folding). Sub-network 4 was correlated with the embryonic MyHC and muscle glycogen content and mostly involved in the proteasome complex and the mRNA metabolic process. An isoform (isoform 1) of PSMC5 had the highest betweenness. This protein and DDAH1 (Dimethylarginine dimethylaminohydrolase - involved in arginine metabolic process) had the highest degree of this sub-network. Muscle glycogen content and adult fast MyHC were the two biological phenotypes correlated with sub-network 5. PGAM1 (Phosphoglycerate Mutase 1 - involved in the glycolytic process) was the protein with the highest degree, whereas one isoform (isoform 2) of LDB3 (LIM domain binding 3 - involved in sarcomere organization) had the highest betweenness. Sub-network 6 was significantly correlated with muscle glycogen content, which is consistent with the enriched GO terms (*e.g.* gluconeogenesis and ATP catabolic/anabolic process). An isoform (isoform 2) of PSMC5 had the highest betweenness and the highest degree of this sub-network. Thus, PSMC5 isoform 2 had a major influence on the structure of sub-network 6 at 110 dg and sub-network 2 at 90 dg, whereas PSMC5 isoform 1 played a key role in sub-network 4 at 110 dg. Two isoforms (isoform 2 and 4) of GPD1 were also present in sub-network 6 and under-expressed in LW fetuses (Figure 7).

Figure 8 about here.

Identification of mRNAs correlated with proteome networks

For each protein (with a corresponding gene in the transcriptomic dataset) within each extreme fetal genotype, a Pearson's correlation was computed between mRNA and protein expression levels in order to identify proteins with possible transcriptional regulation. The transcriptomic dataset was previously used in [18] and 60 fetuses were represented in both the transcriptomic and proteomic datasets. Figure 9 shows the correlation between transcriptome and proteome in MS (x axis) and LW (y axis) genotypes. Among the 81 proteins available in the transcriptomic dataset, 31 proteins exhibited a significant positive correlation between mRNA and protein expression levels in both extreme fetal genotypes, 8 proteins in LW only (mostly involved in muscle filament sliding and located in cytosol) and 13 proteins in MS only (with no specific biological enrichment) (Figure 9 and Additional File 11). On the other hand, one protein and its associated isoform (TNNT3 isoform 6) was also found to have a significant negative

correlation between mRNA and protein expression levels in both extreme genotypes, two proteins in LW only (TNNT3 isoform 1 and LDB3 isoform 1) and two proteins (TNNT3 isoform 7 and GSN isoform 1) in MS only (Additional File 11).

Figure 9 about here.

For a more detailed biological interpretation, we chose to highlight the 31 proteins with a significant positive correlation between mRNA and protein expression in both MS and LW pure fetuses. Sub-network 2 at 110 dg, already highlighted because of its relevant spatial correlation with two biological phenotypes of interest (adult fast and embryonic MyHC), was particularly interesting. Thanks to the correlation analysis, we showed that 10 nodes (seven belonging to the mitochondrion cellular component) among the 24 nodes that comprised this sub-network exhibited a positive correlation between mRNA and protein expression and were likely transcriptionally regulated (Additional File 11), especially CKMT2 and ATP5A1. An IPA analysis of this sub-network (using all nodes) identified possible transcription factors (TFs) affecting these proteins, in particular PPARGC1A (peroxisome proliferator-activated receptor gamma coactivator 1-alpha) with an effect on CKM (creatine kinase, muscle, cytoplasm), ATP5A1, CKMT2 and ACADVL (very long-chain specific acyl-CoA dehydrogenase, mitochondrial - involved in fatty acid beta-oxidation) (Figure 10). The PPARGC1A gene (available in the transcriptomic dataset) was over-expressed at 110 dg compared to 90 dg. It is interesting to note that at 90 dg, the expression of PPARGC1A was lower in LW than in MS. This difference between genotypes, which was only visible at 90 dg, suggests that the increase in PPARGC1A expression was delayed at 90 dg in LW, which could have subsequently reduced protein expression of CKM, ATP5A1, CKMT2 and ACADVL at 110 dg in LW vs. MS fetuses. ESR1 (Estrogen receptor alpha) was also seen to affect targets mostly involved in proliferation and differentiation of muscle cells such as MACROD1 (MACRO Domain Containing 1), PDIA3 (isoform 2), SEPT2 (Septin 2), ANXA2 (Annexin A2) and GSN (Figure 10). The ESR1 gene (also available in the transcriptomic dataset) was over-expressed at 110 dg with a higher expression in MS than in LW at both 90 and 110 dg. IPA identified KCNJ11, a subunit of the ATP-sensitive K⁺ channel (KATP), as a regulator of many proteins involved in energy metabolism that were up-regulated between 90 and 110 dg and over-expressed in MS compared to LW at 110 dg. Expression of KCNJ11 at the mRNA level increased between 90 and 110 dg but did not differ between genotypes. However, KCNJ11 is not a transcription factor but is involved in the regulation of K⁺ ions in response to the ATP/ADP ratio [52] and cannot be considered as a potential upstream transcriptional regulator of muscle maturity. Finally, IPA analysis identified several myogenic factors whose expression decreased between 90 and 110 dg, and was lower in MS than in LW at 110 dg (see insert in Figure 10), in accordance with better maturity of MS at birth.

Figure 10 about here.

Discussion

Embryonic MyHC, adult fast (IIa + IIx) MyHC and muscle glycogen content are good descriptors of neonatal muscle maturity

In the present experiment, the adult fast IIa and IIx MyHC were already expressed before birth, whereas no IIb was detected, which is in accordance with results obtained by [53]. In a previous study, we also found that α -cardiac MyHC was transiently expressed in pig skeletal muscle shortly after birth [29], and the present results show that the α -cardiac MyHC is already expressed in late gestation. In all genotypes, embryonic and perinatal MyHC (mRNA and protein levels) decreased at the end of gestation, whereas fast (IIa + IIx) and slow (I + α -cardiac) MyHC increased during the maturation process. These results are in agreement with those of a previous review [54] which exhaustively describes the ontogenesis of skeletal muscle in farm species. In addition, because of a high correlation between mRNA and protein expression

(Figure 2), our results suggest transcriptional regulation of MyHC expression during the maturation process. Lefaucheur et al. [55] already showed that expression of adult MyHC is very similar at the mRNA and the protein levels in pig skeletal muscle at 60 kg body weight. This transcriptional regulation has also been observed for other MyHC isoforms and in accordance with previous results in skeletal muscle in other species [56].

The maturation of skeletal muscle contractile and metabolic properties has been reported to influence piglet maturity at birth [57, 58]. Among the MyHC phenotypes, embryonic and adult fast (IIa + IIx) MyHC appear to be good markers of muscle maturity at 90 and 110 dg, respectively (Table 1). Indeed, because embryonic MyHC was highly expressed at 90 dg but decreased drastically in late gestation, it is likely a good marker of muscle immaturity at 90 dg (easily measurable because highly expressed) and its higher expression in LW than MS points to a lower maturity of LW than MS fetuses at 90 dg. Conversely, fast (IIa + IIx) MyHC was highly expressed at 110 dg and increased dramatically in late gestation, making it a potentially good descriptor of muscle maturity around birth, and its higher expression in MS than in LW fetuses points to higher maturity of MS than in LW fetuses at 110 dg. Interestingly, the hybrid fetuses were mostly intermediate between pure genotypes. In pigs, the adult fast IIa, IIx and IIb MyHC progressively replace the developmental MyHC (embryonic and perinatal MyHC) in late gestation and early after birth [54]. Because MS neonates are known to be more mature than LW, with a lower rate of mortality at birth [15], the lower expression of embryonic MyHC at 90 dg and the higher expression of adult fast (IIa and IIx) MyHC at 110 dg are representative of a better muscle maturity in MS than in LW fetuses in late gestation, the hybrid fetuses being in an intermediate position.

In addition, good correspondence was found between the increased expression of adult fast (IIa + IIx) MyHC and muscle glycogen content between 90 and 110 dg (Figure 2). Energy reserves, e.g. glycogen and lipids, must be maximal in the neonatal period because piglets are not able to oxidize protein efficiently before 5-7 days of life [57]. In pig neonates, muscle glycogen content is very high (about 9% of fresh muscle vs. 1% in growing pigs [57]) and plays a key role in whole body glycogen storage (89% of total body glycogen at birth [57]) and thermoregulation because neonatal pigs are devoid of brown adipose tissue and only have a small amount of adipose tissue (about 1.5% of body weight). The animal's energy requirement is maximum in the neonatal period to promote locomotion, thermoregulation and growth. Therefore, a high level of muscle glycogen at birth is likely involved in better neonatal maturity and piglet survival. The higher muscle glycogen content found in MS fetuses in the present study is in agreement with the better maturity previously reported in MS vs. LW neonates [15]. Interestingly, a difference between the two extreme breeds in expression of genes involved in glycogen metabolic processes was already observed during the same fetal period in the same individuals [18]. Thus, some genes were over-expressed at 110 dg in MS only, such as *PCK2* (phosphoenolpyruvate carboxykinase 2, mitochondrial also known as PEPCK 2) likely involved in gluconeogenesis from precursors derived from the citric acid cycle [18]. Therefore, the significant difference we observed in muscle glycogen content between MS and LW can be, at least partly, explained by these differences in the transcriptome. We hypothesize that piglets with high value for survival, like MS, have a higher ability to maintain glucose levels during and after farrowing, and are more able to maintain body temperature. Moreover, oxidative metabolism tends to increase during late gestation in all farm species making it an increasingly important source of energy during fetal life [54]. Skeletal muscle appears to play a key role in glycogen storage and oxidative metabolism around birth. Therefore, the muscle metabolic maturity or immaturity is indicative of the whole body metabolic maturity at the time of birth.

All three biological phenotypes were chosen for further data integration analysis. Taken together, the present data showed that embryonic MyHC, adult fast (IIa + IIx) MyHC and muscle glycogen content are good descriptors of muscle maturity in pig fetuses during late gestation, demonstrate that MS fetuses are

more mature than LW fetuses, and that hybrid fetuses are intermediate between pure MS and LW breeds. All these three biological phenotypes were used in the further data integration analysis to help find new biological markers of neonatal maturity and to advance our understanding of the underlying mechanisms.

Mitochondrial oxidative metabolism is a key determinant of neonatal muscle maturity

To identify the biological processes underlying muscular maturity, the GO terms classification was performed on the 89 unique identified proteins. A large number of proteins associated with metabolic processes were identified. However, 2D gel electrophoresis is known to reveal a limited collection of highly abundant and soluble proteins [59] covering a limited number of cell functions, mainly dealing with cell structure and metabolism, which is why proteomic and transcriptomic approaches are not always comparable [60].

Network analysis was performed and computed at the two developmental time points to identify proteins with a relevant role in the network and to identify the biological processes. It is important to note that, at both 90 and 110 dg, a large number of biological processes and cellular components overlapped between sub-networks and, within each sub-network, the enriched biological analysis often highlighted a mixture of GO terms. Indeed, it is sometimes difficult to assign a single biological function to a protein because they often play several roles in interdependent metabolic pathways. Several isoforms were found in the proteomic dataset, and were diversely intercorrelated meaning two isoforms of the same protein could be found in two different sub-networks.

The GO terms over-expressed at 90 dg were mostly involved in the actin cytoskeleton, muscle development and mRNA processing (Figure 4). As already highlighted in a previous transcriptomic study using the same individuals [18], these results are consistent with the second phase of myofiber genesis known to occur between 55 and 90 dg in pigs [54]. The total number of muscle fibers increases up to approximately 90 dg and further muscle development mostly occurs through hypertrophy and maturation of existing muscle fibers [54]. In addition, it has already been reported that the total number of myofibers is lower in MS than LW [55] which helps explain the lower postnatal muscle growth capacity of MS than LW pigs. Among the proteome sub-networks that were correlated with the biological phenotypes of interest (sub-networks 1, 2 and 4), all were characterized by GO terms involved in muscle filament development and sliding, underlining the importance of filamentous proteins and likely of the cytoskeleton at this stage. Notably, sub-network 2 also contained three isoforms of GPD1 (isoforms 1, 2 and 4) that were over-expressed in MS fetuses (Figure 7). GPD1 is a cytoplasmic enzyme involved in the glycerol phosphate shuttle that can transfer cytoplasmic glycolytic reducing NADH equivalents to mitochondrial FADH at complex 2, thus providing ATP to the cell through the respiratory chain [61]. The higher expression of these three GPD1 isoforms in MS at 90 dg could suggest increased mitochondrial oxidation of cytosolic NADH in MS, which already contributes to the advanced maturity of MS fetuses. Surprisingly, sub-networks 3 and 5, which were mostly involved in energy metabolism, had no significant correlation with MyHC and muscle glycogen content at 90 dg. On the whole, our data suggest that muscle immaturity at 90 dg is primarily related to the high proportion of cytoskeletal proteins and proteins involved in myofibril assembly, even though some metabolic enzymes such as GPD1 could also be positively correlated with the maturation process.

At 110 days of gestation, the most striking differences between genotypes concerned the mitochondria cellular component, in particular the mitochondrial oxidation/reduction molecular function (Figure 4). Moreover, the proteome sub-network analysis identified gluconeogenesis and glycolysis as important co-expressed pathways that could explain the higher muscle glycogen content in MS than in LW at

110 dg. Analysis of the proteome sub-network identified five sub-networks that were correlated with biological phenotypes of interest (sub-networks 2, 3, 4, 5 and 6). Most were characterized by GO terms primarily involved in energy metabolism (sub-networks 2, 3, 5 and 6), whereas GO terms dealt mostly with the proteasome complex and mRNA processing in sub-network 4 with PSMC5 (isoform 1) as the most important node in the cluster. The most relevant sub-network was sub-network 2 with 24 nodes and 65 edges. This sub-network was highly correlated with adult fast and embryonic MyHC, and its GO terms primarily concerned mitochondrial energy metabolism. Interestingly, CKMT2 and ATP5A1 were identified as important nodes in this cluster which could be physiologically relevant for the production of energy such as ATP. Indeed, CKMT2 is a creatine kinase and is responsible for the transfer of high energy phosphate from the mitochondria to the cytosolic compartment, and at the same time for returning ADP to the respiratory system, thereby stimulating oxidative phosphorylation. Moreover, cytoplasmic muscle CKM was also over-expressed in MS at 110 dg, suggesting that the energy metabolism of muscle contraction was higher in MS than LW. ATP5A1 encodes a subunit of the mitochondrial ATP synthase that converts the mitochondrial electrochemical H^+ gradient to ATP production, thereby supplying ATP to the muscle. CKMT2 and ATP5A1 are likely good biological markers of muscle maturity at 110 dg (Figure 7). The greater expression of CKMT2 and ATP5A1 in MS at 110 dg indicates that MS fetuses possess greater oxidative capacity, in accordance with previous results showing greater enzyme activities of citrate synthase and hydroxy acyl-CoA dehydrogenase in MS than LW muscle at birth [62]. Sub-network 3 was the second most important cluster with 19 nodes and 46 edges. It was highly correlated with glycogen content as well as embryonic and adult fast MyHC, and its GO terms mostly concerned glycolytic energy metabolism and gluconeogenesis. Sub-network 6 was highly correlated with muscle glycogen content and logically, one of its GO terms corresponded to gluconeogenesis. Moreover, two GPD1 isoforms (isoforms 2 and 4) were also present in sub-network 6 and under-expressed in LW fetuses, as already observed at 90 dg (Figure 7). Notably, expression of the four GPD1 isoforms was strongly influenced by the genotype with no effect of age, and no interaction between genotype and development time points was observed (Additional File 3). This suggests that GPD1 is genetically determined but not related to maturity as influenced by age. At the mRNA level, as already demonstrated in [18], GPD1 gene expression was lower in LW than MS at 90 dg and did not increase in LW during the maturation process, whereas it did increase in MS. Therefore, the gestational time point only had an effect on GPD1 at the mRNA level in MS. Altogether, the present data show that muscle energy metabolism, in particular mitochondrial energy metabolism and muscle glycogen storage, increased dramatically between 90 and 110 dg and were higher in MS than in LW fetuses at 110 dg, meaning the mitochondrial oxidation/reduction process and glycogen storage play a crucial role in the late fetal muscle maturation process.

Biological data integration identified PPARGC1A as potential upstream transcriptional regulator of muscle maturity at birth

To gain further insight into the molecular machinery underlying the muscle maturation process in pigs, we undertook a guided and integrated analysis of the transcriptomic and proteomic datasets. Among the 89 unique proteins identified, 81 were available in the transcriptomic dataset. Using Pearson's correlations, mRNA and protein expression levels were analyzed to find proteins with possible transcriptional regulation. Thirty-one proteins were identified with a significant positive correlation between mRNA and protein expression levels in both extreme fetal genotypes (Pearson's correlation $|r| > 0.7$) (Figure 9). The correlation between mRNA and protein abundances in the cell has been reported to be notoriously poor. A number of transcriptome and proteome data integration studies have already been reported *e.g.* in cell lines [63–65], plants [66], and mammal models [6,67–70]. At 110 dg, sub-network 2 was composed of 10 nodes (out of 24) with possible transcriptional regulation in both MS and LW. Proteins belonging to this sub-network generally exhibited similar expression profiles at 110 dg, were mostly located in the same cellular component and may possibly be regulated by common transcription factors (TFs). Based on a

bibliographic network computed using IPA software (using the upstream regulator function), we were able to find possible upstream TFs. PPARGC1A was one of the TFs identified. It was a regulator of CKMT2, ACADVL and ATP5A1 (Figure 10). CKMT2 has already been identified as an important node (the highest betweenness for CKMT2 at 110 dg) and as a likely good biological marker of muscle maturity at 110 dg.

PPARGC1A (also known as PGC1-alpha) is a transcriptional coactivator that controls the expression of many genes through a whole range of nuclear hormone receptors and other TFs [71]. Moreover, the activity of PPARGC1A is regulated by phosphorylation and deacetylation through the coaction of two upstream metabolic sensors of energy deficiency: AMPK and SIRT1, respectively [72]. PPARGC1A is abundant in muscle [73] where it is involved in several biological functions such as mitochondrial biogenesis, and oxidative metabolism, which play a key role in ATP production and the adaptation of muscle to exercise and exposure to cold [74]. It has been shown to drive the formation of slow twitch fibers [75] and to be more highly expressed in oxidative myofibers [72]. Therefore, together with AMPK and SIRT1, PPARGC1A is involved in carbohydrate and lipid metabolisms to maintain energy homeostasis. In post-natal growing pigs, PPARGC1A is more highly expressed in LM of Erhualian than LW pigs, in accordance with a higher proportion of oxidative fibers in Erhualian pigs [76], as well as in MS pigs [55]. Voillet et al. [18] showed that PPARGC1A gene expression was differentially expressed depending on the interaction between fetal genotypes and gestational time points, with higher expression in MS than LW at 90 dg. Therefore, our results suggest that PPARGC1A could have a precocious effect on the subsequent increase in expression of genes such as CKMT2, ATP5A1, ACADVL and CKM in MS than LW at 110 dg. Mitochondrial CKMT2 and cytoplasmic CKM are both involved in the creatine metabolic process and are known to play a key role in the transfer of energy within the muscle fiber for muscle contraction and thermogenesis. Interestingly, over-expression of PPARGC1A in the skeletal muscle of transgenic mice has been shown to increase glycogen synthesis and storage [77], which could also explain the greater glycogen content observed in MS compared to LW fetuses at 110 dg in the present experiment. In addition, greater variability of PPARGC1A expression in LW than in MS was also observed at 110 dg. Therefore, PPARGC1A could be a relevant upstream regulator involved in the accelerated muscle maturation observed in MS compared to LW in late gestation.

Cell culture experiments have shown that PPARGC1A is a coactivator of ESR1 [78]. In our transcriptomic study, ESR1 expression increased dramatically between 90 and 110 dg and was greater in MS than LW fetuses at both stages (Figure 10). In sub-network 2 at 110 dg, ESR1 was found to be a regulator of several proteins mostly involved in protein folding (PDIA) and in cytoskeleton (SEPT2 and GSN). This estrogen receptor has been reported to be involved in estrogen-mediated regulation of substrate metabolism [79] and lower mRNA expression has been observed in the adipose tissue of obese compared to lean women [80]. In ESR1 knock out mice, body weight was about 30% higher than in wild-type mice: mice were obese and their oxidative metabolism was impaired [81]. In our study, under-expression of ESR1 in LW was accompanied by reduced oxidative metabolism but not by increased fatness (LW leaner than MS). Altogether, ESR1 increased between 90 and 110 dg, and its over-expression in MS compared to LW could be involved, in interaction with PPARGC1A, in the improved maturity of MS compared to LW animals at birth.

Conclusion

In this study, we developed an innovative strategy combining state-of-the-art statistical and computational methods to integrate multi-omics datasets to gain a deeper insight into the late fetal maturation process. Three biological phenotypes of interest (i.e. adult fast and embryonic MyHC and muscle glycogen content) were identified as good descriptors of maturity. Some proteins, involved in oxidative metabolism and

with a possible transcriptional regulation, were also emphasized as possible biomarkers of the maturation process. As one of the main sources of energy during late fetal life, muscle oxidative metabolism is very important at birth [54]. In pigs, a higher rate of death at birth has already been observed in genotypes with a lower percentage of oxidative fibers [15,55]. In recent decades, the pig industry has been focusing on selection for rapid production of lean meat, low back-fat thickness and a rapid growth rate, thus influencing muscle fiber properties [82]. More precisely, a selection for lean tissue is more associated with muscles containing a low percentage of oxidative myofibers and a high percentage of large glycolytic myofibers [83]. This kind of the selection could have an effect on important genes, such as PPARC1A, with lower expression in skeletal muscle, which could contribute to an alteration or delay of mitochondrial gene/protein expression in late gestation and lead to muscle metabolic immaturity at birth.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We are grateful to P. Ecolan and S. Tacher for 2D gel analyses, C. Tréfeu and S. Daré for RT-PCR analyses and glycogen determination, respectively, and D. Viala for LC-MS/MS protein identification. We also wish to thank D. Goodfellow for English revision. We also thank Kim-Anh Lê Cao for her relevant comments.

Funding

Animal Genetics Division (INRA, <http://www.ga.inra.fr/en/>), Animal Physiology and Livestock Systems (INRA, <http://www.phase.inra.fr/en/>) and *Région Languedoc Roussillon Midi Pyrénées* (<http://www.regionlrmp.fr/>) (to V.V.); French 'Agence Nationale de la Recherche' Porcine grant [ANR-09-GENM005, <http://www.agence-nationale-recherche.fr/>]. Funding for open access: INRA, UMR1388 Génétique, Physiologie et Systèmes d'Élevage, Castanet-Tolosan, F-31326, France.

Author contributions

LLi, LC and LL conceived and designed the study. YB supervised the performance testing, from animal production to biological sampling. VV analyzed the expression datasets supervised by MSC, LL and LLi. VV drafted the manuscript with help of MSC, LL and LLi. LLi and LL supervised the project. All authors read and approved the final manuscript.

References

1. Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11: 31-46.
2. Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 12: 87-88.
3. Altelaar AFM, Munoz J, Heck AJR (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet* 14: 35-48.
4. Shulaev V (2006) Metabolomics technology and bioinformatics. *Brief Bioinform* 7: 128-139.

5. Haider S, Pal R (2013) Integrated analysis of transcriptomic and proteomic data. *Curr Genomics* 14: 91-110.
6. Cox B, Kislinger T, Emili A (2005) Integrating gene and protein expression data: pattern analysis and profile mining. *Methods* 35: 303-314.
7. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 27: 117-124.
8. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. (2013) Global quantification of mammalian gene expression control. *Nature* 473: 337-342.
9. Ritchie M, Holzinger E, Li R, Pendergrass S, Kim D (2015) Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* 16: 85-97.
10. Barabási A, Oltvai Z (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5: 101-113.
11. Luscombe N, Babu M, Yu H, Snyder M, Teichmann S, et al. (2004) Genomics analysis of regulatory network dynamics reveals large topological changes. *Nature* 431: 308-312.
12. Mitra K, Carvunis A, Ramesh S, Ideker T (2013) Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet* 14: 719-732.
13. Villa-Vialaneix N, Liaubet L, Laurent T, Cherel P, Gamot A, et al. (2013) The structure of a gene co-expression network reveals biological functions underlying eqtls. *PloS ONE* 8: e60045.
14. Canario L, Cantoni E, Le Bihan E, Caritez JC, Billon Y, et al. (2006) Between-breed variability of stillbirth and its relationship with sow and piglet characteristics. *J Anim Sci* 84: 3185-3196.
15. Canario L, Pere MC, Tribout T, Thomas F, David C, et al. (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.
16. Leenhouwers J, Knol E, de Groot P, Vos H, van der Lende T (2002) Fetal development in the pig relation to genetic merit for piglet survival. *J Anim Sci* 80: 1759-1770.
17. Leenhouwers J, Knol E, van der Lende T (2002) Differences in late prenatal development as an explanation for genetic differences in piglet survival. *J Anim Sci* 78: 57-62.
18. Voillet V, SanCristobal M, Lippi Y, Martin P, Iannuccelli N, et al. (2014) Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity. *BMC Genomics* 15: 797.
19. Miller DR, D B, Jackson RB, Downie EF, R RJ (2010) Metabolic maturity at birth and neonate lamb survival: Association among maternal factors, litter size, lamb birth weight, and plasma metabolic and endocrine factors on survival and behavior. *J Anim Sci* 88: 581-593.
20. Lawn JE, Cousens S, Zupan J (2005) 4 million neonatal deaths: when? where? why? *The Lancet* 365: 891-900.
21. Basso O, Wilcox A (2010) Mortality risk among preterm babies: immaturity versus underlying pathology. *Epidemiology* 21: 521-527.
22. Foxcroft GR, Dixon WT, Novak S, Putman CT, Town SC, et al. (2006) The biological basis for prenatal programming of postnatal performance in pigs. *J Anim Sci* 84: E105-E112.

23. Good CA, Kramer H, Somogyi M (1933) The determination of glycogen. *J Biol Chem* 100: 485-491.
24. Montagne L, Loisel F, Le Naou T, Gondret F, Gilbert H, et al. (2014) Difference in short-term responses to a high-fiber diet in pigs divergently selected for residual feed intake. *J Anim Sci* 92: 1512-1523.
25. Perruchot MH, Lefaucheur L, Louveau I, Mobuchon L, Palin MF, et al. (2015) Delayed muscle development in small pig fetuses around birth cannot be rectified by maternal early feed restriction and subsequent overfeeding during gestation. *Animal* 9: 1996-2005.
26. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acid Res* 29: e45.
27. Schiaffino S, Reggiani C (2011) Fiber types in mammalian skeletal muscles. *Physiol Rev* 91: 1447-1531.
28. Lefaucheur L, Edom F, Ecolan P, Butler-Browne GS (1995) Pattern of muscle fiber type formation in the pig. *Dev Dynam* 203: 27-41.
29. Lefaucheur L, Hoffman R, Okamura C, Gerrard D, Leger JJ, et al. (1997) Transitory expression of alpha cardiac myosin heavy chain in a subpopulation of secondary generation muscle fibers in the pig. *Dev Dynam* 210: 106-116.
30. Lefaucheur L, Ecolan P, Lossec G, Gabillard JC, Butler-Browne GS, et al. (2001) Influence of early postnatal cold exposure on myofiber maturation in pig skeletal muscle. *J Muscle Res Cell M* 22: 439-452.
31. Vincent A, Louveau I, Gondret F, Trefeu C, Gilbert H, et al. (2015) Divergent selection for residual feed intake affects the transcriptomic and proteomic profiles of pig skeletal muscle. *J Anim Sci* 93: 2745-2758.
32. Breiman L (2001) Random forests. *Mach Learn* 45: 5-32.
33. Lê Cao KA, Boitard S, Besse P (2011) Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* 12: 253.
34. Lê Cao KA, Rossouw D, Robert-Granie C, Besse P (2008) A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol* 7: article 35.
35. Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10: 515-534.
36. Francois Y, Marie-Etancelin C, Vignal A, Viala S D an Davail, Molette C (2014) Mule duck 'foie gras' shows different metabolic states according to its quality phenotype by using a proteomic approach. *J Agr Food Chem* 62: 7140-7150.
37. Team RC (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
38. Reverter A, Chan E (2008) Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* 24: 2491-2497.
39. Hudson N, Reverter A, Wang Y, Greenwood P, Dalrymple B (2009) Inferring the transcriptional landscape of bovine skeletal muscle by integrating co-expression networks. *PLoS ONE* 4: e7249.

40. Pérez-Montarelo D, Hudson N, Fernández A, Ramayo-Caldas Y, Dalrymple B, et al. (2012) Porcine tissue-specific regulatory networks derived from meta-analysis of the transcriptome. *PLoS ONE* 7: e46159.
41. Watson-Haigh N, Kadarmideen H, Reverter A (2010) PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches. *Bioinformatics* 26: 411-413.
42. Newman M, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E* 69: 026113.
43. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. *Phys Rev E* 74: 016110.
44. Montastier E, Villa-Vialaneix N, Caspar-Bauguil S, Hlavaty P, Tvrzicka E, et al. (2015) System model network for adipose tissue signatures related to weight changes in response to calorie restriction and subsequent weight maintenance. *PLoS Comput Biol* 11: e1004047.
45. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acid Res* 1: 214-220.
46. Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A (2012) GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acid Res* 10: 478-483.
47. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* 57: 289-300.
48. Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal Complex Systems*: 1695.
49. Laurent T, Villa-Vialaneix N (2011) Using spatial indexes for labeled network analysis. *Information, Interaction, Intelligence* : 11.
50. Noack A (2009) Modularity clustering is force-directed layout. *Phys Rev* 79: 026102.
51. Bastian M, Heymann S, Jacomy M (2009) Gephi: An open source software for exploring and manipulating networks. In: International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence.
52. Tricarico D, Selvaggi M, Passantino G, De Palo P, Dario C, et al. (2016) ATP sensitive potassium channels in the skeletal muscle function: Involvement of the KCNJ11(kir6.2) gene in the determination of mechanical warner bratzer shear force. *Front Physiol* 7: 167.
53. Chang KC, Fernandes K (1997) Developmental expression and 5' end cDNA cloning of the porcine 2x and 2b myosin heavy chain genes. *DNA Cell Biol* 16: 1429-1437.
54. Picard B, Lefaucheur L, Berri C, Duclos J (2002) Muscle fibre ontogenesis in farm animal species. *Reprod Nutr Dev* 42: 415-431.
55. Lefaucheur L, Milan D, Ecolan P, Le Callennec C (2004) Myosin heavy chain composition of different skeletal muscles in large white and meishan pigs. *J Anim Sci* 82: 1931-1941.
56. Cox RD, Buckingham ME (1992) Actin and myosin genes are transcriptionally regulated during mouse skeletal muscle development. *Dev Biol* 149: 228-234.

57. Herpin P, Damon M, LeDividich J (2002) Development of thermoregulation and neonatal survival in pigs. *Livest Prod Sci* 78: 25-45.
58. Rehfeldt C, Lefaucheur L, Block J, Stabenow B, Pfuhl R, et al. (2012) Limited and excess protein intake of pregnant gilts differently affects body composition and cellularity of skeletal muscle and subcutaneous adipose tissue of newborn and weanling piglets. *Eur J Nutr* 51: 151-165.
59. Petrak J, Ivanek R, Toman O, Cmejla R, Cmejlova J, et al. (2008) Déjà vu in proteomics. a hit parade of repeatedly identified differentially expressed proteins. *Proteomics* 8: 1744-1749.
60. Wang P, Bouwman FG, Mariman ECM (2009) Generally detected proteins in comparative proteomics - a matter of cellular stress response? *Proteomics* 9: 2955-2966.
61. Kornberg A, Pricer W (1953) Enzymatic esterification of alpha-glycerophosphate by long chain fatty acids. *J Biol Chem* 204: 345-357.
62. Bonneau M, Mourot J, Noblet L, Lefaucheur L, Bidanel JP (1990) Tissue development in meishan pigs: Muscle and fat development and metabolism and growth regulation by somatotropic hormone. *Chinese Pig Symp* : 202-213.
63. Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 19: 1720-1730.
64. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, et al. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292: 929-934.
65. Schmidt MW, Houseman A, Ivanov AR, Wolf DA (2007) Comparative proteomic and transcriptomic profiling of the fission yeast *Schizosaccharomyces pombe*. *Mol Syst Biol* 3: 79.
66. Yin L, Tao Y, Zhao K, Shao J, Li X, et al. (2007) Proteome and transcriptomic analysis of rice mature seed-derived callus differentiation. *Proteomics* 7: 755-768.
67. Tian Q, Stepaniants SB, Mao M, Weng L, Feethal MC, et al. (2004) Integrated genomic and proteomic analyses of gene expression in mammalian cells. *Mol Cell Proteomics* 3: 960-969.
68. Xun Z, Sowell RA, Kaufman TC, Clemmer DE (2007) Protein expression in a drosophila model of parkinson's disease. *J Proteome Res* 6: 348-357.
69. Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, et al. (2011) Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet* 7: e1001393.
70. Waters KM, Liu T, Quesenberry RD, Willse AR, Bandyopadhyay S, et al. (2012) Network analysis of epidermal growth factor signaling using integrated genomic, proteomic and phosphorylation data. *PLoS ONE* 7: e34515.
71. Wu Z, Puigserver P, Andersson U, Zhang C, Adelmant G, et al. (1999) Mechanisms controlling mitochondrial biogenesis and respiration through the thermogenic coactivator PGC-1. *Cell* 98: 155-124.
72. Jäger S, Handschin C, St-Pierre J, Spiegelman BM (2007) Amp-activated protein kinase (ampk) action in skeletal muscle via direct phosphorylation of pgc-1alpha. *P Natl Acad Sci USA* 104: 12017-12022.
73. Oberkofler H, Esterbauer H, Linnemayr V, Strosberg A, Krempler F, et al. (2002) Peroxisome proliferator-activated receptor (ppar) alpha coactivator-1 recruitment regulates ppar subtype specificity. *J Biol Chem* 277: 16750-16757.

74. Chan MC, Arany Z (2014) The many roles of PGC-1alpha in muscle—recent developments. *Metabolism* 63: 441-451.
75. Lin J, Wu H, Tarr P, Zhang C, Wu Z, et al. (2002) Transcriptional co-activator pgc-1alpha drives the formation of slow-twitch muscle fibres. *Nature* 418: 797-801.
76. Zhao R, Yang X, Xu Q, Wei X, Xia D, et al. (2004) Expression of GHR and PGC-1alpha in association with changes of myhc isoform types in longissimus muscle of erhualian and large white pigs (*sus scrofa*) during postnatal growth. *Anim Sci* 79: 203-211.
77. Wende AR, Schaeffer PJ, Parker GJ, Zechner C, Han DH, et al. (2007) A role for the transcriptional coactivator PGC-1alpha in muscle refueling. *J Biol Chem* 282: 36642-51.
78. Tcherepanova I, Puigserver P, Norris JD, Spiegelman BM, McDonnell DP (2000) Modulation of estrogen receptor-alpha transcriptional activity by the coactivator PGC-1. *J Biol Chem* 275: 16302-16308.
79. Heine PA, Taylor JA, Iwamoto GA, Lubahn DB, Cooke PS (2000) Increased adipose tissue in male and female estrogen receptor-alpha knockout mice. *Proc Natl Acad Sci USA* 97: 12729-12734.
80. Nilsson M, Dahlman I, Rydén M, Nordström EA, Gustafsson JA, et al. (2007) Oestrogen receptor alpha gene expression levels are reduced in obese compared to normal weight females. *Int J Obesity* 31: 900-907.
81. Ribas V, Audrey Nguyen MT, Henstridge DC, Nguyen AK, Beaven SW, et al. (2010) Impaired oxidative metabolism and inflammation are associated with insulin resistance in er alpha-deficient mice. *Am J Physiol-Endoc M* 298: 304-319.
82. Lefaucheur L (2010) A second look into fibre typing—relation to meat quality. *Meat Sci* 84: 257-270.
83. Brocks L, Klont R, Buist W, de Greef K, Tieman M, et al. (2000) The effects of selection of pigs on growth rate vs leanness on histochemical characteristics of different muscles. *J Anim Sci* 78: 1247-1254.

Figure

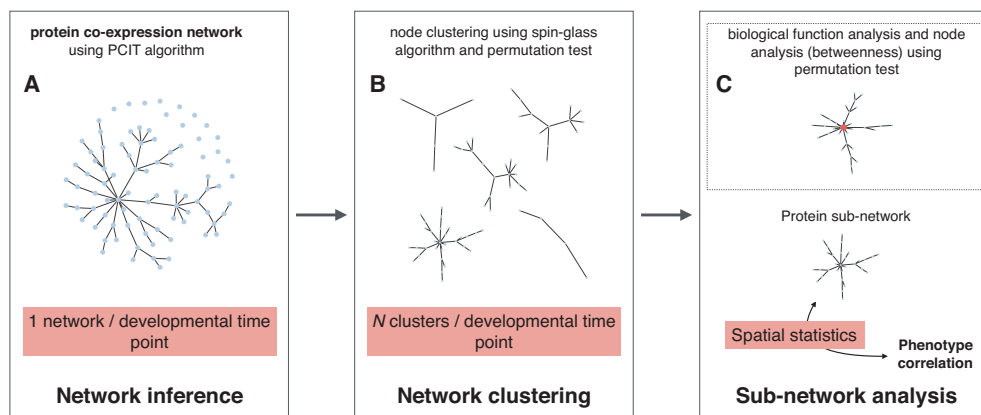


Figure 1. Overview of the network data integration analysis. (A) Proteome networks were first inferred using the PCIT (partial correlation and information theory) algorithm for each developmental time point: 90 days of gestation and 110 days of gestation. PCIT, which belongs to the family of weighted network algorithms, is based on the combination of the concept of partial correlation coefficient and the information theory to identify meaningful associations. (B) Network clustering was then performed using a spin-glass algorithm. A permutation test was performed to assess the significance of clustering. (C) Sub-networks (clusters) were analyzed. (i) GeneMANIA was used to assess the relevance of the proposed sub-networks and GO enrichment analysis was performed to analyze the biological functions of the sub-networks. (ii) Centrality (or betweenness) of nodes was analyzed using a permutation test. (iii) Using spatial statistics tools, correlations between sub-networks and biological phenotypes of interest were analyzed.

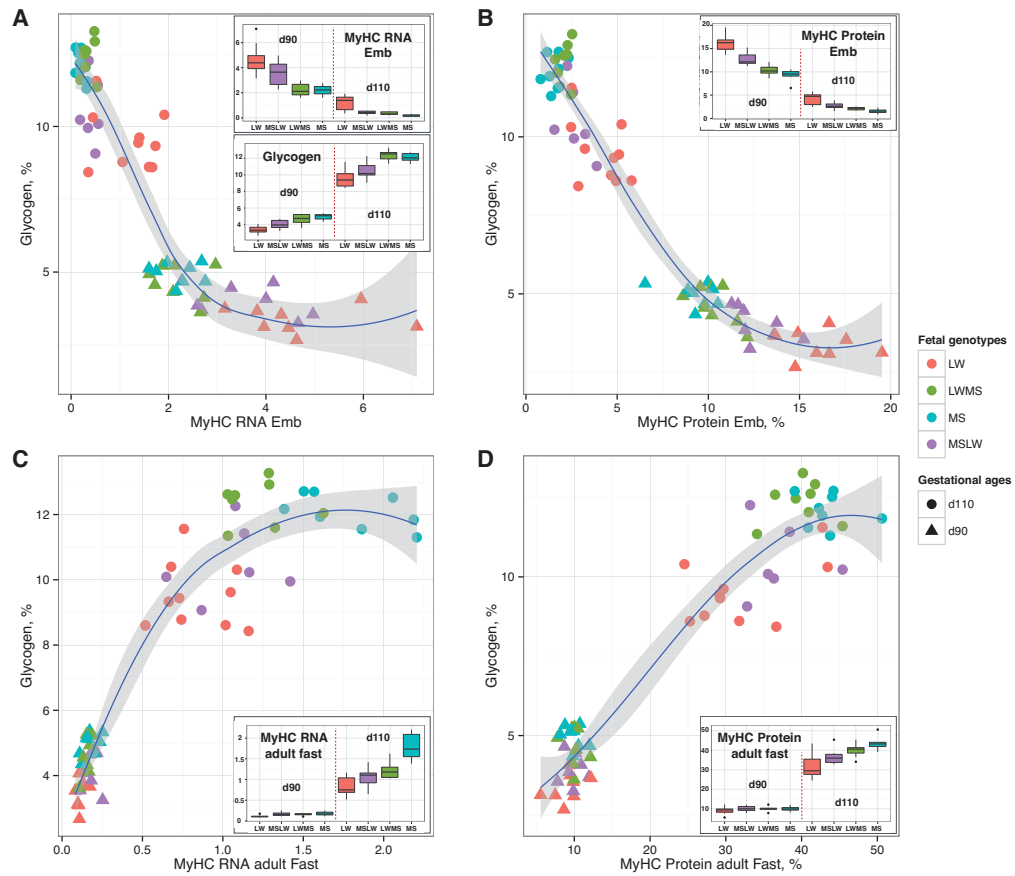


Figure 2. Co-variation of muscle glycogen and expression profiles (mRNA and proteins) of myosin heavy chain (MyHC) genes (adult fast (IIa + IIb + IIx) and embryonic MyHC). Lowess curves and confidence intervals are in blue and grey, respectively. Box-plots of mRNA or protein expression and glycogen content are also shown. **(A)** mRNA embryonic MyHC. **(B)** Protein embryonic MyHC. **(C)** mRNA adult Fast (IIa + IIb + IIx) MyHC. **(D)** Protein adult Fast (IIa + IIb + IIx) MyHC.

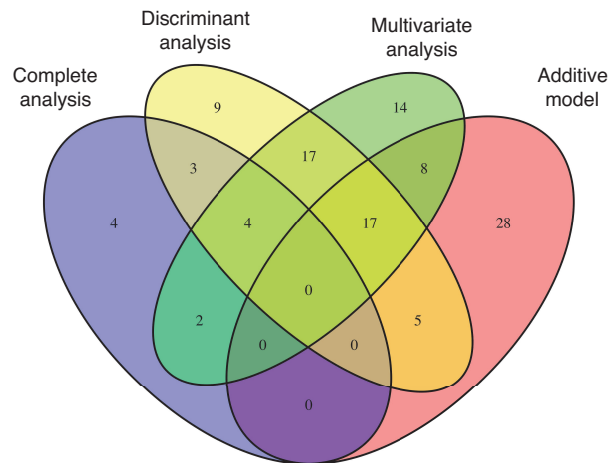


Figure 3. Global analysis of the 113 identified proteins. Venn diagram of the statistical methods chosen to select spots for identification. 13 proteins (in blue) were differentially expressed according to the interaction between developmental time points and fetal genotypes. 58 proteins (in red) were differentially expressed in an additive manner between development time points and fetal genotypes. 62 proteins (in green) were selected using sparse multivariate methods (both sCCA and sPLS). 55 proteins (in yellow) were selected using discriminant analyses (both RF and sPLS-DA). See the Materials and Methods section for further information.

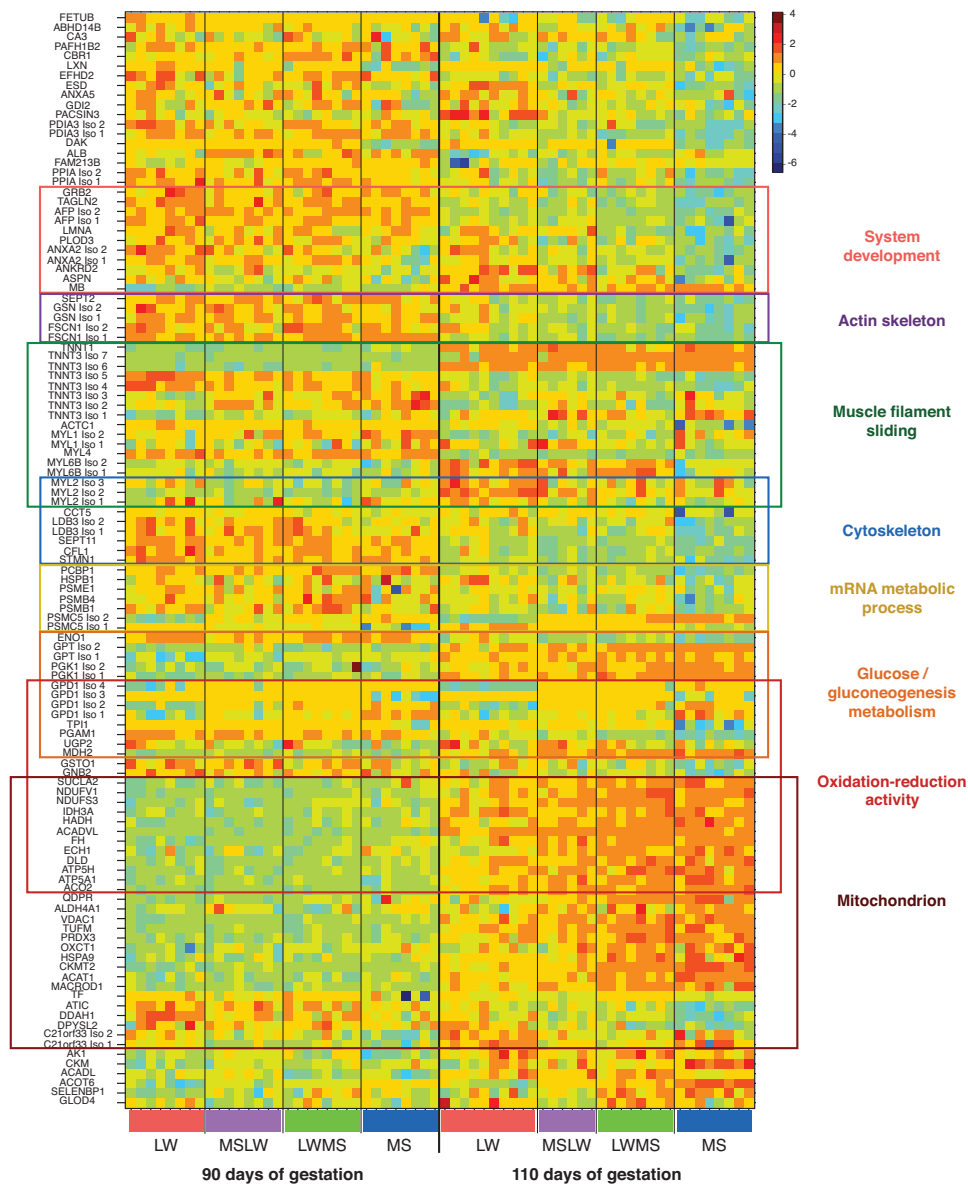


Figure 4. Level plot of the 113 identified spots. All spots were ranged according to the biological function. Muscle development functions and components, *e.g.* system development, actin skeleton, muscle filament sliding, cytoskeleton and mRNA metabolic process, were mostly over-expressed at 90 dg. Energy metabolism functions and components, *e.g.* glucose and gluconeogenesis metabolism, oxidation-reduction activity and mitochondrion, were mostly over-expressed at the end of gestation (110 dg). All the biological functions and components are not included in this figure.

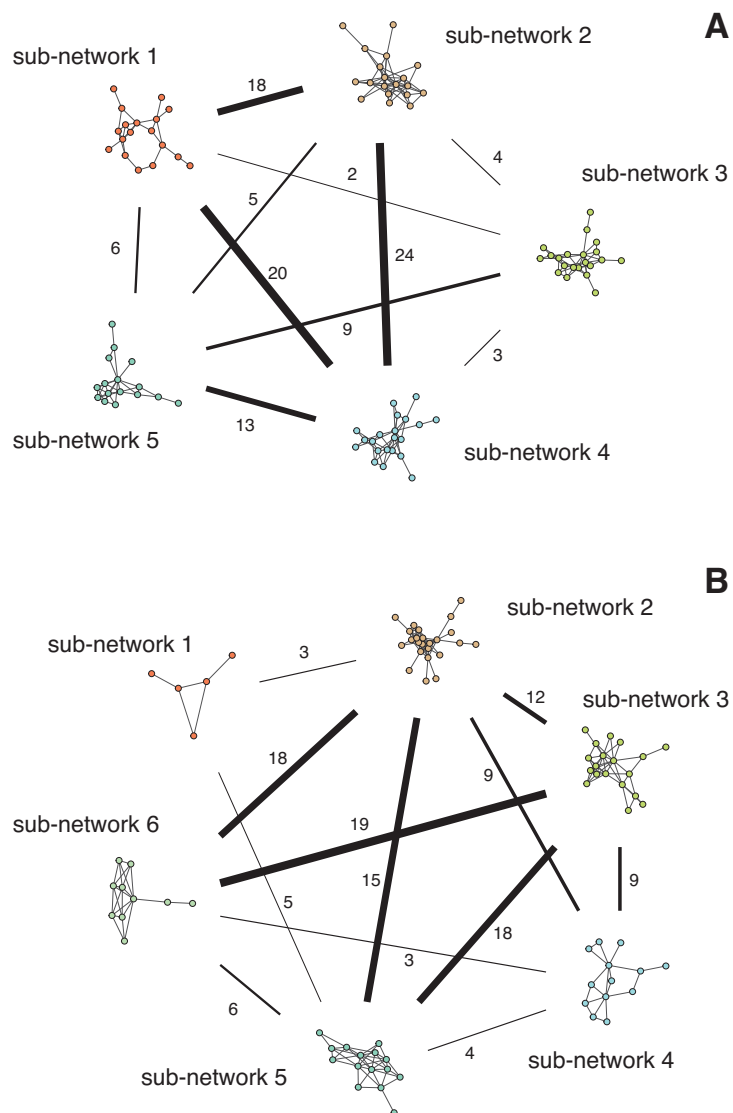


Figure 5. Decomposition of the proteome networks. (A) Each sub-network represents a cluster in the d90-proteome network. (B) Each sub-network represents a cluster in the d110-proteome network. The width of the edge is proportional to the number of edges between the two corresponding sub-networks.

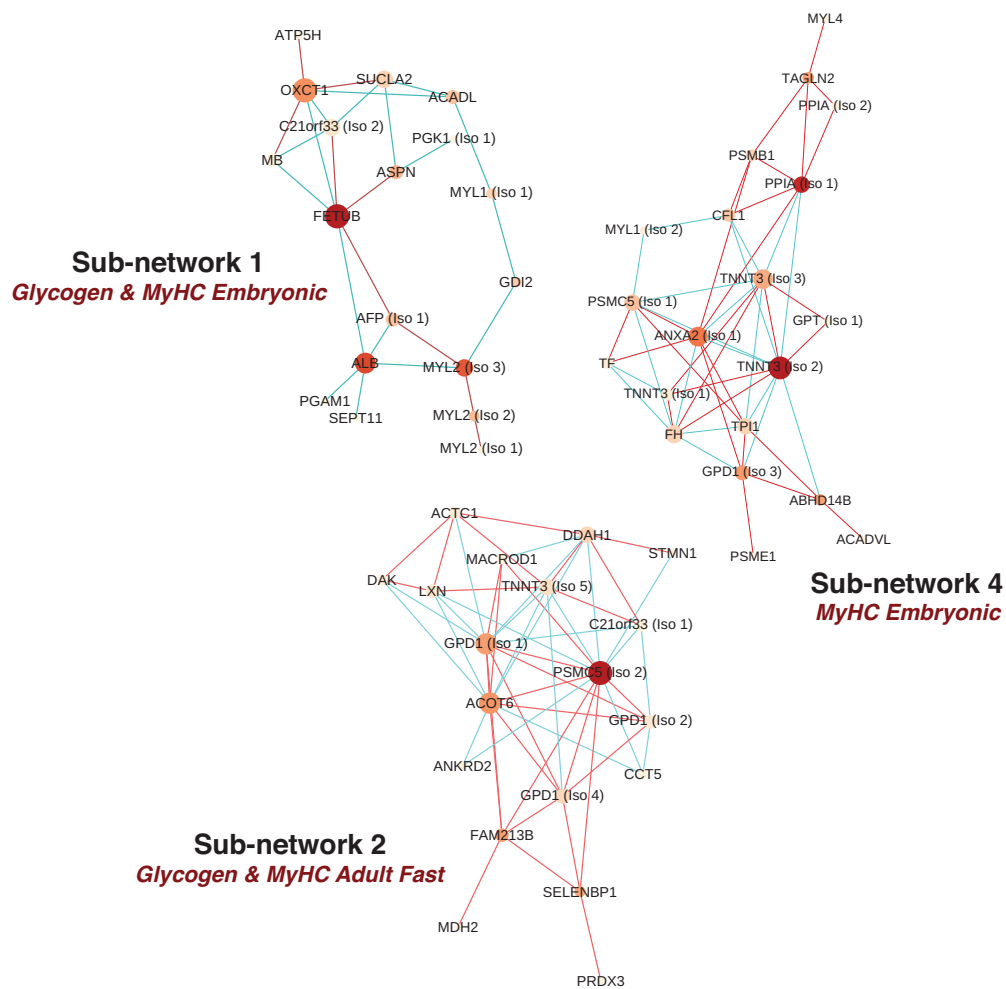


Figure 6. Three sub-networks obtained at 90 days of gestation. A PCIT model was used to infer a co-expression network. Nodes were clustered using a spin-glass model. The figure shows three (out of five) sub-networks (1, 2 and 4) obtained by clustering. The color of the nodes indicates betweenness centrality. The red nodes have the highest betweenness. The size of the nodes indicates degree. The color of the edges indicates the correlation sign: red for positive correlations and blue for negative ones.

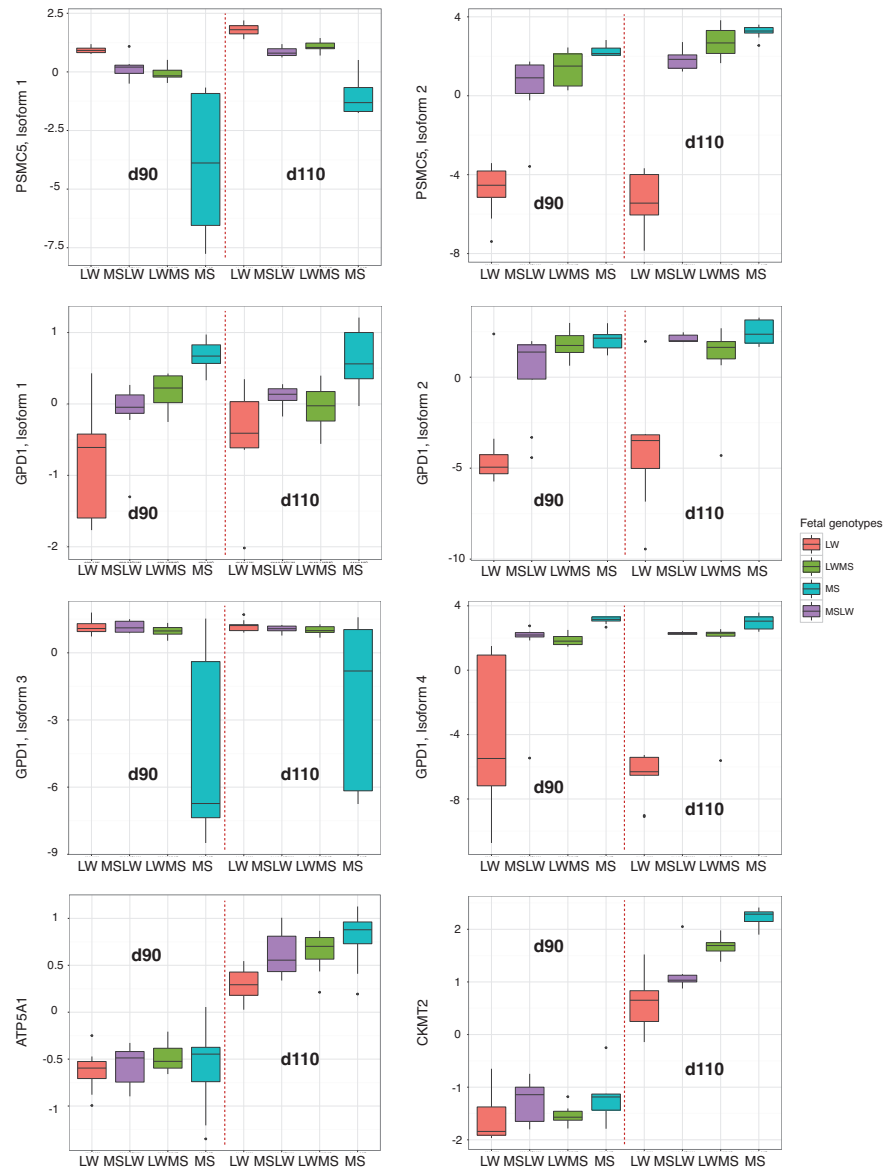


Figure 7. Box plots of protein level expression. PSMC5 with two isoforms, GPD1 with four isoforms, ATP5A1 and CKMT2.

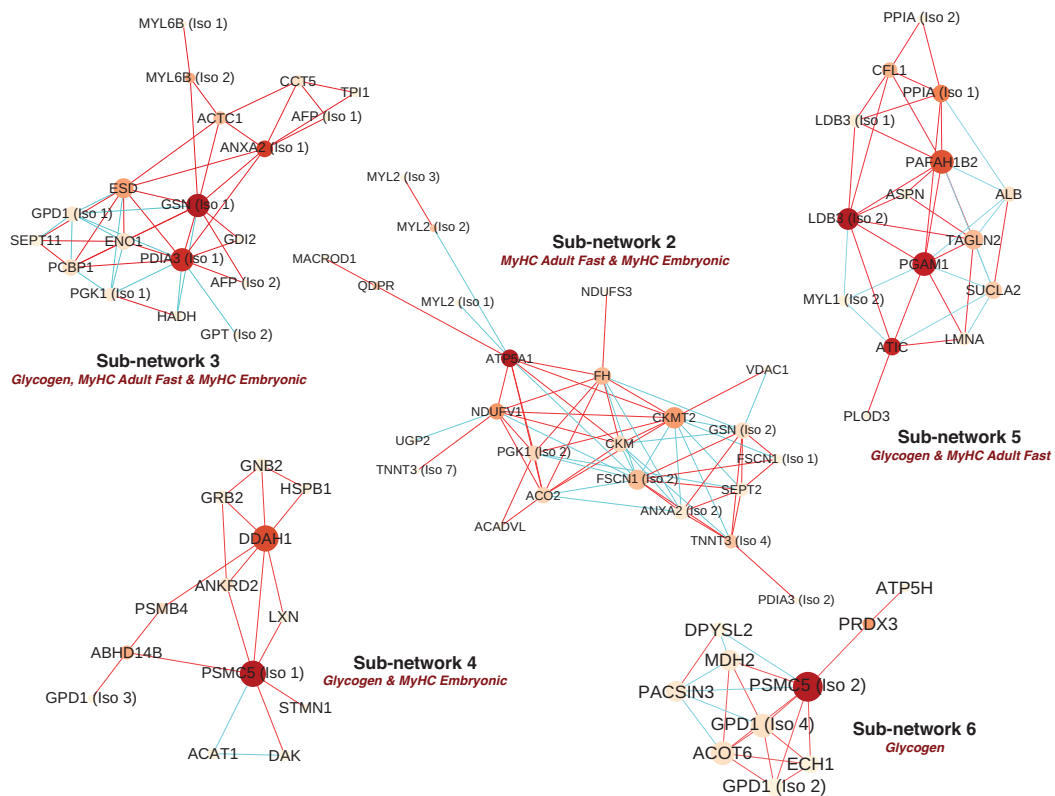


Figure 8. Five sub-networks obtained at 110 days of gestation. A PCIT model was used to infer a co-expression network. Nodes were clustered using a spin-glass model. This figure shows five (out of six) sub-networks (2, 3, 4, 5 and 6) obtained by clustering. The color of the nodes indicates betweenness centrality. The red nodes have the highest betweenness. The size of the nodes indicates degree. The color of the edges indicates the correlation sign: red for positive correlations and blue for negative ones.

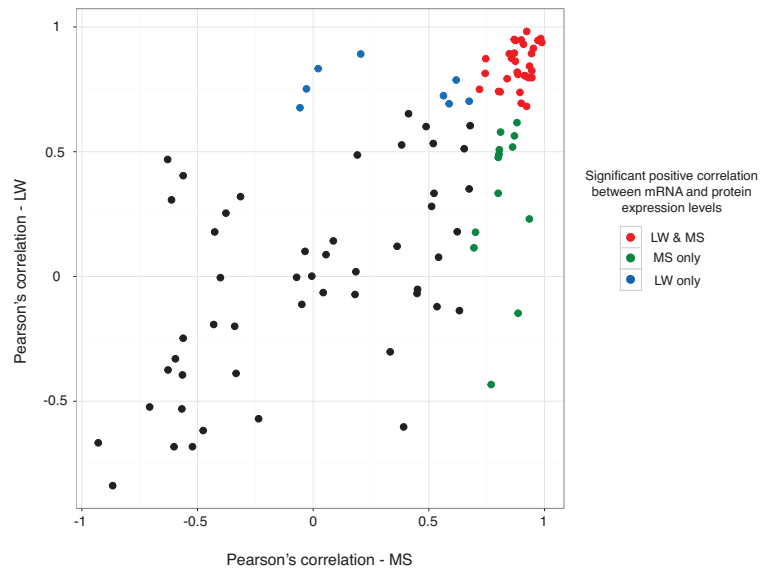


Figure 9. Correlation between transcriptome and proteome in both extreme fetal genotypes (LW and MS). Scatter plot of correlations between mRNA and protein expression levels in MS (x-axis) and LW (y-axis). Each dot represents the correlation in MS and LW of a pair of gene-protein. Red dots show a significant positive correlation in LW and MS, the green dots show a significant positive correlation only in MS, and blue dots show a significant positive correlation only in LW. It is important to note that more than 81 points were represented because some proteins had several isoforms.

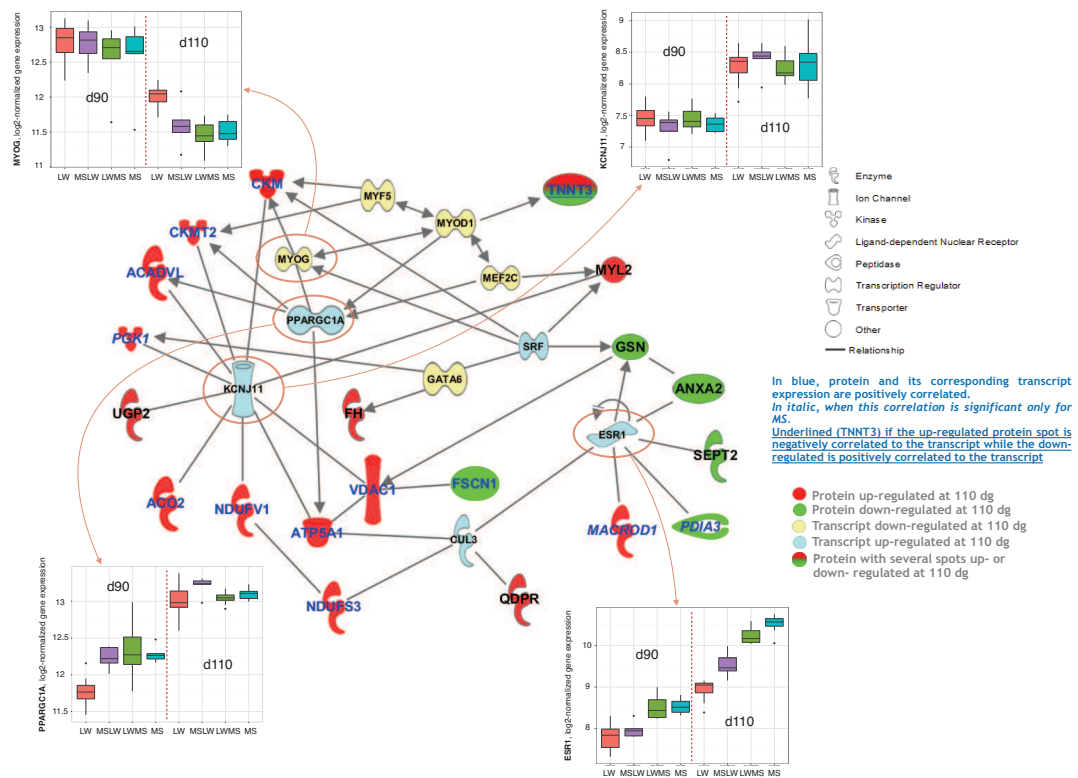


Figure 10. Bibliographic network obtained with the d110-sub-network 2. This network is composed of 30 nodes. Ten nodes out of the 24 composing sub-network 2 were found to be possibly transcriptionally regulated. Box plots of PPARGC1A, KCN11, MYOG and ESR1 mRNA expression from [18] in the present experiment are also shown.

Tables

	90 days of gestation				110 days of gestation				RSD ¹	P-values ²		
	LW	MSLW	LWMS	MS	LW	MSLW	LWMS	MS		A	G	AxG
MyHC Protein level, % total MyHC												
Embryonic	16.20 ^a	12.71 ^b	10.31 ^c	9.25 ^c	4.10 ^d	2.62 ^e	2.10 ^e	1.53 ^e	0.94	<.001	<.001	<.001
Perinatal	65.81 ^a	68.32 ^b	71.51 ^c	71.56 ^c	47.52 ^d	45.38 ^d	43.88 ^d	39.31 ^e	2.57	<.001	0.120	<.001
Fast (IIa + IIx + IIb)	9.06 ^a	10.07 ^a	9.99 ^a	9.90 ^a	32.28 ^b	37.10 ^c	39.91 ^c	43.77 ^d	2.82	<.001	<.001	0.003
I + α -cardiac	8.93 ^a	8.89 ^a	8.20 ^a	9.30 ^a	16.14 ^b	14.92 ^b	14.10 ^b	15.41 ^b	1.17	<.001	0.013	0.448
MyHC RNA level												
Embryonic	4.66 ^a	3.58 ^b	2.23 ^c	2.21 ^c	1.18 ^d	0.39 ^d	0.36 ^d	0.19 ^d	0.54	<.001	<.001	0.008
Perinatal	1.49 ^a	1.73 ^a	1.66 ^a	1.76 ^a	0.82 ^b	0.62 ^b	0.68 ^b	0.59 ^b	0.21	<.001	0.987	0.010
Fast (IIa + IIx + IIb)	0.11 ^a	0.17 ^a	0.16 ^a	0.18 ^a	0.84 ^b	1.05 ^c	1.21 ^c	1.80 ^d	0.17	<.001	<.001	<.001
IIa	0.12 ^a	0.16 ^a	0.14 ^a	0.17 ^a	0.76 ^b	1.09 ^c	1.27 ^c	1.87 ^d	0.22	<.001	<.001	<.001
IIx	0.12 ^a	0.18 ^a	0.18 ^a	0.19 ^a	0.92 ^b	0.95 ^b	1.14 ^b	1.73 ^c	0.27	<.001	0.002	0.006
I	0.53 ^a	0.60 ^a	0.47 ^a	0.61 ^a	1.02 ^b	0.99 ^b	0.98 ^b	1.60 ^c	0.20	<.001	<.001	<.001
α -cardiac	0.09 ^a	0.13 ^a	0.14 ^a	0.14 ^a	1.85 ^b	2.72 ^b	3.48 ^b	6.46 ^c	1.46	<.001	<.001	<.001
Glycogen, % wet muscle	3.38 ^a	4.03 ^a	4.68 ^b	4.99 ^b	9.55 ^c	10.52 ^d	12.36 ^e	12.11 ^e	0.58	<.001	<.001	0.033

Table 1. Biological phenotype characteristics of the 64 fetuses according to the eight experimental conditions.

The 64 fetuses used for this analysis were the same fetuses as those used in the proteome and transcriptome analyses. We considered eight conditions according to two developmental time points (d90 and d110) and four fetal genotypes (MS, LW, MSLW and LWMS). ¹Residual Standard Deviation (or Standard error of estimate). ²A, Age; G, Genotype; AxG, Age x Genotype interaction. ^{a,b,c,d,e} Means with different letters are significantly different (P < 0.05).

90 days of gestation					
Community	Nb. of nodes	Nb. of edges	Important node ¹	Phenotype link ²	Gene Ontology ³
1	18	26	FETUB	Glycogen & Emb. MyHC	glycolysis, myofibril, mitochondrion
2	19	55	PSMC5 (Iso 2)*	Glycogen & Fast MyHC	gluconeogenesis, actin filaments, mitochondrion
3	21	47	PDIA3 (Iso 1)	-	respiratory electron transport chain, cell redox homeostasis, mitochondrion
4	20	49	TNNT3 (Iso 2)*	Emb. MyHC	gluconeogenesis, muscle filament sliding and development, cell cycle checkpoint
5	16	33	ANXA2 (Iso 2)	-	creatine metabolic process, sarcolemma

Table 2. Network characteristics at 90 days of gestation.

Five sub-networks (communities) were obtained at 90 days of gestation. ¹Nodes with high betweenness. ²The phenotype correlations, with MyHC embryonic, adult fast and muscle glycogen content, were analyzed using spatial statistics tools. ³Additional File 6 shows all significantly enriched GO terms for each sub-network. *Using a permutation test, betweenness significantly high ($P < 0.05$).

110 days of gestation					
Community	Nb. of nodes	Nb. of edges	Important nodes ¹	Phenotype correlation ²	Gene Ontology ³
1	5	5	TUFM, DLD	-	tricarboxylic acid cycle, mitochondrion
2	24	65	ATP5A1*	Fast & Emb. MyHC	tricarboxylic acid cycle, glucose metabolic process, muscle filament sliding
3	19	46	GSN (Iso 1)	Glycogen, Fast & Emb. MyHC	gluconeogenesis, glycolysis, muscle filament sliding
4	13	19	PSMC5 (Iso 1)	Glycogen & Emb. MyHC	mRNA metabolic process, proteasome complex
5	15	39	LDB3 (Iso 2)	Glycogen & Fast MyHC	muscle development, lipid metabolism
6	10	22	PSMC5 (Iso 2)	Glycogen	gluconeogenesis, ATP catabolic/anabolic process, mitochondrion

Table 3. Network characteristics at 110 days of gestation.

Six sub-networks (communities) were obtained at 110 days of gestation. ¹Nodes with a high betweenness. ²The phenotype correlations, with MyHC embryonic, adult fast and muscle glycogen content, were analyzed using spatial statistics tools. ³Additional File 8 shows all significantly enriched GO terms for each sub-network. *Using permutation test, betweenness significantly high ($P < 0.05$).

Supplementary data

All the additional files are available at the url:

<https://www.dropbox.com/sh/zbscxnom8dyzkl9/AABW9MTiXhYTTY2V1qpsHZGPa?dl=0>

Additional File 1. An example of gel 2D electrophoresis .pdf file.

Additional File 2. Characteristics of the 120 spots identified by LS-MS/MS .xlsx file.

Additional File 3. Complete list of protein spots .xlsx file. Features of 113 spots (Protein name, Spot name, P-value, Discriminant analysis and Multivariate analysis). P-values were obtained by a F-type test testing the effect of gestational time point and/or fetal genotype.

Additional File 4. The largest extracted connected components at 90 and 110 days of gestation .pdf file. The d90-proteome network was composed of 94 nodes and 314 edges (density of 7.2%). The d110-proteome network was composed of 86 nodes and 313 edges (density of 8.6%).

Additional File 5. Features of the largest proteomic connected component at 90 days of gestation .xlsx file.

Additional File 6. Features of the largest proteomic connected component at 110 days of gestation .xlsx file. .

Additional File 7. Cluster description at 90 days of gestation .pdf file. Description of the five sub-networks obtained using spin-glass algorithm.

Additional File 8. Complete list of enriched Gene Ontology in the five sub-networks of the d90-proteome network .xlsx file.

Additional File 9. Cluster description at 110 days of gestation .pdf file. Description of the six sub-networks obtained using spin-glass algorithm.

Additional File 10. Complete list of enriched Gene Ontology in the five sub-networks of the d110-proteome network .xlsx file.

Additional File 11. Pearson's correlation between mRNA and protein expression levels .xlsx file.

Chapitre 4

Intégration de données hétérogènes

4.1 Intégration de données musculaires pour différentes espèces

4.1.1 Introduction

Dans cette partie, nous allons décrire les travaux effectués durant ma mobilité (août à décembre 2015) au CSIRO à Brisbane en Australie, dans le laboratoire de Brian Dalrymple. Cette équipe travaille notamment sur l'analyse de l'expression génique du muscle squelettique chez le bovin durant et après la période de gestation (Hudson *et al.*, 2009; Gu *et al.*, 2011; Guo *et al.*, 2015). L'objectif de cette mobilité fut d'analyser et d'intégrer des transcriptomes musculaires provenant de trois espèces d'intérêt agronomique différentes (porc, mouton et bovin), afin de déceler l'existence de mécanismes et fonctions biologiques communes ou différentes entre ces espèces durant le dernier tiers de gestation.

Les trois transcriptomes musculaires utilisés ont des conditions plus ou moins similaires : deux génotypes extrêmes pour le développement musculaire et plusieurs jours de gestation, notamment des points correspondant au dernier tiers de gestation. Différents ensembles de gènes impliqués dans des mêmes processus biologiques importants pour la maturité à la naissance ont été comparés entre ces trois espèces. Des ensembles de cinq à sept gènes impliqués dans le cycle cellulaire, la matrice extra-

cellulaire, les fibres musculaires rapides, le transport mitochondrial, le métabolisme lipidique et le métabolisme du glycogène ont été identifiés dans les trois tables de données. Globalement, les profils d'expression sont relativement similaires au cours de la gestation pour toutes les espèces et génotypes. En revanche, par des comparaisons entre génotypes pour chaque espèce, quelques différences significatives ont été observées. Des mécanismes impliqués dans le développement musculaire ou le métabolisme musculaire semblent avoir des effets différents en fonction du génotype et de l'espèce étudiée, et donc sur la maturité musculaire. Le développement pour une forte musculature et l'immaturité métabolique semblent être des caractéristiques relativement séparables, la maturité métabolique ne semblant pas être étroitement liée à des différences dans la division cellulaire. La sélection pour la maturité métabolique du muscle à la naissance est donc recommandée pour augmenter la survie des animaux.

Tous ces résultats sont décrits dans l'Article 3, présenté dans la partie suivante. Cet article est en préparation et sera soumis au journal *BMC Genomics* très bientôt.

4.1.2 Article 3 : Voillet et al., *En préparation*, 2016

Cette section correspond au manuscrit suivant :

- **V. Voillet**, M. San Cristobal, L. Liaubet, B.P. Dalrymple. Comparative transcriptomic analysis of fetal muscle development in cattle, sheep and pigs. *In preparation* (2016).

Comparative transcriptomic analysis of fetal muscle development in cattle, sheep and pigs

Valentin Voillet¹, Magali San Cristobal^{1,2}, Laurence Liaubet¹, Brian P. Dalrymple³

¹ Université de Toulouse, INRA, INPT, INP-ENVT, UMR1388, GenPhySE, F-31326 Castanet-Tolosan, France

² Université de Toulouse, INSA, UMR5219, Institut de Mathématiques, F-31077 Toulouse, France

³ CSIRO, Agriculture Flagship, QLD 4067 St-Lucia, Australia

E-mail: laurence.liaubet@toulouse.inra.fr

Abstract

Background Selection of animals for high production efficiency can have negative impacts on the survival rates of neonates. Maturity of the muscles at birth has been identified as a potential contributor to poor survival. Here we compare and contrast the expression of potential markers of muscle maturity during late stage gestation in high and low muscling breeds from cattle, sheep and pigs using a number of previously described datasets.

Results Gene sets of five to seven genes for estimation of cell cycle, extracellular matrix (ECM), fast twitch muscle fibre structural proteins, mitochondrial transport, lipid metabolism and glycogen metabolism applicable in all three datasets were identified. Overall the expression profiles of all sets of genes was very similar across gestation in all species and breeds examined. In a pairwise comparison of a cattle cross heterozygous for a myostatin mutation (Piedmontese-derived) and a high marbling (Wagyu-based) cattle cross some significant differences in the gene sets were observed such as cell cycle ($p < 0.01$), ECM ($p < 0.05$) and lipid metabolism at birth ($p < 0.05$). In contrast, in a comparison of a sheep breed (Texel) with a different mutation in myostatin and a relatively unselected sheep breed (Ujumqin) a difference in expression of the cell cycle genes ($p < 0.01$) and muscle structure fast subunits ($p < 0.05$) were observed in the latter stages of gestation. Although also exhibiting a big difference in muscling potential the two breeds of pigs, Large White and Meishan, did not show a significant difference in cell cycle gene expression. However, significant differences in expression of the ECM ($p < 0.01$), the fast twitch muscle fibre structural proteins ($p < 0.01$), mitochondrial ($p < 0.01$), lipid ($p < 0.01$) and glycogen metabolism ($p < 0.05$) gene sets was observed between the two pig breeds at birth.

Conclusions Different mechanisms for increased muscling appear to have different effects on muscle gene expression and thus different effects on the maturity of muscle at birth. High muscling and metabolic immaturity are separable characteristics and metabolic maturity does not appear to be tightly linked to differences in cell division. Selection for metabolic maturity of muscle at birth is recommended for increased survival of neonatal production animals.

Background

The efficient production of meat, with minimal impact on animal welfare and the environment, is a significant goal of most large scale production systems across all farm species. Welfare includes not only the conditions of growing and transporting animals and conditions at slaughter, but also the reduction of premature death of animals [1,2]. The latter is of particular importance around birth when the major

losses occur. Faster growing animals are generally more efficient as proportionally less energy is required for maintenance. However, the strong selection of farm animals, such as cattle, sheep and pigs, for increased progeny per pregnancy, increased muscle mass and reduced fat has been reported to affect neonatal survival to a greater or lesser extent in all three species [3–5].

Many breeds of cattle and some sheep breeds have been selected for high muscling due to mutations in the myostatin gene (*MSTN*). *MSTN* is a TGF-beta family member and a negative regulator of muscle mass [6]. Mutations in the *MSTN* gene cause a hypertrophic muscle phenotype via inhibition of both myoblast proliferation [7] and differentiation [8]. It has been reported in many species including cattle [9], sheep [10], mice [11] and human [12]. In cattle, mutations in *MSTN*, which lead to significantly increased muscle mass and less fat, tend to increase the weight of calves at birth leading to calving difficulties (dystocia) and to increase mortality of young animals [13–16]. In mice, it has also been shown that the pre-weaning mortality of a *MSTN* mutant was increased [17]. However, the reasons of a direct *MSTN* effect for increased neonatal mortality are still unclear.

Lamb survival is known to be a significant contributor to reproductive inefficiencies [18]. However there is no study analyzing the link between *MSTN* mutation and mortality at birth in sheep. Birth weight differences, which can be due to *MSTN* gene mutations such as in the Texel breed, are known to explain differences in the variation of lamb survival and is suggested as one of the main causes of death in sheep. Birth weight would have an effect on survival although it is likely to have a nonlinear effect [5]. Moreover, lambs exhibiting a greater metabolic maturity at birth seem to have a better rate of survival [19].

In pigs, increased muscle mass production is also associated with a higher rate of postnatal deaths [3]. Birth weight and maturity (amongst other factors), have been shown to be associated with variations in postnatal survival [3]. Unlike sheep and cattle, there is no production of pig breed with a *MSTN* mutation, but there are extreme breeds for muscle mass and intramuscular fat. For example, Large White (LW) is a European pig breed highly selected for muscular content and with a low intramuscular fat content, whereas the Chinese Meishan breed (MS) produces less muscular pigs with a higher intramuscular fat content than LW [20, 21]. These two breeds are also known to be extremely divergent for maturity at birth. MS piglets have a better survival rate than LW piglets although they are lighter at birth [3]. The high selection for growth in LW has led to a lower maturity of these piglets at birth, and a higher rate of death [22].

The objective of this study was to identify transcriptomic similarities for potential markers of maturity of muscular development in these three major mammalian meat producing livestock species and between genotypes within each species. Three gene expression datasets with sampling time points from mid to late fetal gestation to birth from three species were analyzed in the *longissimus dorsi* (LD) muscle.

Results and Discussion

Muscular development datasets from farm animal species

The cattle dataset comprised samples from a high muscling genotype (Piedmontese x Hereford (PH)), and a high marbling genotype (Wagyu x Hereford (WH)) [23] (Table 1). The Piedmontese breed has a *MSTN* inactivating mutation [9, 24]. The Wagyu breed, without a known mutation in *MSTN* affecting its expression or function, is less muscular and has a high intramuscular fat content [25].

Table 1 about here.

The sheep dataset also included samples from a high muscling genotype due to a *MSTN* mutation present in the Texel (TX) breed [10, 26], and a less muscular indigenous genotype (Ujumqin (UJ)) [27]. UJ has

also a higher intramuscular fat content than TX (Table 1). These two breeds, like the two cattle breeds, provide a natural model for studying muscle and fat development in sheep.

Two purebred pig breeds, Large White (LW) and Meishan (MS), were also studied (Table 1). Unlike the cattle and sheep, the high muscling in LW is not due to a *MSTN* mutation, but due to strong selection for high muscle content and low intramuscular fat content [20,21]. LW has a higher birth weight than MS but also a lower survival rate. Indeed, MS piglets are known to be more mature at birth even if they are lighter at birth than LW [3].

Time of gestation across species

Data from three gestational time points (60, 135 and 195 days) and birth (around 280 days of gestation) were available in the cattle muscle transcriptome dataset (Figure 1). In cattle, primary myogenesis occurs around 60 days post conception, followed by the secondary myogenesis around 90 days [28]. A third generation of fibres has been observed at about 40% of the gestation period [28]. The onset of functional differentiation of myofibres starts at around 195 days post-gestation. Preparation of the musculature for birth, the implications for locomotor and the metabolic independence occurs at the end of gestation during the maturation process [29]. Moreover, a study of Sudre *et al.* [30] showed that crucial development changes occur during the maturation process (the final trimester of gestation - around 210 days of gestation).

Figure 1 about here.

Five gestational time points (70, 85, 100, 120 and 135 days) were investigated for sheep (Figure 1). As also reviewed in [28], the skeletal muscle development in sheep is characterized by the formation of primary, secondary and tertiary myofibres beginning approximately 32, 38 and 62-76 days of gestation respectively. Birth is around 147 days of gestation. During the maturation process (from around 120 days of gestation), a major change (affecting oxidative metabolic processes) has also been demonstrated in sheep [31].

In the pig muscle dataset, only two time points, corresponding to the end of gestation (90 and 110 days of gestation), were available (birth is around 114 days of gestation) (Figure 1). Primary myogenesis occurs at 35 days of fetal life and the secondary myogenesis starts at 55 days. Unlike ruminants, the third generation of myogenesis does not take place during gestation, but just after birth [28]. Otherwise, like in the other species, a major change in the gene expression program has been shown during the maturation process [32].

To compare muscle development across species, the sampling points in the different datasets need to be standardized to a common scale. To achieve this sampling time points for each species were scaled between 0 and 1, with 1 corresponding to 100% gestation completed (*i.e.* birth) (Figure 1). The three datasets had some equivalent time points, but most were not exactly equivalent on this scale. However, primary, secondary (in all three species) and tertiary (in sheep and cattle) myogenesis and maturation (in all three species) occurred at equivalent times on this scale (Figure 1). In all species, the number of fibres is fixed before birth (around 80% of gestation) [28] and later increase of muscle mass occurs by hypertrophy of the existing fibres (Figure 1).

Refining the selection of genes and combining multiple genes

Co-expressed genes are more likely to be involved in the same or very closely related biological processes [33] and increasing sensitivity of detection of differential expression in muscle transcriptomic data. To estimate the impact of genotype during the maturation process on muscle mass and metabolic processes, a number of small and robust sets of genes was extracted. The chosen genes were annotated with the same biological process GO-term (except for genes for fast twitch muscle fibre structural protein genes chosen according to [29]), were expressed in all three datasets, and were co-expressed throughout development within each

genotype within each species, but they were not necessarily co-expressed between genotypes of the same species, or between species. As a result seven genes were identified for the cell cycle process gene set, six genes were identified for the extracellular matrix (ECM) gene set, and five genes were identified for the fast twitch muscle fibre structural subunits, the mitochondrial transport gene set, the lipid metabolism gene set and the glycogen metabolism gene set (Table 2). As already noted, combining data from multiple genes can reduce the impact of gene expression noise [34]. Previous analyses have shown that five genes is a good compromise between the number of genes and stability of the profile [34]. However, in our study, when it was possible, we chose to increase the number of genes to increase the power of the analysis.

Table 2 about here.

For all datasets, the individual gene expression values were standardized using z-scores to provide zero mean and unit variance across all values of each species and to enable combination of different genes. Therefore, for each genotype in each species, we averaged the gene expression values of the genes in each gene set.

The general development process during gestation is conserved across species

The expression patterns of each of the six gene sets was plotted across gestation for each breed of each species (Figure 2). All three species had very similar expression profiles across muscle development for each gene set, but the different gene sets had different expression profiles (Figure 2).

Figure 2 about here.

Expression of the cell cycle and ECM gene sets decreased during gestation, in particular during muscle maturation, whereas expression of the fast twitch muscle fibre structural proteins, mitochondrial transport, lipid metabolism and glycogen metabolism gene sets increased markedly towards the end of gestation (Figure 2). Figure 3 summarizes the fetal development in all farm species.

Figure 3 about here.

These changes in gene expression reflect a decrease in cell division as the number of muscle fibers is set by the end of the second trimester [28] and the muscle preparing itself metabolically to ensure the establishment of essential mechanisms for survival at birth and the postnatal environment where the muscle is important for locomotion and at least in pigs for thermogenesis [30–32]. It has already been shown in all three species that a major change in the gene expression program of skeletal muscle occurs during the last trimester of gestation. One of the most basic roles of the extracellular matrix is to provide a supportive scaffold for cells and tissue [35]. The relative contributions of the division of muscle or ECM stem cells to the cell cycle gene activity is unknown, but is likely to vary across fetal development. A recent analysis of the cattle dataset used here suggests that the cell division and proliferation of fibroblasts is a significant contributor to the cell cycle signal [34]. The decline in expression is consistent with a decrease in cell division with age as the rate of growth declines and the muscle program switches from hyperplasia (more cells) to hypertrophy (larger cells) [28].

The increase in expression of the fast twitch fibre muscle structural protein genes is consistent with the succession from embryonic and fetal myosins, through slow twitch to fast twitch fibre structural protein genes during prenatal development. The common increase in mitochondrial transport, and lipid and

glycogen metabolism is in agreement with contractile and metabolic maturation cattle [30], sheep [31] and pigs [32]. Oxidative metabolism tends to increase during gestation in all the three farm species and represents the principal source of energy during fetal life and probably reflects new functional demands postnatally [28].

Differences between breeds of different species

The design of the analysis also allowed us to compare expression between the high and low muscling breeds within species, especially during the last stages of gestation (Figure 2). In pigs, no significant differences for the cell cycle gene set were found between MS and LW at the end of gestation (90 days and 110 days of gestation), whereas significant differences between breeds were observed in sheep ($p < 0.01$ at d120 and d135) and in cattle ($p < 0.01$ at d280) (Figure 2A and Table 3). The major determinants of skeletal muscle mass are the number and the size of muscle fibers [28]. These two factors are controlled during the gestation by many events such as myoblast proliferation, myotube formation and myofiber maturation. Cattle and sheep are evolutionarily closely related [36] and hypertrophic animals with *MSTN* gene mutations were used in this study. Previous analysis of the sheep data set [27] showed that the myostatin mutation in TX appeared to disrupt the gene expression profile in prenatal skeletal muscle, in particular changing pivotal signaling pathways governing muscle development, explaining this difference in myofiber phenotypes between TX and UJ sheep. They demonstrated that before the mid-fetal stage the proliferation of muscle fibers was faster in TX than in UJ, but that the reverse was true in late-fetal stages [27]. In a network analysis, these results were also confirmed by a higher connection of genes involved in cell cycle in UJ than in TX at the late-stage in gene co-expression network [37]. Our analysis of the dataset seems to be consistent with these results. However, corresponding differences in the ECM genes were not observed in sheep and cattle (Figure 2B), raising the question of the source of the differences in the cell cycle signal, especially in sheep; muscle stem cells or fibroblasts, or another cell type again?

Table 3 about here.

On the other hand, significant differences between breeds for the ECM gene set were found in pigs ($p < 0.01$ at d90 and d110) and in cattle ($p < 0.05$ at birth) (Figure 2B and Table 3). The development and growth of skeletal muscle is a complex process, including not only the muscle contractile cells, but also the expansion of the ECM [38]. ECM genes are involved in the regulation of many cellular events during myogenesis. ECM is a dynamic mixture of structural and functional macromolecules and plays a crucial role in tissue and organ morphogenesis and the maintenance of tissue structure and function [38]. In pigs, the effects of ECM genes on muscle development remain poorly understood, in particular toward the end of gestation. Ma *et al.* [39] have compared three genotypes including an obese- (Tongcheng) and a lean-type (Landrace) to observe the role of ECM genes in muscle development and the formation of phenotypic variation in pigs. They observed that the expression of many ECM genes was significantly higher in the prenatal than postnatal periods and that expression of ECM genes was higher during gestation in the lean-type breed than the obese-type breed, as we observed here in LW. In our data set, several ECM genes more highly expressed in Landrace than Tongcheng, were also more highly expressed in LW than in MS. In pigs, the higher expression of ECM gene set in LW than in MS may reflect an extended muscle development period leading to a possible state of immaturity of muscle at birth as expected in LW. In contrast the *MSTN* sheep have evidence for an accelerated myogenesis compared to non-*MSTN* sheep. Perhaps due to milder effect of the genotypes little difference in either cell cycle or ECM gene expression was observed between the two cattle crosses.

We only observed significant differences in expression of the mitochondrial transport ($p < 0.01$ at d90 and d110), lipid metabolism ($p < 0.05$ at d90 and $p < 0.01$ at d110) and glycogen metabolism ($p < 0.05$ at d110) gene sets between pig breeds (Figure 2D, 2E, 2F and Table 3). It appears that the extended

muscle development late in gestation, which leads to more muscular animals by a different mechanism from *MSTN* mutant (increased myogenesis early in gestation), results in delayed metabolic maturity of the muscle. If we assume that all species evolve *at the same time*, the differences in timing may also have a larger impact on muscle maturity at birth in pigs due to the generally less mature muscle at birth of pigs than in cattle and sheep, perhaps as a consequence of the much later tertiary myogenesis in pigs. In pigs, a greater physiological maturity at birth is responsible for higher survival [40,41]. As already highlighted in [32], the selection on growth traits in LW generated a high divergence during the maturation process between MS and LW. In large mammals, the major events of contractile and metabolic differentiation occur during the last third of gestation and are fully achieved just after birth [28]. In these species, foetal life represents a primordial step for muscle maturation and could explain why there is no difference between breeds in sheep and cattle. Moreover, brown adipose tissue is absent in pigs at birth [42], but present in sheep and cattle [43]. Therefore, in pigs, muscle plays also a key role in thermoregulation via muscle glycogen accumulation [44]. In other species, there is a high contribution of the brown adipose tissue to the energy balance changes at birth and is in relationship with the stage of maturity at birth [43]. In sheep, we observed a significant difference in expression of the lipid metabolism gene set between the two breed ($p < 0.05$ at d135), and it is in agreement with a study which has already shown that lipid related genes in muscle were highly connected during the mid-stage in TX but more highly connected in UJ in later stages [37]. In cattle, we also observed a significant difference in expression of lipid metabolism gene set at birth between the two breed ($p < 0.05$). In our study, it seems that there is minor effect of *MSTN* gene mutation on leanness during the maturation process in cattle and sheep. Moreover, as already highlighted, ECM genes play important roles in the regulation of muscle cell proliferation and differentiation [45]. The extracellular environment regulate the differentiation of stem cells and satellites cells [46], providing, via cellular signalling, myogenic differentiation which increases muscle mass, but also an increase of intramuscular adipocytes. Our results were totally in accordance with these assumption with a significant differences between breeds in expression of the lipid metabolism and ECM gene sets in pigs and cattle. The function of ECM has an important influence in muscle development but also in future meat quality by affecting the intramuscular fat deposition during the end of gestation. The maturation of skeletal muscle metabolic and contractile properties are indicative of piglet maturity at the time of birth.

Conclusion

Selection for increased muscle mass and muscle maturity seem to be separable processes in at least some circumstances. Cattle and sheep containing at least one mutant myostatin allele do not differ from the myostatin wild type animals for gene expression profiles for muscle structure fast subunits, mitochondrial transport and glycogen metabolism. A long term metabolic difference between the pig lines exists, but is not really evident in sheep and cattle (despite the fact that the cattle and sheep examples are contrasting normal and myostatin mutations). The relative timing of increased activity prenatally leading to increased muscling postnatally and of the different stages of muscle development may have a profound impact on the maturity of muscle at birth. Post natal mortality is a complex process potentially involving a number of different processes. Whilst muscle metabolic maturity appears to be more important in pigs this may reflect the datasets analyzed as well as developmental differences between the species. However, there is no evidence for a large effect of the *MSTN* mutation on metabolic maturity of cattle and sheep muscle. Rather the impact of mutant alleles of *MSTN* on post-natal mortality may be due to the size of the fetus or other impacts of *MSTN* on the physiological state of the new born. These results provide valuable information about possible mechanisms determining the phenotypic differences on growth and meat quality between the genetic types studied, mainly related to the development and function of the extracellular matrix and also to some metabolic processes as glycogen and lipid metabolism. In addition, the highlighted genes across all species in this study confer a good reliability to name them biomarkers in future analysis.

Materials and Methods

Gene expression datasets

All animal use (pig, sheep and cattle) were respectively approved by European Union legislation (directive 86/609/EEC) and French legislation of région Midi-Pyrénées in France; by the Biological Studies Animal Care and Use Committee, Shanxi Province, Peoples Republic of China and by the Industry & Investment New South Wales (NSW), Orange Agriculture Institute Animal Ethics Committee, Commonwealth Scientific and Industrial Organisation (CSIRO) Rockhampton Animal Experimentation Ethics Committee, and the Department of Agriculture and Food, Western Australia (WA) Animal Ethics Committee.

Pig gene expression dataset

The generation of gene expression data from the *Sus scrofa* Large White (LW) and Meishan (MS) genotypes have been previously described in [32]. Briefly, LD mRNA were acquired at two development stages (90 and 110 days of gestation) from four fetal genotypes. These genotypes consisted in two extreme breeds for mortality at birth (MS and LW) and two crosses (MSLW from MS sows and LWMS from LW sows). MS and LW sows were inseminated with mixed semen so that each litter was composed of purebred and crossbred fetuses. For each genotype, gene expression value were collected for to five and nine individuals (with $n_{d90MS} = 8$, $n_{d90LW} = 8$, $n_{d110MS} = 7$ and $n_{d110LW} = 9$). The Agilent microarray GPL16524 (Agilent Technologies) was used in this experiment. The raw data were normalized using the quantile method [47], and the data were transformed to a base-2 logarithm for further statistical analysis [32]. Only purebreds MS and LW were analyzed during this analysis.

The gene expression data is available in the Gene Expression Omnibus (GEO) NCBI platform through the accession number GSE56301.

Cattle gene expression dataset

The generation of gene expression data from the *Bos taurus* Wagyu x Hereford (WH) and Piedmontese x Hereford (PH) genotypes have already been described in [48, 49]. In brief, LD biopsy samples were collected from nine developmental stages (pre and postnatal) and a post slaughter sample: 60, 135 and 195 days post conception, at birth (around 280 days) and 3, 7, 12, 20, 25 and 30 months of age. Only prenatal (d60, d135, d195 and d280) and birth samples were analyzed during this analysis. From 60 days to birth, samples of each cross were recovered by cesarian from different fetuses, afterward samples were obtained by biopsy from the same individuals. For each stage, gene expression values were acquired for three or four individuals on the Agilent Bovine microarray platform (Agilent Technologies) [50] (with $n_{d195WH} = 3$, $n_{d195PH} = 3$, $n_{d280WH} = 3$ and $n_{d280PH} = 3$). The data were normalized using a linear mixed model as previously described in [34, 50].

The gene expression data is available through the GEO accession number GSE44030.

Sheep gene expression dataset

The generation of gene expression data from the *Ovis aries* Ujumquin (UJ) and Texel (TX) genotypes have been previously described in detail [27]. Fetuses were collected by caesarean from three pregnant ewes at five developmental stages: 70, 85, 100, 120 and 135 days post conception (birth around 150 days). Each development stage for each breed contained three individuals except the d100 group in Texel breed with four individuals. Gene expression values of the LD were acquired on the Agilent Sheep Gene Expression Microarray (Agilent Technologies). The data were normalized using the quantile method [47], and transformed to a base-2 logarithm [27].

The gene expression data is available through the GEO accession number GSE23563.

Statistical analysis

All analyses were performed using R computing environment [51]. As previously described in [34], z-scores were performed to minimize the impact of differences in levels of expression and dynamic range of expression of genes on combining the expression data from two or more genes. Z-score corresponds to the difference from the mean of each measurement divided by the relevant standard deviation. Therefore, pig gene expression dataset were rescaled to a mean of 0 across the whole set including all four fetal genotypes (MS, LW, MSLW and LWMS) for all developmental stages, and cattle gene expression dataset were rescaled to 0 across the whole set including both WH and PH genotypes for only prenatal time points. Sheep gene expression dataset were rescaled to 0 throughout development and including both UJ and TX genotypes for all only prenatal development stages.

Instead of using a linear model as already shown in [32], we chose to perform Student's t-test to compare the differences in the expression of a combination of genes (average of genes that belong to the same biological function) between genotypes at the same time points in each species separately. We only analyzed time points included in the maturation process at a time. Indeed, we did not include all time points because we were focused only on the end of gestation. The threshold for significance was $p < 0.05$. Regarding the gene selection, as explained in the Results and Discussion section, we chose genes which were annotated with the same biological process GO-term (except for genes for fast twitch muscle fibre structural protein genes chosen according to [29]), expressed in all three datasets, and co-expressed throughout development within each genotype within each species, but they were not necessarily co-expressed between genotypes of the same species, or between species.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

BPD supervised the study. MSC and LL contributed to important ideas in the study. VV carried out all statistical analysis. VV and BPD wrote the manuscript incorporating comments from all other authors. All authors read and approved the final manuscript.

Acknowledgements

VV is a PhD fellow supported by the INRA GA (Génétique Animale), the INRA PHASE (Physiologie Animale et Systèmes d'Élevage) and the region Midi-Pyrénées. We also would like to thank Kim-Anh Lê Cao for her really useful comments.

References

1. Broom, D.M.: Animal welfare: concepts and measurement. *Journal of Animal Science* **66**, 4167–4175 (1991)
2. Hewson, C.J.: What is animal welfare? common definitions and their practical consequences. *The Canadian Veterinary Journal* **44**(6), 496–499 (2003)
3. Canario, L., Cantoni, E., Le Bihan, E., Caritez, J.C., Billon, Y., Bidanel, J.P., Foulley, J.L.: Between-breed variability of stillbirth and its relationship with sow and piglet characteristics. *Journal of Animal Science* **84**(12), 3185–3196 (2006)

4. Casas, E., Bennett, G.L., Smith, T.P.L., Cundiff, L.V.: Association of myostatin on early calf mortality, growth, and carcass composition traits in crossbred cattle. *Journal of Animal Science* **82**, 2913–2918 (2004)
5. Everett-Hincks, J.M., Mathias-Davis, H.C., Greer, G.J., Auvray, B.A., Dodds, K.G.: Genetic parameters for lamb birth weight, survival and death risk traits. *Journal of Animal Science* **92**, 2885–2895 (2014)
6. McPherron, A., Lawler, A., Lee, S.J.: Regulation of skeletal muscle mass in mice by a new TGF- β superfamily member. *Nature* **387**, 83–90 (1997)
7. Joulai, D., Bernardi, H., Garandel, V., Rabenoelina, F., Vernus, B., Cabello, G.: Mechanisms involved in the inhibition of myoblast proliferation and differentiation by myostatin. *Experimental Cell Research* **286**, 263–275 (2003)
8. Rios, R., Carneiro, I., Arce, V., Devesa, J.: Myostatin is an inhibitor of myogenic differentiation. *American Journal of Physiology* **282**, 993–999 (2003)
9. Kambadur, R., Sharma, M., Smith, T.P.L., Bass, J.L.: Mutations in myostatin (GDF8) in double-muscled belgian blue and piedmontese cattle. *Genome Research* **7**, 910–915 (1997)
10. Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibe, B., Bouix, J., Caiment, F., Elsen, J.M., Eychenne, F., Larzul, C., Laville, E., Meish, F., Milenkovic, D., Tobin, J., Charlier, C., Georges, M.: A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nature Genetics* **38**(7), 813–818 (2006)
11. Lee, S.J., McPherron, A.: Regulation of myostatin activity and muscle growth. *Proceedings of the National Academy of Sciences of the United States of America* **98**(16), 9306–9311 (2001)
12. Schuelke, M., Wagner, K.R., Stolz, L.E., Hübner, C., Riebel, T., Kömen, W., Braun, T., Tobin, J.F., Lee, S.J.: Myostatin mutation associated with gross muscle hypertrophy in a child. *The New England Journal of Medicine* **350**, 2682–2688 (2004)
13. Morris, C.A., Bennett, G.L., Baker, R.L., Carter, A.H.: Birth weight, dystocia and calf mortality in some new zealand beef herds. *Journal of Animal Science* **62**, 327–343 (1986)
14. Berger, P.J., Cubas, A.C., Koehler, K.J., Healey, M.H.: Factors affecting dystocia and early calf mortality in angus cows and heifers. *Journal of Animal Science* **70**, 1775–1786 (1992)
15. Johanson, J.M., Berger, P.J., Tsuruta, S., Misztal, I.: A bayesian threshold-linear model evaluation of perinatal mortality, dystocia, birth weight, and gestation length in a Holstein herd. *Journal of Dairy Science* **94**, 450–460 (2011)
16. Casas, E., Thallman, R.M., Cundiff, L.V.: Birth and weaning traits in crossbred cattle from Hereford, Angus, Brahman, Boran, Tuli, and Belgian Blue sires. *Journal of Animal Science* **89**, 979–987 (2011)
17. Li, Z.F., Shelton, G.D., Engval, E.: Elimination of myostatin does not combat muscular dystrophy in dy mice but increases postnatal lethality. *The American Journal of Pathology* **166**(2), 491–497 (2005)
18. Hinch, G.N., Brien, F.: Lamb survival in australian flocks: a review. *Animal Production Science* **54**, 656–666 (2014)

19. Miller, D.R., Blache, D., Jackson, R.B., Downie, E.F., Roche, J.R.: Metabolic maturity at birth and neonate lamb survival: association among maternal factors, litter size, lamb birth weight, and plasma metabolic and endocrine factors on survival and behavior. *Journal of Animal Science* **88**, 581–592 (2010)
20. Bidanel, J.P., Caritez, J.C., Legault, C.: Ten years of experiments with chinese pigs in france. *Pig News Info* **11**(1), 345–348 (1990)
21. White, B.R., Lan, Y.H., McKeith, F.K., Novakofski, J., Wheeler, M.B., McLaren, D.G.: Growth and body composition of meishan and yorkshire barrows and gilts. *Journal of Animal Science* **73**, 738–749 (1995)
22. Canario, L., Père, M.C., Tribout, T., Thomas, F., David, C., Herpin, P., Bidanel, J.P., Le Dividich, J., Gogué, J.: Estimation of genetic trends from 1977 to 1998 of body composition and physiological state of Large White pigs at birth. *Animal* **1**, 1409–1413 (2007)
23. Hudson, N.J., Reverter, A., Greenwood, P.L., Guo, B., Cafe, L.M., Dalrymple, B.P.: Longitudinal muscle gene expression patterns associated with differential intramuscular fat in cattle. *Animal* **9**(4), 650–659 (2014)
24. Grobet, L., Poncelet, D., Royo, L.J., Brouwers, B., Pirottin, D., Michaux, C., Menissier, F., Zanotti, M., Dunner, S., Georges, M.: Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double-muscling in cattle. *Mammalian Genome* **9**, 210–213 (1998)
25. Zembayashi, M., Nishimura, K., Lunt, D.K., Simth, S.B.: Effect of breed type and sex on the fatty acid composition of subcutaneous and intramuscular lipids of finishing steers and heifers. *Journal of Animal Science* **73**, 3325–3332 (1995)
26. Johnson, P.L., Dodds, K.G., Bain, W.E., Greer, G.J., McLean, N.J., McLaren, R.J., Galloway, S.M., van Stijn, T.C., McEwan, J.C.: Investigations into the GDF8 g+6723G-A polymorphism in new zealand texel sheep. *Journal of Animal Science* **87**(6), 1856–1864 (2009)
27. Ren, H., Li, L., Su, H., Xu, L., Wei, C., Zhang, L., Li, H., Liu, W., Du, L.: Histological and transcriptome-wide level characteristics of fetal myofiber hyperplasia during the second half of gestation in texel and ujumqin sheep. *BMC Genomics* **12**, 411 (2011)
28. Picard, B., Lefaucheur, L., Berri, C., Duclos, M.J.: Muscle fibre ontogenesis in farm animal species. *Reproduction Nutrition Development* **42**, 415–431 (2002)
29. Hudson, N.J., Lyons, R.E., Reverter, A., Greenwood, P.L., Dalrymple, B.P.: Inferring the in vivo cellular program of developing bovine skeletal muscle from expression data. *Gene Expression Patterns* **13**, 109–125 (2013)
30. Sudre, K., Leroux, C., Piétu, G., Cassar-Malek, I., Petit, E., Listrat, A., Auffray, C., Picard, B., Martin, P., Hocquette, J.F.: Transcriptome analysis of two bovine muscles during ontogenesis. *J Biochem* **133**(6), 745–756 (2003)
31. Byrne, K., Vuocolo, T., Gondro, C., White, J.D., Cocket, N.E., Hadfield, T., Bidwell, C.A., Waddell, J.N., Tellam, R.L.: A gene network switch enhances the oxidative capacity of ovine skeletal muscle during late fetal development. *BMC Genomics* **11**, 378 (2010)
32. Voillet, V., San Cristobal, M., Lippi, Y., Martin, P.G.P., Iannuccelli, N., Lascor, C., Vignoles, F., Billon, Y., Canario, L., Liaubet, L.: Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity. *BMC Genomics* **15**, 797 (2014)

33. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein: Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**(25), 14863–14868 (1998)
34. Guo, B., Greenwood, P.L., Cafe, L.M., Zhou, G., Zhang, W., Dalrymple, B.P.: Transcriptome analysis of cattle muscle identifies potential markers for skeletal muscle growth rate and major cell types. *BMC Genomics* **16**, 177 (2015)
35. Frantz, C., Stewart, K.M., Weaver, V.M.: The extracellular matrix at a glance. *Journal of Cell Science* **123**, 4195–4200 (2010)
36. Kijas, J.W., Menzies, M., Ingham, A.: Sequence diversity and rates of molecular evolution between cattle and sheep genes. *Anim Genet* **37**, 171–174 (2006)
37. Xu, L., Zhao, F., Ren, H., Li, L., Lu, J., Liu, J., Zhang, S., Liu, G.E., Song, J., Zhang, L., Wei, C., Du, L.: Co-expression analysis of fetal weight-related genes in ovine skeletal muscle during mid and late fetal development stages. *International journal of biological sciences* **10**, 1039–1050 (2014)
38. Velleman, S.G.: Extracellular matrix regulation of skeletal muscle formation. *Journal of Animal Science* **90**, 936–941 (2012)
39. Ma, X., Tang, Z., Wang, N., Zhao, S., Wang, R., Tan, L., Mu, Y., Li, K.: Identification of extracellular matrix and cell adhesion molecule genes associated with muscle development in pigs. *DNA Cell Biology* **30**(7), 469–479 (2011)
40. Leenhouwers, J.I., Knol, E.F., de Groot, P.N., Vos, H., van der Lende, T.: Fetal development in the pig in relation to genetic merit for piglet survival. *Journal of Animal Science* **80**(7), 1759–1770 (2002)
41. Leenhouwers, J.I., F, K.E., van der Lende, T.: Differences in late prenatal development as an explanation for genetic differences in piglet survival. *Journal of Animal Science* **78**, 57–62 (2002)
42. Trayhurn, P., Temple, N.J., van Aerde, J.: Evidence from immunoblotting studies on uncoupling protein that brown adipose tissue is not present in the domestic pig. *Canada Journal of Physiology and Pharmacology* **67**, 1480–1485 (1989)
43. Symonds, M.E., Popre, M., Budge, H.: The ontogeny of brown adipose tissue. *Annual Review of Nutrition* **35**, 295–320 (2015)
44. Herpin, P., Damon, M., Le Dividich, J.: Development of thermoregulation and neonatal survival in pigs. *Livestock Production Science* **78**(1), 25–45 (2002)
45. Clause, K.C., Barke, T.H.: Extracellular matrix signaling in morphogenesis and repair. *Current Opinion in Biotechnology* **24**(5), 830–833 (2013)
46. Watt, F., Huck, W.T.S.: Role of the extracellular matrix in regulating stem cell fate. *Nature Reviews Molecular Cell Biology* **14**, 467–473 (2013)
47. Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**(2), 185–193 (2003)
48. Cafe, L.M., Hennessy, D.W., Hearnshaw, H., Morris, S.G., Greenwood, P.L.: Influences of nutrition during pregnancy and lactation on birth weights and growth to weaning of calves sired by Piedmontese and Wagyu bulls. *Australian Journal of Experimental Agriculture* **46**(2), 245–255 (2006)

49. Lehnert, S.A., Reverter, A., Byrne, K.A., Wang, Y., Nattrass, G.S., Hudson, N.J., Greenwood, P.L.: Gene expression studies of developing bovine longissimus muscle from two different beef cattle breeds. *BMC Developmental Biology* **7**, 95 (2007)
50. Reverter, A., Barris, W., McWilliam, S., Byrne, K.A., Wang, Y.H., Tan, S.H., Hudson, N.J., Dalrymple, B.P.: Validation of alternative methods of data normalization in gene co-expression studies. *Bioinformatics* **21**(7), 1112–1120 (2005)
51. Team, R.C.: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2013). R Foundation for Statistical Computing

Figures

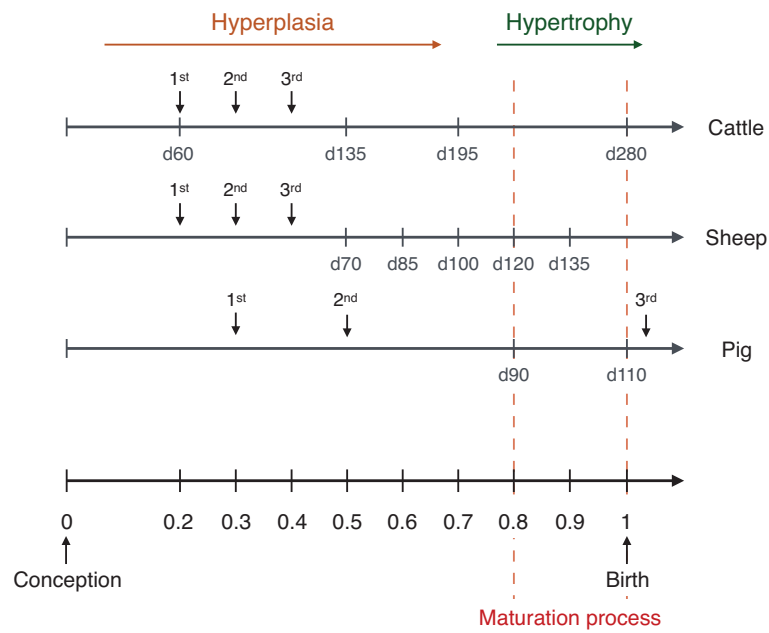


Figure 1. Overview of gestation for each species. A scale between 0 and 1 with 1 corresponding to birth was used. Time points with gene expression data used in the datasets analysed are shown for each species. The timing of primary (1st), secondary (2nd) and tertiary (3rd) myogenesis is indicated for each species.

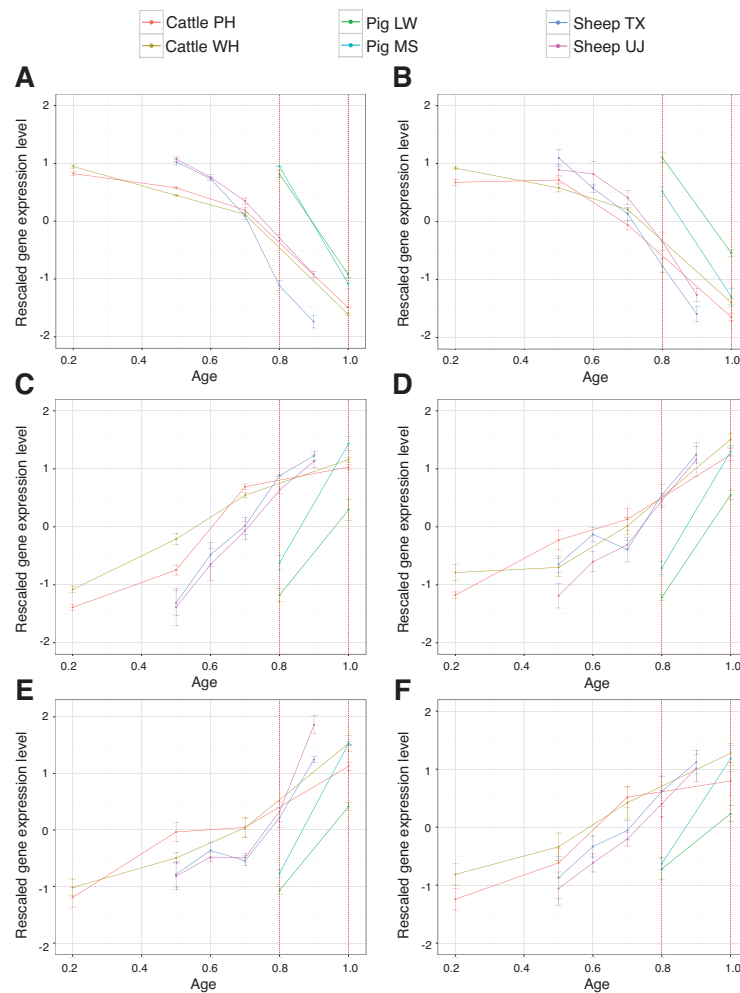


Figure 2. The expression profiles of six gene sets through development in cattle, sheep and pig. The expression levels of gene sets are the average z-scores of the genes in each gene set: after a z-transformation, genes were averaged across GO-ontology for each genotype. The maturation process is represented according to red lines. **(A)** Cell cycle gene set. **(B)** ECM gene set. **(C)** Fast twitch muscle fibre structural protein gene set. **(D)** Mitochondrial transport gene set. **(E)** Lipid metabolism gene set. **(F)** Glycogen metabolism gene set.

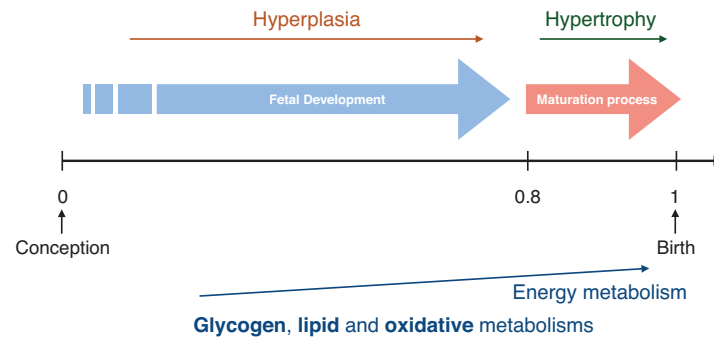


Figure 3. Global fetal development in all three farm species. A scale between 0 and 1 with 1 corresponding to birth was used. The maturation process occurs during the last trimester of gestation. A change of gene expression is observed with a decrease in cell division as the number of muscle fibers is set by the end of the second trimester and a increase of metabolic process to ensure the establishment of essential mechanisms for survival at birth.

Tables

		Muscular content	Intramuscular fat content	Birth weight	Postnatal survival
Pig	Large White	+++	+	+++	+
	Meishan	+	+++	+	+++
Cattle	Piedmontese x Hereford	+++	+	+++	+
	Wagyu x Hereford	+	+++	+	+++
Sheep	Texel	+++	+	+++	+
	Ujumqin	+	+++	+	+++

Table 1. Muscle genotype characteristics for each specie (pigs, cattle and sheep)

Gene Set	Biological process	Genes
Cell cycle ¹	Cell cycle process	<i>CDC6, CDC20, CDCA3, CCNB1, CCNB2, KIF20A, CKAP2</i>
ECM ¹	ECM organization	<i>BGN, COL5A2, TGFB2, SERPINH1, SDC3, SH3PXD2B</i>
Muscle structure fast subunits ²	Fast twitch	<i>TMOD4, MYOZ1, MYH1, CKM, ENO3</i>
Mitochondrial transport ¹	Respiratory electron transport chain	<i>ATP5B, PPARGC1A, COX5B, UCP3, NDUFV2</i>
Lipid metabolism ¹	Cellular lipid metabolic process	<i>ACADS, CRAT, SLC25A20, CPT1B, MCEE</i>
Glycogen metabolism ¹	Glycogen metabolic process	<i>GAA, PYGM, PPP1R1A, PPP1R3A, GSK3A</i>

Table 2. Gene sets identified to compare species and breeds ¹GO term were the source of these gene sets. ²This gene set was generated using a previous work [29].

		Cell cycle process	ECM organization	Muscle structure fast subunits	Mitochondrial transport	Lipid metabolism	Glycogen metabolism
Pig	d90	No diff.*	< 0.01	< 0.01	< 0.01	< 0.05	No diff.
	d110	No diff.	< 0.01	< 0.01	< 0.01	< 0.01	< 0.05
Sheep	d120	< 0.01	No diff.	< 0.05	No diff.	No diff.	No diff.
	d135	< 0.01	No diff.	No diff.	No diff.	< 0.05	No diff.
Cattle	d280	< 0.01	< 0.05	No diff.	No diff.	< 0.05	No diff.

Table 3. P-values obtained between breeds for each gene sets in each specie (pigs, cattle and sheep) *corresponding to no significant differences between breeds.

4.2 Intégration de données issues de différents tissus pour une espèce

4.2.1 Introduction

Dans cette partie, nous allons décrire l'intégration de données de même type, c'est-à-dire plusieurs transcriptomes (provenant de plusieurs tissus) prélevés à partir des mêmes individus. Dans le projet Porcinet, nous disposons de données transcriptomiques musculaires, mais également de données transcriptomiques issues d'autres tissus comme les surrénales, le foie et le sang du cordon. Toutes ces informations ont été obtenues sur un même groupe d'individus. Cependant, il est important de noter que quelques données omiques manquent pour certains individus, engendrant ainsi des problèmes de lignes manquantes lors de l'intégration de données, notamment pour l'utilisation de stratégies de statistiques multidimensionnelles. Les méthodes statistiques multidimensionnelles sont utilisées pour intégrer différents tableaux de données obtenues à partir des mêmes individus. Je vais, dans un premier temps, brièvement décrire quelques méthodes statistiques multidimensionnelles qui ont été utilisées durant la thèse, et notamment la méthode d'analyse factorielle multiple (AFM), permettant d'intégrer plus de deux tableaux de données (section 4.2.2).

Nous avons proposé une extension à l'AFM visant à contourner les lignes manquantes lors de l'estimation des composantes de l'AFM. Notre stratégie, appelée MI-MFA (pour *multiple imputation - MFA*), consiste en l'application du principe de l'imputation multiple dans le cadre de l'AFM. Brièvement, les lignes manquantes pour un groupe de variables donné sont imputées par des valeurs plausibles de façon aléatoire et une AFM est effectuée sur ces données imputées. Cette opération est réalisée m fois afin d'obtenir m configurations AFM. Par la suite, l'utilisation de la méthode STATIS (Structuration des Tableaux A Trois Indices de la Statistique) permet de déterminer un ensemble consensus entre ces m configurations. Un article décrivant cette extension a été accepté dans le journal *BMC Bioinformatics* en septembre 2016 et est présenté dans la section 4.2.3. Pour finir, la MI-MFA sera illustrée avec les données Porcinet.

4.2.2 Aperçu de quelques méthodes d'analyse multidimensionnelle

Afin d'analyser conjointement plusieurs tableaux de données, les méthodes statistiques de projections multidimensionnelles peuvent être utilisées. Ces stratégies, généralement choisies comme première analyse statistique exploratoire, ont des objectifs plus ou moins différents en fonction de la question biologique posée, comme structurer et résumer les données ou encore comparer et extraire des groupes de variables. L'objectif peut être de dégager les caractéristiques principales issues des données comme des effets biologiques ou des effets techniques non-désirés. Plusieurs méthodes ont été développées et proposées dans la littérature, dont l'analyse canonique des corrélations (ACC) (ou *canonical correlation analysis* (CCA)) (Hotelling, 1936), la régression des moindres carrés partiels (ou *partial least square* (PLS)) (Wold, 1966; Wold *et al.*, 2001) ou encore l'analyse factorielle multiple (AFM) (ou *multiple factor analysis* (MFA)) (Escofier & Pagès, 1988-1998; Pagès, 2002).

Dans cette partie, j'ai choisi de décrire les méthodes ACC et PLS car elles ont été appliquées (dans leurs versions *sparse*) à nos données lors de la sélection de spots protéiques à identifier (voir Voillet *et al.* (2016b), section 3.3). Ces méthodes ont pour objectif d'explorer les relations entre deux groupes de variables quantitatives observées sur un même groupe d'individus. La description de ces deux stratégies est largement inspirée du document *Multivariate projection methodologies for the exploration of large biological data sets* écrit par Lê Cao (2014). La méthode AFM sera également présentée ; c'est une méthode factorielle adaptée à l'étude de tableaux (deux ou plus) dans lesquels un ensemble d'individus est décrit par plusieurs ensembles de variables. Son objectif, proche de celui de l'analyse en composantes principales, est donc de représenter un ensemble de groupes de variables observées sur un ensemble d'individus, avec l'aide de variables latentes (non observées), afin de réduire le nombre de dimensions (et donc le nombre de variables). Une amélioration de cette méthode a également été proposée afin de contourner les problèmes des individus manquants lors de l'estimation des composantes de l'AFM (voir Voillet *et al.* (2016a), section 4.2.3.2). Toutes les stratégies d'analyse multidimensionnelle ne sont pas exhaustivement présentées ici. Par exemple, l'analyse de co-inertie (ACO), déjà utilisée dans le cadre de données omiques (Culhane *et al.*, 2003; Fagan *et al.*, 2007), et son extension (analyse de co-inertie multiple (ACOM)) capable d'intégrer plus

de deux tables de données simultanément (Meng *et al.*, 2014), ne sont pas décrites dans cette partie. L'ACO est une méthode itérative dans laquelle, à chaque étape, on cherche un couple de vecteurs (correspondant à chacun des ensembles de données) vérifiant certaines contraintes d'orthogonalités et en maximisant un critère de covariance (c'est une méthode assez proche de la PLS). L'ACOM généralise à plus de deux tableaux l'ACO. Contrairement à l'ACC et la PLS, ces deux méthodes ne proposent pas de version *sparse* permettant la sélection de variables. C'est pour cette raison que nous ne les avons pas retenues dans notre étude.

4.2.2.1 Rappels sur l'analyse en composantes principales (ACP)

En amont de la description de l'ACC, de la PLS et de l'AFM, les bases de l'analyse en composante principale (ACP) vont être succinctement présentées. L'ACP est une méthode exploratoire très souvent utilisée comme première analyse des données. Son objectif est de réduire le nombre de variables et d'identifier les sources majeures de variation dans la table de données. L'idée principale est donc de trouver des variables latentes, appelées composantes principales, combinant de façon linéaire les variables initiales afin de rechercher le sous-espace le plus représentatif des données.

Nous désignerons par X la matrice de dimension n individus et p variables. La réduction de dimension est obtenue par la projection des individus dans un sous-espace de plus petite dimension, espace déterminé par les composantes principales. La détermination de l'axe principal correspond à la combinaison linéaire des variables où un coefficient est appliqué, tel que la variance du nuage de points (des n individus) autour de cet axe soit maximale. Ces coefficients sont appelés les facteurs principaux (ou *loading vectors* en anglais), et sont associés à une composante principale (Figure 4.1). Une image est bonne lorsque l'on observe un maximum de la variabilité des données. Ainsi, pour une matrice de données X , l'objectif est de déterminer la composante principale expliquant le maximum de variance :

$$\arg \max_{\|a^h\|=1} \text{var}(Xa^h)$$

avec a^h le facteur principal associé à la composante principale t^h avec $h = 1, \dots, H$; et sous la contrainte que a^h ait une norme de 1. La composante principale est définie comme $t^h = Xa^h$. En outre, les composantes principales sont orthogonales entre elles.

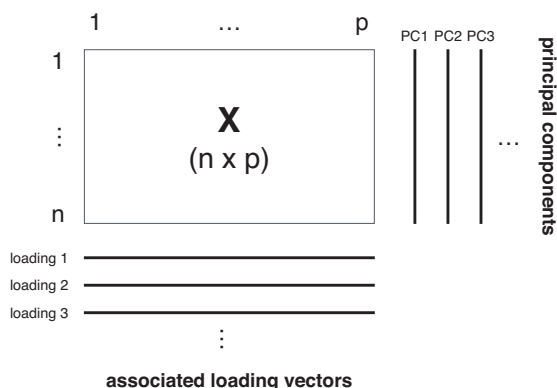


Figure 4.1 – **Vue schématique de la décomposition de la matrice X pour l’ACP.** La décomposition de la matrice X pour l’ACP consiste en un ensemble de composantes principales, chacune étant associée à des vecteurs de coefficients appelés facteurs principaux (ou *loading vectors*).

Les composantes principales sont calculées en maximisant l’inertie du nuage projeté. Cela passe par la détermination des valeurs et vecteurs propres de la matrice de variance-covariance $X^T X$, ou alors par la décomposition en valeurs singulières de X lorsque le nombre de variables est plus grand que le nombre d’individus. La première composante principale est définie comme étant la combinaison linéaire des variables expliquant la plus grande quantité de variation. Ce sous-espace, passant par le centre de gravité, a comme direction le vecteur propre associé à la plus grande valeur propre de la matrice de variance-covariance. La deuxième composante principale est considérée comme étant la combinaison linéaire expliquant la plus grande quantité de variations résiduelles, et étant orthogonale à la première composante principale. Les composantes peuvent également être calculées par l’utilisation de l’algorithme NIPALS (*Non-Linear Iterative Partial Least Square*). Le principe de cet algorithme (procédure de régression locale de façon itérative) peut être utilisé afin de réaliser une ACP avec des données manquantes sans à avoir supprimer les individus à données manquantes, ni à estimer les données manquantes. Cet algorithme permet d’estimer les paramètres d’un modèle (ici, a^h et t^h) à l’aide d’une suite de régressions simples entre les données et une partie des paramètres.

4.2.2.2 Analyse canonique des corrélations (ACC)

L’ACC présente des analogies avec l’ACP pour la construction et l’interprétation des graphiques. Cependant, contrairement à l’ACP, l’ACC permet aussi, de façon

symétrique, d'étudier les relations entre deux groupes de variables. Son objectif est donc d'explorer les relations entre deux groupes de variables quantitatives observées sur un même ensemble d'individus. Nous désignerons par X la matrice de dimension $n \times p$ contenant les données relatives au premier ensemble de variables et par Y la matrice de dimension $n \times q$ contenant celles du second ensemble de variables (Figure 4.2).

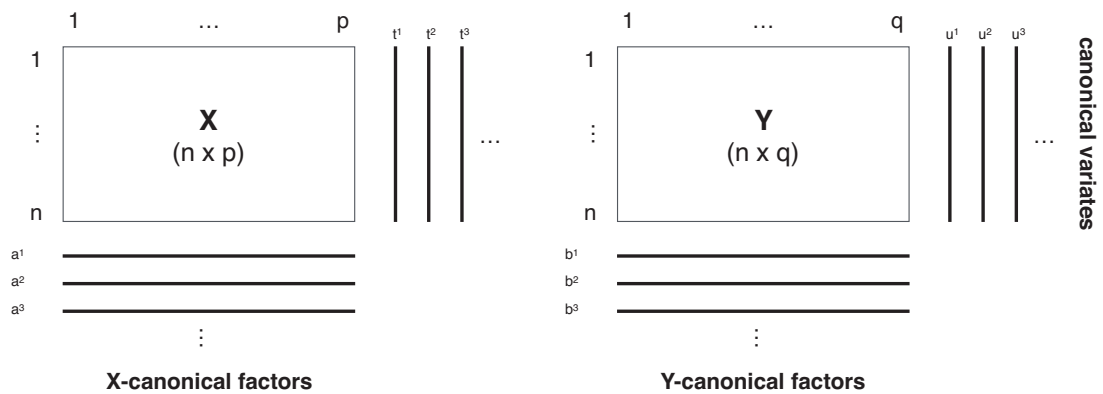


Figure 4.2 – **Vue schématique de la décomposition des matrices X et Y pour l'ACC.** La décomposition des matrices X et Y consiste en un ensemble de variables canoniques (t^h et u^h) et de facteurs canoniques associés (a^h et b^h) à ces vecteurs.

L'ACC cherche à maximiser la corrélation entre une combinaison linéaire des variables de X et une combinaison linéaire des variables de Y :

$$\arg \max_{a^h, b^h} \text{cor}(Xa^h, Yb^h)$$

où $\text{var}(Xa^h) = \text{var}(Yb^h) = 1$ avec $h = 1, \dots, H$, les vecteurs ($t^h = Xa^h, u^h = Yb^h$) sont les variables canoniques et a^h, b^h sont les facteurs canoniques associés à ces vecteurs.

Il est important de noter que, contrairement à l'ACP et la PLS présentée plus loin (section 4.2.2.3), les facteurs canoniques ne sont pas directement interprétables pour identifier l'importance des variables dans la mise en relation de X et Y . Cependant, on peut représenter les variables sur un cercle de corrélation, en projetant X et Y sur les espaces engendrés par les variables canoniques t^h et u^h avec $h = 1, \dots, H$. Les coordonnées obtenues sont les corrélations entre les variables initiales et les variables canoniques. Ainsi, il est possible de déterminer des clusters de variables

fortement corrélées entre les deux tables de données. En revanche, lorsque le nombre de variables est très grand, l'ACC tend à donner plusieurs corrélations canoniques proches de 1, indiquant ainsi que le sous-espace canonique ne semble pas couvrir toutes les observations pertinentes. En outre, la mise en œuvre de l'ACC nécessite le calcul des inverses des matrices de variance-covariance empirique $X^T X$ et $Y^T Y$ (S_{XX}^{-1} et S_{YY}^{-1}). Or, lorsque le nombre de variables (p et q) est plus grand que le nombre d'individus n , ou lorsque les variables sont fortement corrélées entre elles, ces matrices ont tendance à être singulières et donc non-inversibles, ou mal conditionnées (avec des matrices inverses instables). Une version régularisée de l'ACC (rACC) a été développée pour traiter ces problèmes de haute dimension (González *et al.*, 2008). Cette méthode calcule des paramètres de pénalités par validation croisée afin d'obtenir des vecteurs canoniques stables. Cette stratégie est disponible dans le package R **mixOmics** et utilisée lors d'analyses de données de type omique (González *et al.*, 2012).

4.2.2.3 Régression des moindres carrés partiels (PLS)

Comme l'ACC, la PLS est une approche permettant d'observer, par l'utilisation d'un modèle multivarié, les relations entre deux matrices de données X et Y ayant les mêmes individus. Cette approche possède l'avantage de ne pas être limitée par le nombre et la colinéarité des variables. Plusieurs modes de PLS existent :

- **mode régression** (ou classique) permettant d'expliquer les relations entre X et Y par la prédiction de Y selon X (Lê Cao *et al.*, 2008). L'algorithme NIPALS est à la base de la régression PLS ;
- **mode canonique** permettant d'expliquer les relations entre X et Y de façon symétrique (Lê Cao *et al.*, 2009).

Le principe est basé sur la décomposition des matrices X et Y en variables latentes et vecteurs de poids associés (Figure 4.3). Comme pour l'ACC, les variables latentes sont des combinaisons linéaires de variables, mais ici l'objectif est de maximiser la covariance et non la corrélation entre ces variables latentes :

$$\arg \max_{\|a^h\|=1, \|b^h\|=1} \text{cov}(Xa^h, Yb^h)$$

avec $h = 1, \dots, H$ et où les vecteurs ($t^h = Xa^h, u^h = Yb^h$) sont les variables latentes et les vecteurs a^h, b^h sont les poids (ou *loading vectors*) associés à ces variables. L'idée

principale de la PLS est d'effectuer successivement des régressions en utilisant des projections sur les variables latentes. Les composantes de la PLS (t^h et u^h) sont des combinaisons linéaires des variables initiales. Les coefficients définissant ces variables latentes sont déterminés par des régressions locales. Comme avec l'ACP, les variables latentes sont directement interprétables, indiquant comment certaines variables de X et Y peuvent expliquer les relations entre X et Y . Ces variables contiennent des informations concernant les similarités ou dissimilarités entre les individus ou échantillons.

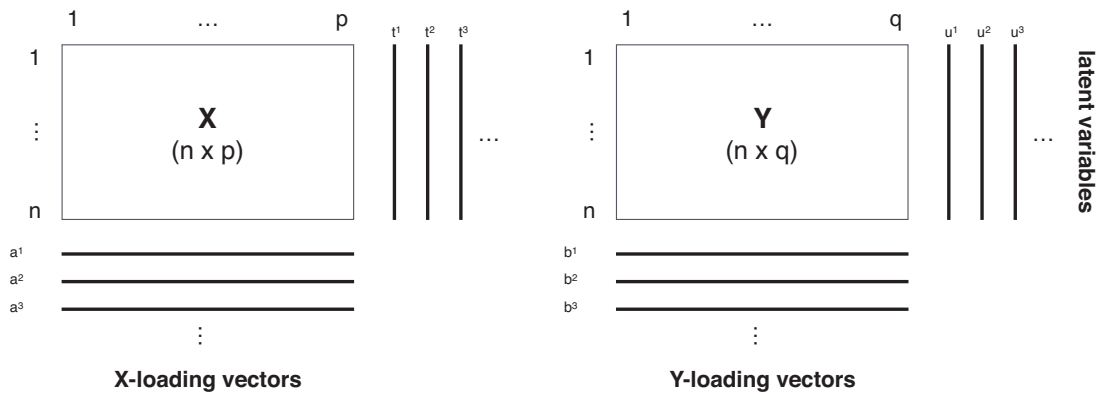


Figure 4.3 – **Vue schématique de la décomposition des matrices X et Y pour la PLS.** La décomposition des matrices X et Y consiste en un ensemble de variables latentes (t^h et u^h) et de vecteurs de poids des variables (ou *loading vectors*) (a^h et b^h).

Comme souligné précédemment, il existe une version en mode régression où la PLS modélise une relation asymétrique entre X et Y , Y étant prédit par X ; et une version en mode canonique où, comme pour l'ACC, les relations entre X et Y sont bi-directionnelles (ou symétriques).

Afin d'améliorer l'interprétation de cette méthode, une version parcimonieuse (*sparse*) a également été développée (Lê Cao *et al.*, 2008, 2009). L'objectif est de permettre la sélection de variables pour la modélisation entre X et Y . La sélection de variables est effectuée via l'utilisation d'une pénalisation de type LASSO (*least absolute shrinkage and selection operator*) sur les vecteurs *loadings* simultanément. La version *sparse* de la PLS existe pour les deux modes, canonique et régression. Comme pour l'ACC, la PLS est disponible dans le package R **mixOmics**.

4.2.2.4 Analyse factorielle multiple (AFM)

Les méthodes capables de traiter plus de deux tableaux de données sont dites de type k -tableaux, comme l'AFM (Escofier & Pagès, 1988-1998; Pagès, 2002). L'objectif de l'AFM est l'exploration simultanée de groupes de variables ayant les mêmes individus. Le coeur de l'AFM est une ACP sur le jeu de données global (concaténé), où les différents groupes de variables ont préalablement été pondérés. Cette pondération rend possible l'analyse des différents groupes de variables en s'assurant qu'aucun de ces groupes n'influencera l'analyse plus qu'un autre. La Figure 4.4 illustre de façon schématique les étapes de l'AFM.

Considérons un jeu de données $K = [K_1, K_2, \dots, K_J]$ où K_j correspond à un groupe de variables j . Lors d'une AFM, trois étapes se succèdent :

- (i) plusieurs analyses séparées sont effectuées par la réalisation d'ACP sur chaque groupe de variables K_j ;
- (ii) les groupes de variables K_j sont pondérées par $1/\sqrt{\lambda_1^j}$, λ_1^j correspondant à la première valeur propre de la matrice de variance-covariance associée au groupe de variables K_j ;
- (iii) une analyse globale est effectuée. Cette analyse est une ACP sur le jeu de données global K avec les poids $1/\sqrt{\lambda_1^j}$ pour chaque table K_j . Les graphiques (projections des individus et variables) de l'AFM sont ceux d'une ACP.

L'AFM permet de chercher des facteurs communs en fournissant une représentation graphique de chaque groupe de variables (Figure 4.5). Cette méthode donne la visualisation de la structure commune et spécifique, émergeant des différents groupes de variables K_j . Ainsi, elle permet de comparer les principaux facteurs de variabilité en reliant les représentations des groupes et des variables. La superposition des représentations des J nuages d'individus dans un même sous-espace engendré par l'AFM est possible par la propriété :

$$F_s^j(i) = F_s(i^j) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\sqrt{\lambda_1^j}} \sum_{k \in K_j} x_{ik} G_s(k)$$

où s est le numéro de la composante (axe) de l'AFM étudié, $F_s(i^j)$ est le vecteur des coordonnées de l'individu i de K_j le long de l'axe u_s , λ_1^j est la première valeur

propre de l'ACP de K_j , λ_s est la $s^{ième}$ valeur propre de l'AFM (ACP globale) et $G_s(k)$ les coordonnées de la variable k de K_j pour l'axe u_s .

L'AFM est disponible dans le package R **FactoMineR**. Cette méthode est généralement appliquée à des données de type sensoriel, comme des données traitant de la qualité des aliments, mais elle a déjà été utilisée pour des données de métabonomique (Dumas *et al.*, 2005) ou des données de génomique et transcriptomique (de Tayrac *et al.*, 2009).

Au niveau des différences concernant les méthodes ACC/PLS et AFM, l'AFM est principalement une analyse descriptive globale de données concaténées et pondérées, où plusieurs jeux de variables sont étudiées simultanément. Cette méthode peut notamment être utilisée afin d'analyser les structures conjointes entre plusieurs jeux de données. Les structures communes sont mises en avant, et des sorties graphiques facilement analysables sont fournies afin d'identifier une signification biologique (s'il y en a une). La CCA et la PLS sont, quant à elles, des méthodes de maximisation de corrélation ou covariance (respectivement), où l'objectif principal est d'observer les relations entre variables des différents jeux de données. En outre, les versions *sparse* de ces méthodes permettent également la sélection des variables qui sont les plus importantes pour la modélisation.

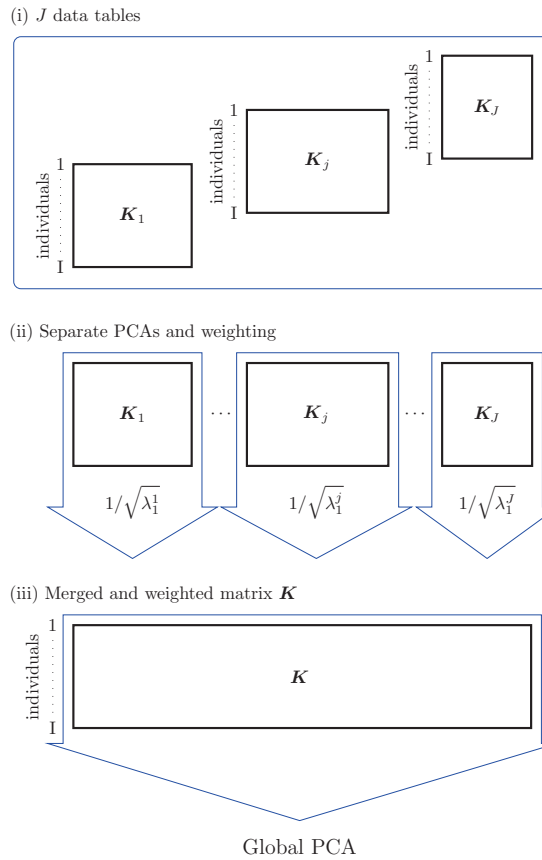


Figure 4.4 – **Vue schématique de l’AFM.** (i) J groupes de variables mesurées sur les mêmes individus. (ii) Des ACPs sont réalisées pour chaque groupe de variables K_j , suivi d’une pondération par $1/\sqrt{\lambda_1^j}$, où λ_1^j est la première valeur propre de l’ACP de K_j . (iii) Réalisation d’une ACP globale sur le jeu de données K .

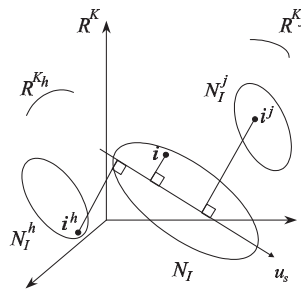


Figure 4.5 – **Principe de la superposition des représentations fournie par l’AFM (de Pagès (2002)).** Chaque nuage de points partiel N_I^j est projeté sur les axes obtenus par le nuage de points moyen N_I .

4.2.3 Contribution de la thèse : les valeurs manquantes dans le cadre de l'AFM

4.2.3.1 Les valeurs manquantes en biologie

Quels que soient les types de données, les valeurs manquantes sont souvent présentes pour diverses raisons. En outre, les risques d'être confronté à des valeurs manquantes augmentent de façon concomitante avec le nombre de sources d'informations. Très souvent les valeurs manquantes sont supposées être manquantes au hasard. Dans cette thèse, nous allons traiter des cas spécifiques pour lesquels nous avons toute une ligne de valeurs manquantes dans un ou plusieurs tableaux, typiquement des valeurs manquantes d'un individu pour tout un groupe de variables. Avec les données omiques, généralement prélevées sur un même groupe d'individus, il est fréquent que des individus soient manquants pour tout un groupe de variables. A l'heure actuelle, il existe encore très peu d'articles scientifiques dans la littérature abordant cette thématique.

L'intégration de données hétérogènes étant un challenge dans le domaine de la biologie des systèmes, notamment avec les données de type omique, le problème des lignes manquantes pour tout un groupe de variables peut entraîner des pertes d'informations et donc des modifications au niveau des conclusions biologiques. Les méthodes de statistique multidimensionnelle, comme l'ACC, la PLS ou l'AFM, ne peuvent pas être appliquées directement sur des jeux de données ayant des valeurs manquantes. Lorsque des lignes manquantes sont présentes, deux options s'offrent aux scientifiques : soit supprimer l'individu, ceci entraînant une perte d'information ; soit les remplacer par des valeurs plausibles (comme la moyenne des observations). Dans la littérature, peu de stratégies ont été développées dans le cadre d'analyses multidimensionnelles (Van de Velden & Bijmolt, 2006; Van de Velden & Takane, 2011). L'idée principale de ces méthodes est d'essayer d'estimer des paramètres (comme les dimensions ou les axes) en imputant les valeurs manquantes par l'utilisation d'algorithmes itératifs.

Dans le cadre de l'AFM, une stratégie a été développée par Husson & Josse (2013), appelée la *regularized iterative MFA* (RI-MFA). Cette méthode, inspirée de la *regularized PCA* (Josse & Husson, 2012), consiste en une alternance entre l'estimation des axes et composantes de l'AFM, et l'estimation des valeurs manquantes. Dans

cette thèse, nous avons souhaité proposer une méthode alternative, appelée *multiple imputation - MFA* (MI-MFA), dans laquelle nous avons implémenté l'algorithme d'imputation multiple dans le cadre de l'AFM. Notre objectif est d'imputer les lignes manquantes dans une analyse AFM afin d'estimer les composantes de cette dernière. Notre méthode est présentée dans la partie suivante 4.2.3.2 et est appliquée à nos données dans la partie 4.2.3.3.

4.2.3.2 Article 4 : Voillet et al., *BMC Bioinformatics*, 2016

Cette section correspond à l'article suivant accepté en septembre 2016 dans le journal *BMC Bioinformatics*:


- **V. Voillet**, P. Besse, L. Liaubet, M. San Cristobal and I. González. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics* **17**, 402 (2016).

METHODOLOGY ARTICLE

Open Access



Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework

Valentin Voillet^{1†}, Philippe Besse², Laurence Liaubet¹, Magali San Cristobal^{1,2}
and Ignacio González^{3*†} 

Abstract

Background: In omics data integration studies, it is common, for a variety of reasons, for some individuals to not be present in all data tables. Missing row values are challenging to deal with because most statistical methods cannot be directly applied to incomplete datasets. To overcome this issue, we propose a multiple imputation (MI) approach in a multivariate framework. In this study, we focus on multiple factor analysis (MFA) as a tool to compare and integrate multiple layers of information. MI involves filling the missing rows with plausible values, resulting in M completed datasets. MFA is then applied to each completed dataset to produce M different configurations (the matrices of coordinates of individuals). Finally, the M configurations are combined to yield a single consensus solution.

Results: We assessed the performance of our method, named MI-MFA, on two real omics datasets. Incomplete artificial datasets with different patterns of missingness were created from these data. The MI-MFA results were compared with two other approaches i.e., regularized iterative MFA (RI-MFA) and mean variable imputation (MVI-MFA). For each configuration resulting from these three strategies, the suitability of the solution was determined against the true MFA configuration obtained from the original data and a comprehensive graphical comparison showing how the MI-, RI- or MVI-MFA configurations diverge from the true configuration was produced. Two approaches i.e., confidence ellipses and convex hulls, to visualize and assess the uncertainty due to missing values were also described. We showed how the areas of ellipses and convex hulls increased with the number of missing individuals. A free and easy-to-use code was proposed to implement the MI-MFA method in the R statistical environment.

Conclusions: We believe that MI-MFA provides a useful and attractive method for estimating the coordinates of individuals on the first MFA components despite missing rows. MI-MFA configurations were close to the true configuration even when many individuals were missing in several data tables. This method takes into account the uncertainty of MI-MFA configurations induced by the missing rows, thereby allowing the reliability of the results to be evaluated.

Keywords: Multiple omics data integration, Multivariate factor analysis, Missing individuals, Multiple imputation, Hot-deck imputation

*Correspondence: Ignacio.Gonzalez@toulouse.inra.fr

†Equal contributors

³INRA, UR875 Mathématiques et Informatiques Appliquées, F-31326
Castanet-Tolosan, France

Full list of author information is available at the end of the article



Background

Due to the increase in available data information [1], integrating large amounts of heterogeneous data is currently one of the major challenges in systems biology. Biological data integration provides scientists with a deeper insight into complex biological processes. However, when dealing with multiple data tables, the presence of missing values is a common situation for a variety of reasons. In omics data integration studies, it is common for some individuals to not be present in all data tables, resulting in a specific missing data pattern for multiple tables, as shown in Fig. 1. For instance, in clinical studies, this can occur when a patient forgets to fill out a form. It also can be attributable to the study design if individual data are expensive or difficult to measure.

Missing row values for a table of variables are challenging to handle because most statistical methods cannot be directly applied to incomplete datasets. In the multiple multivariate framework, several approaches have already been proposed to deal with missing row values [2]. The only methods widely available for analyzing incomplete data focus on removal of the missing rows, either by ignoring subjects with incomplete information or by replacing the missing items with plausible values (e.g., means of the observed cases). In multivariate statistical analysis, case deletion procedures can be very inefficient, discarding an unacceptably high proportion of subjects because even if the per-table rates of missing rows are low, only a few subjects may have complete data for all tables. In addition, case-deletion procedures may bias the results if the remaining subjects providing the complete data are unrepresentative of the entire sample [3, 4]. On the other hand, simple mean substitution can seriously distort the marginal and joint distribution of the variables [5] and be an issue because many statistical methods rely on estimation of the variance-covariance matrix.

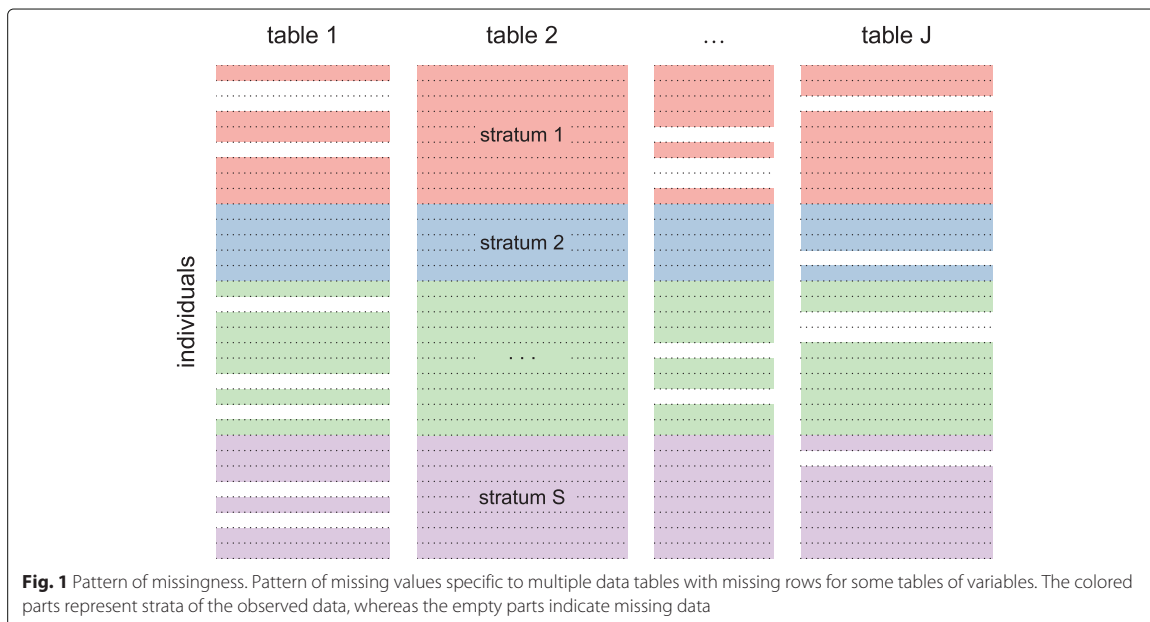
Recently, two approaches have been proposed to deal with missing row values in multiple multivariate analysis. The first method, introduced by Van de Velden and Bijmolt (2006) [6], was developed in the context of generalized canonical correlation analysis. Its application in the omics framework is often limited by the size, noise and multicollinearity of the data [7, 8]. The second method, described in Husson and Josse (2013) [9], was developed in the context of multiple factor analysis (MFA). This method, designated regularized iterative MFA (RI-MFA), was derived from a method available in principal component analysis (PCA) and consists of alternating the estimation of axes and components, and the estimation of missing values [10, 11]. Here we consider an alternative method, involving a multiple imputation approach adapted to the MFA framework, and called MI-MFA.

Multiple imputation (MI) was proposed by Rubin (1987) [3] in order to estimate both the parameters of interest

and their variability in a data missingness framework. It relies on the principle that a single value cannot reflect the uncertainty of the estimation of a missing value. First, MI is used to generate plausible synthetic data values, called imputations, for missing values in the data. This step results in a number (M) of imputed datasets in which the missing data are replaced by random draws of plausible values according to a specific statistical model. The second step consists of analyzing each imputed dataset using a statistical method that estimates the parameters of interest. This step results in M analyses (instead of just one) which differ only because the imputations differ. Finally, MI combines all the results together to obtain a single consensus estimate, thereby combining variation within and across the M imputed datasets. Under fairly liberal conditions, this last step results in statistically valid estimates that properly reflect sampling variability.

The major challenge in MI involves generating possible values for each missing observation. Statistically advanced imputation procedures can therefore be used for this. Two general approaches are often used for imputing multivariate data: joint modeling (JM) [12] and fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE) [13, 14]. JM involves specifying a multivariate distribution for the missing data and drawing imputations from their conditional distributions by Markov chain Monte Carlo (MCMC) techniques. FCS specifies the multivariate imputation model on a variable-by-variable basis using a set of conditional models, one for each incomplete variable.

The key issue in JM is appropriate specification of the multivariate distribution. A multivariate normal model has often been used as it is computationally tractable (because only the mean vector and the variance-covariance matrix need to be estimated). This model has even been used when some of the variables are not Gaussian. However, the main weakness of JM is that it can only be applied when the imputation involves a small number of variables. This is not very common in omics datasets that are often composed of tens of thousands of variables, or more. FCS allows greater flexibility than JM in creating multivariate models. Indeed, FCS can use specialized imputation models by separately defining the conditional densities for each variable, even if this can require a considerable amount of work. When the number of variables is large, it is often impractical and too computer-intensive to test and develop the best models for each variable. As an alternative to the JM and FCS approaches, we propose using the hot-deck imputation approach [5]. This approach is a nonparametric imputation method that resolves the most important limitation of the JM and FCS approaches as it can be applied to data tables containing more than just a few variables.



When using the MI method, special attention must be given to the process that gave rise to the missing data, referred to as the missing data mechanism. Most methods for generating multiple imputations, fully-, semi- and non-parametric methods, assume that the mechanism responsible for missing data is ignorable [5, 15]. Briefly, if the missing data mechanism is ignorable, then the analysis can focus on the observed values rather than also having to model the process that resulted in certain values being observed and certain values being missing. If the assumption of an ignorable missing-data mechanism is valid, then statistical methods that rely on that assumption can be expected to produce results with minimal bias. One way that the missing-data mechanism can be viewed as ignorable is if the missing data are missing completely at random (MCAR). For data to be MCAR, there must not be any systematic differences between the cases that have missing items and the cases that are fully observed. In microarray experiments, technical failure, low signal-to-noise ratio and measurement errors can for instance be considered as sources of MCAR patterns. The missing-data mechanism can also be viewed as ignorable under the less restrictive missing at random (MAR) scenario, which allows missingness to depend on observed variables but not on unobserved variables. Late-stage cancer patients, as compared to early-stage cancer patients, unfortunately have more chance of dropping out of follow-up studies, which may result in a MAR pattern in a clinical data table. When the ignorability assumption does not hold, the imputation needs to be drawn from the posterior

distribution of the missing data given the complete data and the missingness mechanism. Non-ignorable missing data occurs frequently in mass-spectrometry-based experiments. Measures too close to the limit of detection of the instrument are censored, resulting in a higher rate of missing values. The probability of being missing is, in this particular case, directly dependent on the intensity value. In this paper, we decided to focus on models with ignorable missing-data mechanisms.

Methods

Mathematical basis of multiple factor analysis (MFA)

MFA [16] is devoted to the simultaneous exploration of multiple data tables where the same individuals are described by several tables of variables. In MFA, the number and the type of variables (quantitative or categorical) may vary from one table to another, but within each table the nature of the variables is the same. Here we focus on quantitative variables. The aims of MFA are similar to those of PCA, namely to study the similarities between individuals from a multidimensional point of view, to analyze the relationships between variables and characterize individuals based on these relationships. However, beyond these conventional uses, MFA can also be used to study the links between tables of variables and to compare the information contributed by each table.

MFA analyzes a set of J data tables K_1, \dots, K_J , where each K_j corresponds to a table of quantitative variables measured on the same I individuals (for a schematic overview of MFA see Additional file 1: Figure S1). The

core of MFA is a PCA in which weights are assigned to variables. More formally, the matrix of variance-covariance associated with each data table K_j is decomposed by PCA and its largest eigenvalue λ_1^j is derived. Then, each variable belonging to K_j is weighted by $1/\sqrt{\lambda_1^j}$. Finally, a global PCA is performed on the merged and weighted data table $K = [K_1, \dots, K_J]$ to obtain the configuration F (the scores matrix or principal components). The main reason for the weighting step is to remove from each table all information related to its own dimensionality or variance. Therefore, no single table can dominate the first dimension of the global analysis.

MFA provides the same graphical representations as PCA (i.e., representation of individuals and variables) but also, due to the table structure, specific representations such as the table representation and the superimposed representation are available [16].

The multiple imputation multiple factor analysis approach (MI-MFA)

To deal with multiple tables with missing rows, we propose the MI-MFA approach, a multiple imputation (MI) adapted to the framework of MFA. The aim of our method is not to get the best possible estimations of the missing values, but to replace them with plausible values in order to provide estimates of the MFA configurations. According to MI methodology, the MI-MFA approach is carried out by performing the following three steps:

1. Imputation: generate M different imputed datasets $K^{(1)}, \dots, K^{(m)}, \dots, K^{(M)}$ of K .
2. MFA analysis: perform an MFA on each $K^{(m)}$ imputed dataset leading to M different configurations $F_1, \dots, F_m, \dots, F_M$.
3. Combination: find a consensus configuration between all F_1, \dots, F_M configurations.

These steps are outlined in Fig. 2 and described in detail in the following sections.

Generating imputed data: multiple hot-deck imputation

Hot-deck imputation involves replacing missing values of one or more variables with available values from a similar unit [17]. The observation from which these available values are taken for imputation is called the donor and the observation with the missing value, which receives the donor's value, is the recipient. The donor can be randomly selected from a set of potential donors, called the donor pool. Selection of a suitable donor pool is not an easy task and is beyond the scope of this article [18, 19]. The general principle is to choose donor units that are as close as possible to the recipient with respect to some *affinity score*. Affinity is defined in terms of the degree to

which each potential donor matches the recipient's values across all variables other than the one being imputed. Intuitively, in the framework of stratified multiple omics tables, the donor pool can be formed of available individuals belonging to the same stratum (e.g. cancerous cell line, treatment, etc.) and the same omics table as the recipient.

Multiple hot decking differs from other forms of hot decking by using several donors for a single recipient [20]. Multiple hot-deck imputation proceeds as follows. Let $K = [K_1, \dots, K_J]$ be the merged data table containing missing rows with strata $s = 1, \dots, S$, then carry out the following steps (see Fig. 2, hot-deck imputation step):

- Step 1.** Create donor pools by taking donors belonging to the same stratum s and the same table K_j as the recipient. Recipients within the same stratum have the same donor pool. Suppose always that there is a large enough number of donors for recipients in each stratum.
- Step 2.** For each recipient in K_j , impute the missing individual by drawing randomly with replacement a donor from the corresponding donor pool. Repeat this procedure until all missing individuals in the J tables have been imputed.
- Step 3.** Repeat Step 2 until M different imputed datasets $K^{(1)}, \dots, K^{(M)}$ of K are obtained.

By conducting the imputations in this way, it is reasonable to assume that the within-unit between-variables multivariate relationships are preserved.

The combination procedure: the STATIS method

The question that arises after using MI in an MFA framework is how should all the configurations resulting from the analyses be combined to obtain a single unique estimate of the *consensus* configuration? While averaging is an appropriate combination procedure in many other statistical techniques, it is not recommended for MFA due to possible reflection, dilation or rotation of the different configurations with respect to each other [21]. Here we consider an alternative approach by implementing the STATIS method which provides a compromise configuration balancing all configurations.

The STATIS method [22] (which stands for *Structuration des Tableaux à Trois Indices de la Statistique* in French) is a generalization of PCA used to simultaneously study several tables of variables collected on the same individuals. The goal of this method is to analyze the structure of the individual tables (i.e., the relation between the individual tables) and to derive from this structure an optimal set of weights for computing a common configuration of the observations. The solution obtained, called the compromise, is the

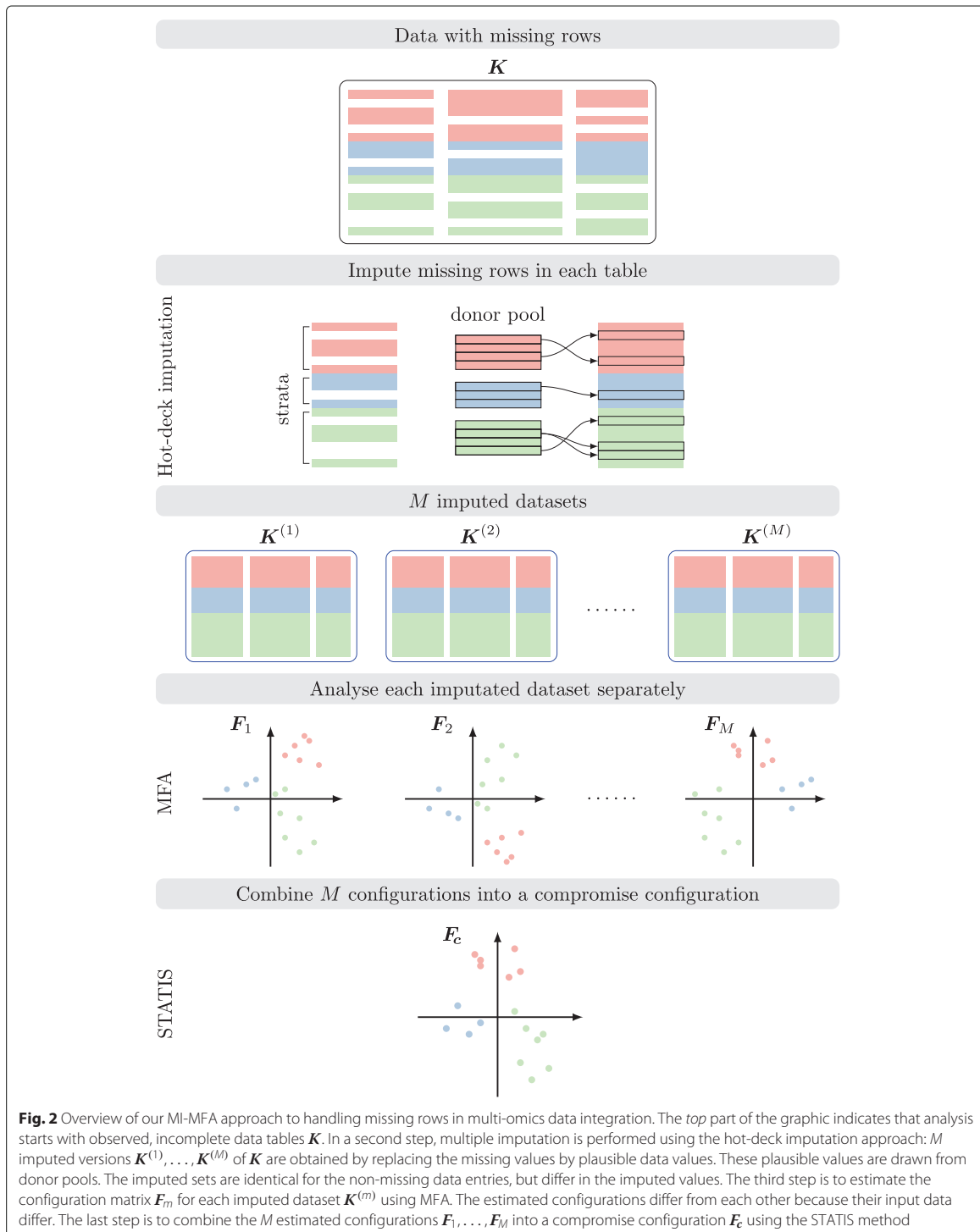


Fig. 2 Overview of our MI-MFA approach to handling missing rows in multi-omics data integration. The top part of the graphic indicates that analysis starts with observed, incomplete data tables K . In a second step, multiple imputation is performed using the hot-deck imputation approach: M imputed versions $K^{(1)}, \dots, K^{(M)}$ of K are obtained by replacing the missing values by plausible data values. These plausible values are drawn from donor pools. The imputed sets are identical for the non-missing data entries, but differ in the imputed values. The third step is to estimate the configuration matrix F_m for each imputed dataset $K^{(m)}$ using MFA. The estimated configurations differ from each other because their input data differ. The last step is to combine the M estimated configurations F_1, \dots, F_M into a compromise configuration E_c using the STATIS method

configuration agrees the most with all other configurations. An overview of the STATIS method is presented in Additional file 1: Figure S2.

STATIS analyzes a set of N tables X_1, \dots, X_N , where each X_n is a table of quantitative variables measured on the same individuals. The first stage of STATIS consists of calculating a matrix of cross-products between individuals for each table $W_n = X_n X_n^T$ (A^T means the transpose of a vector or a matrix A) reflecting the similarities between individuals within this table. The use of matrices W_n instead of X_n simplifies the computation because it obviates the determination of rotations when matching the X_n . The basic idea in STATIS is then to find a compromise space $W_c = \sum_{n=1}^N \alpha_n W_n$ that globally balances these cross-product matrices by choosing a suitable optimal set of weights $\alpha_1, \dots, \alpha_N$. These weights are obtained from the PCA of the matrix R whose generic term R_{jk} gives the cosine between tables (also known as the RV-coefficient [23]) defined as:

$$R_{jk} = \frac{\text{trace}(W_j^T W_k)}{\sqrt{\text{trace}(W_j^T W_j) \cdot \text{trace}(W_k^T W_k)}}$$

where the trace is the sum of the main diagonal elements of a square matrix. The first eigenvector obtained from the PCA of R represents the “agreement between tables”. Its elements are normalized in such a way that their sum is equal to 1 and used as weights α_n in order to define W_c . Tables with larger values of α_n are more similar to the other tables and therefore will have a larger weight, while the weight of the “outlier tables” will be closer to zero with respect to the other weights. The principal components from the PCA of W_c then gives the coordinates of the individuals in the compromise space, called the *compromise configuration*.

Implementation of MI-MFA

The MI-MFA algorithm can be summarized as follows (see Fig. 2):

- Step 0.** Start with an observed, incomplete dataset K . Define the number of imputations M and the dimensionality d of the compromise configuration.
- Step 1.** Perform multiple hot-deck imputation. For $m = 1, \dots, M$:
 - Obtain an imputed version $K^{(m)}$ of K , such that, $K^{(m)} \neq K^{(m')}$ for $m \neq m'$. The imputed datasets are identical for the non-missing data entries, but differ for the imputed values. The imputed version of the data is obtained by using the hot-deck imputation approach.
 - Perform an MFA using d components on the imputed dataset $K^{(m)}$ to obtain the configuration F_m .

- Step 2.** Perform a STATIS on the set of configurations F_1, \dots, F_M to obtain F_c , the compromise configuration.

Note that the number of dimensions d used in the algorithm has to be chosen a priori. However, the number of dimensions does not affect the estimation of the imputed values and the estimation of the compromise configuration. Moreover, for given $K^{(1)}, \dots, K^{(M)}$ imputed datasets, solutions provided by the algorithms are nested (the solution with d dimensions is included in the solution with $d+1$ dimensions). Since the core of MI-MFA is a weighted PCA, the strategies suggested to choose the number of components in PCA can be adapted to MI-MFA, but work needs to be done to validate the quality of these extensions.

How many imputations?

When using MI, one of the uncertainties concerns the number M of imputed datasets needed to obtain satisfactory results. The number of imputed datasets in MI depends to a large extent on the proportion of missing data. The greater the missingness, the larger the number of imputations needed to obtain stable results. However, in multiple hot-deck imputation, the number of imputed datasets is limited by the size of the donor pools. In any case, the total number of possible imputations M_{total} can be calculated before applying the imputation approach (see Additional file 2). If M_{total} is small ($M_{total} \leq 50$), then $M = M_{total}$ can be used in MI-MFA. The appropriate number of imputations can be informally determined by carrying out MI-MFA on N replicate sets of M_l imputations for $l = 0, 1, 2, \dots$, with $M_0 < M_1 < M_2 < \dots < M_{total}$, until the estimate compromise configurations are stabilized. More precisely, this approach can be carried out by applying the following steps:

- Step 0.** Start with an observed, incomplete dataset K . Define the number of imputations M_l with $M_0 < M_1 < M_2 < \dots < M_{total}$ and the number N of replicate sets of M_l imputations.
- Step 1.** Create collections $\mathcal{I}_n^{M_l}$, $n = 1, \dots, N$, each one containing M_l different imputed datasets of K , such that, $\mathcal{I}_n^{M_l} \neq \mathcal{I}_{n'}^{M_l}$, for $n \neq n'$ and $\mathcal{I}_n^{M_{l-1}} \subset \mathcal{I}_n^{M_l}$ for $M_0 < M_1 < M_2 < \dots < M_{total}$.
- Step 2.** For $n = 1, \dots, N$, perform an MI-MFA using $\mathcal{I}_n^{M_0}$, to obtain N different compromise configurations $F_{c1}^{M_0}, \dots, F_{cN}^{M_0}$.
- Step 3.** Let $l = 1$. For $n = 1, \dots, N$,
 - perform an MI-MFA using the collection $\mathcal{I}_n^{M_l}$, to obtain a compromise configuration $F_{cn}^{M_l}$;

- calculate $r_n^l = r(\mathbf{F}_{cn}^{M_l}, \mathbf{F}_{cn}^{M_{l-1}})$, a measure of the distance or correlation between configurations (for example the RV coefficient [23]).

Step 4. Calculate $\bar{r}^l = \frac{1}{N} \sum_{n=1}^N r_n^l$ (or $\sigma(\bar{r}^l)$) the standard error of \bar{r}^l .

Step 5. Repeat steps 3 to 4 for $l = 2, 3, \dots$ until the differences between two subsequent \bar{r}^l (or $\sigma(\bar{r}^l)$) become smaller than a certain convergence criterion.

Uncertainty of MI-MFA solutions

In an MI-MFA framework, after estimating the configurations from the imputed datasets, a new source of variability due to missing values can be taken into account. Here we describe two approaches to visualize the uncertainty of the estimated MFA configurations attributable to missing row values. First, an individual plot for all estimated MFA configurations is constructed. The individual plot is obtained by projecting each estimated MFA configuration onto the compromise configuration (named the trajectories by Lavit [22]). Each individual is represented by M points, each corresponding to one of the M MFA configurations. Confidence ellipses and convex hulls can then be constructed for the M configurations for each individual. The computed convex hull results in a polygon containing all M solutions. All individuals have confidence areas, even those without missing values. Indeed, even if only the estimation of missing values is the only change, this will have a possible impact on all MFA parameters. Therefore, the area of such an ellipse (or convex hull) provides an insight into the uncertainty of the estimated configuration. The larger the area of an ellipse (convex hull), the more uncertain the exact location of the individual. Thus, when the area of an ellipse is large, the scientist should remain really careful regarding its interpretation.

Performance of the method

We conducted two case studies to assess the performance of our method. Instead of using theoretical distributions to generate simulated data, our studies were based on two real datasets, denoted as the original datasets. Subsequently, specific patterns of missingness were created in these datasets as illustrated in Fig. 1, resulting in what we called the incomplete datasets. This approach was used in order to more closely mirror situations that may occur in the omics context. Next, missing row values were estimated and the resulting complete datasets were referred to the imputed datasets.

We then compared our MI-MFA method to the RI-MFA approach [9] and the mean variable imputation MFA (MVI-MFA) method, in which the missing values are simply replaced by the mean of each variable after which an MFA is performed on the imputed dataset. This latter approach was considered as the common base for

comparing the MI-MFA and RI-MFA methods. For each configuration obtained using MI-, RI- and MVI-MFA, the similarity between the configuration solution and the true configuration (based on an MFA using the original dataset) was assessed from the RV coefficient [23]. The RV coefficient, which ranges from zero to one, can be interpreted as a correlation coefficient between two matrices, which allows the relative positions of objects to be compared from one configuration to another.

We also provide comprehensive graphical comparisons of the true vs. the MI-, RI- or MVI- MFA configurations. The individuals from both configurations are drawn in a same plot and connected by an arrow, the length of which indicates the divergence between the two configurations.

Implementation of the analyses

All analyses were performed using the R computing environment [24]. MFA was performed using the *MFA* function of the *FactoMineR* R package [25]. The *statis* function of the *ade4* R package [26] was used to determine the compromise configuration. The RI-MFA method is implemented in the *imputeMFA* function available in the *missMDA* R package [27]. Note that the number of components *ncp* used to predict the missing entries in the *imputeMFA* function has to be chosen a priori. This choice is crucial and difficult [9]. As the true configuration was known in our case, the number of components *ncp* was chosen to minimize the RV coefficient between the true and the *imputeMFA* configurations.

The appropriateness of the results from MI-, RI- and MVI- MFA was then determined by comparing the configurations resulting from these three strategies with the true MFA configuration. Due to a possible lack of alignment (order change, sign reversal of the components and rotation) between two configurations (the true vs. the MI-, RI- or MVI- MFA configuration), it was necessary to align them before being compared. Ordinary Procrustes Analysis [28] was used to align these configurations prior to their comparison.

Datasets

Liver toxicity

The datasets originated from a liver toxicity study [29] in which 64 male rats of the inbred Fisher F344/N strain were exposed to toxic doses of acetaminophen (paracetamol) in a controlled experiment. Necropsies were performed 6, 18, 24 and 48 h after exposure and mRNA was extracted. The data consisted of the expression of 3,116 genes and 10 clinical variables considered to be markers of liver injury. The 64 subjects (rats) were cross-classified in eight strata (or treatments) according to two factors:

- exposure time: 6, 18, 24 and 48 h;

- toxic doses of acetaminophen: high (1500 mg/kg or 2000 mg/kg) or low (50 mg/kg or 150 mg/kg).

Eight subjects per stratum were included. These datasets were downloaded from the *mixOmics* R package [30].

NCI-60 data

The NCI-60 dataset contained transcriptomic [31] and proteomic [32] tables for a collection of 60 cell lines from the National Cancer Institute (NCI-60). The NCI-60 panel included cell lines derived from various cancer types: colon (7 cell lines), renal (8), ovarian (6), breast (8), prostate (2), lung (9) and central nervous system (6), as well as leukemia (6) and melanoma (8). The gene expression profiles used here were generated using an Agilent platform [31] and downloaded from Cellminer [33]. Data were \log_2 -transformed. To facilitate data interpretation and computation, the transcriptomic data were filtered to exclude probes that did not map to an official HUGO gene symbol and to retain only the probe with the highest average value when multiple probes mapped to the same gene, as previously described in [34]. Gene invariants across all 60 cell lines, corresponding to genes without any effect between cancer types, were also removed. Filtering produced a dataset of 1,433 genes. The NCI-60 proteome table was also downloaded from Cellminer [33]. Proteomic data were obtained using high-density reverse-phase lysate microarrays [32]. Data were \log_2 -transformed and protein abundance levels were available for 162 proteins [32].

Results

Liver toxicity data analysis

A specific pattern of missing values was created as illustrated in Fig. 1. To obtain an incomplete dataset, three individuals per stratum were randomly removed from the transcriptomic table. For this specific pattern, there were $3 \times 8 = 24$ missing individuals. MI-MFA was then performed on the incomplete dataset using $M = 30$ imputed datasets. RI-MFA, and MVI-MFA were also performed. Figure 3 shows the divergence of the MI-, RI- and MVI-MFA configurations from the true configuration. As can be seen, the configuration obtained with MI-MFA was very close to the true configuration (Fig. 3, top right). This result was confirmed by the high RV coefficient (0.96 for the first two dimensions). The configurations obtained with RI-MFA and MVI-MFA were more distorted and less close to the true configuration with RV coefficients of 0.77 and 0.84 respectively (Fig. 3, bottom).

The number M of imputed datasets in MI-MFA for the above incomplete liver toxicity data was determined as described in the *How many imputations* section. Collections of size $N = 30$ were generated for each of the following numbers of imputations: $M_l = 10l$, for

$l = 1, \dots, 10$. The stability of the estimated MI-MFA configurations was then determined by calculating the RV coefficient between the configurations obtained using M_l and M_{l+1} imputations (see Fig. 4, left). As the true configuration was known, we also described the stability of the estimated MI-MFA configurations by calculating the RV coefficients between the true configuration and those obtained using M_l imputations (see Fig. 4, right). Although the missing information is substantial, Fig. 4 shows that only a slight increase in precision was obtained by using more than 30 imputations.

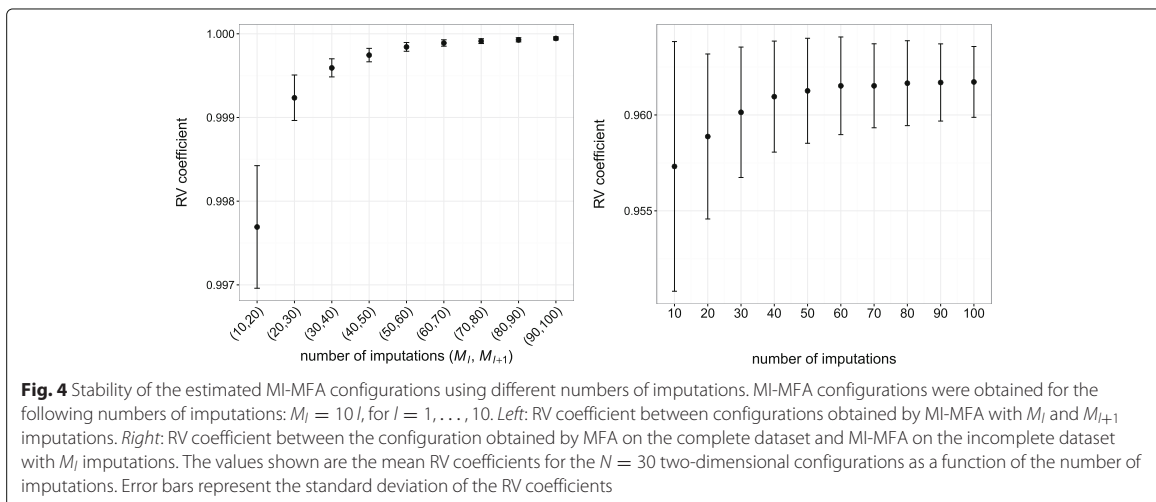
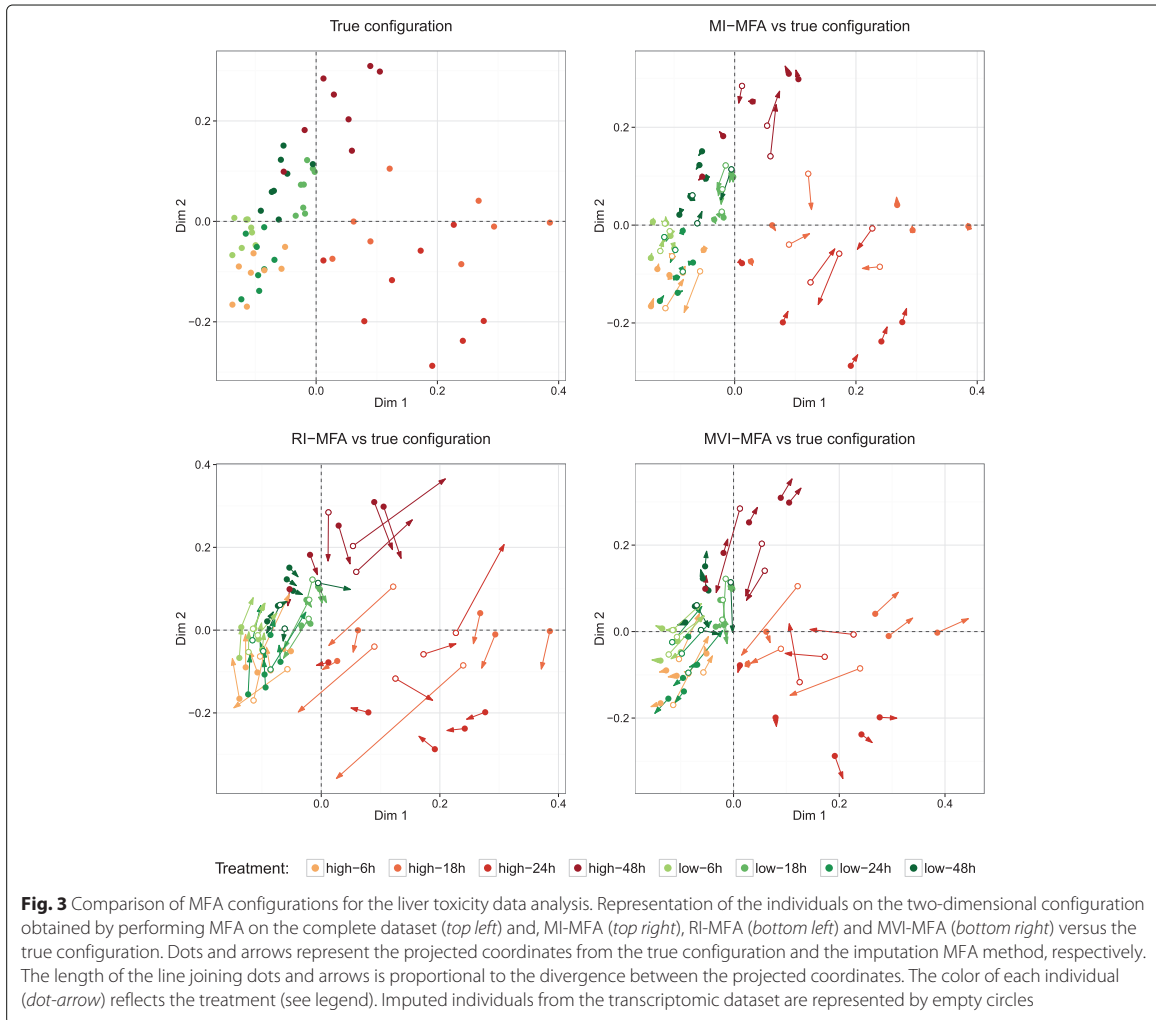
MI-MFA was also applied to different scenarios of missingness. First, MFA was performed on the original dataset to obtain the true configuration. Secondly, individuals were randomly removed from each stratum in the original transcriptomic table. For this, three scenarios were considered for the number of missing rows (i.e. low, medium and high), in which there were respectively one, two and three missing rows per stratum. The total number of missing rows per scenario was therefore 8, 16 and 24 respectively. Thirdly, for each missingness scenario, 50 incomplete datasets were randomly chosen for analysis.

As previously, MI-MFA was performed on each incomplete dataset using $M = 30$ imputed datasets. RI-MFA, as well as MVI-MFA, were also computed. The RV coefficients between the true configuration and the configurations obtained using each method (for the first two dimensions) were then calculated. Figure 5 shows the mean of the RV coefficients for the 50 two-dimensional configurations as a function of the missingness scenario for each method. Note that the average results using MI-MFA were always better than with RI-MFA or MVI-MFA, whatever the scenario. The RV coefficients between the true configuration and the MI-MFA configuration were close to one and remained satisfactory even when the number of missing row values was high, and the results obtained with the RI-MFA and MVI-MFA decreased significantly.

The performance of the MI-MFA procedure was then further investigated for more complex scenarios of missingness. More precisely, missing row values were inserted into each stratum (e.g. high-6h treatment) of the original dataset (including both transcriptomic and clinical tables) according to the scenarios illustrated in Table 1.

Twenty incomplete datasets were then selected at random from each stratum (treatment) and each scenario. All analyses (MI-MFA, RI-MFA and MVI-MFA) and the calculation of the RV coefficient were performed in the same way as previously described. Figure 6 (and Additional file 1: Figure S3) shows the means of RV coefficients for the two-dimensional configurations as a function of the scenarios for each method.

For almost all the scenarios, the average results obtained with MI-MFA were better than with the other methods.



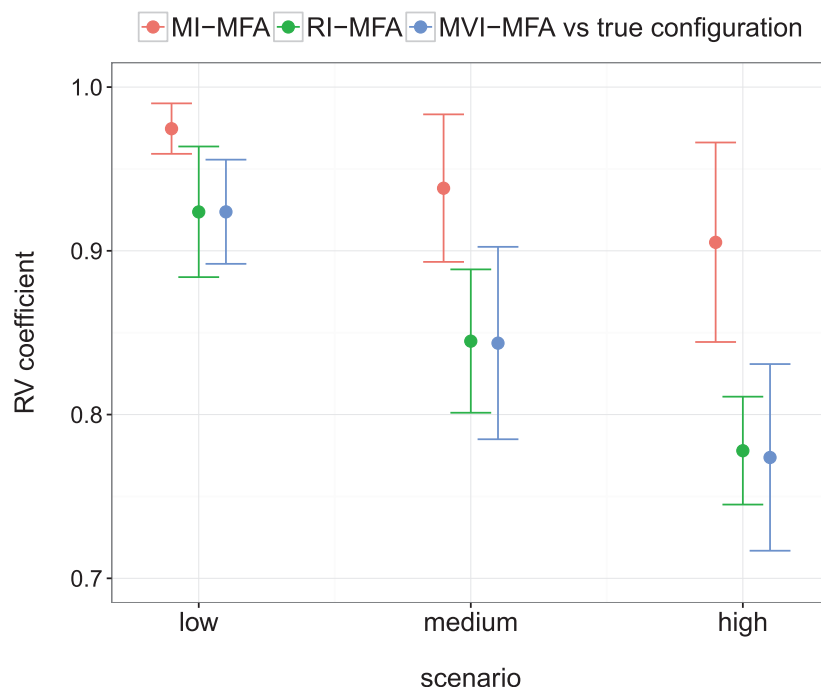


Fig. 5 Performance on liver toxicity data according to the number of missing rows. RV coefficients between the configuration obtained by MFA on the complete dataset and either MI-MFA (red line), RI-MFA (green line) or MVI-MFA (blue line) on the incomplete dataset. The values shown are the mean RV coefficients for the 50 two-dimensional configurations for each missingness scenario: low, medium and high (see text for more details). Error bars represent the standard deviation of the RV coefficients

As the number of missing row values increased (Fig. 6 and Additional file 1: Figure S3), the results obtained with the RI-MFA and MVI-MFA algorithms worsened rapidly, especially when the dimension increased, whereas the results with the MI-MFA approach were still satisfactory.

NCI-60 data analysis

To confirm the performance of our method, MI-MFA was also performed on the NCI-60 dataset. A pattern of

Table 1 Scenarios of missingness for the liver toxicity data analysis

Scenario	Number of missing rows		# cases
	Transcriptome	Clinical	
1	1	1	56
2	2	1	168
3	1	2	168
4	3	1	280
5	2	2	420
6	3	2	560
7	4	1	280

Number of missing rows inserted in each stratum of the original dataset, including both transcriptomic and clinical data, for incomplete data creation. The # cases indicate the number of possibilities of incomplete cases per stratum

missing values was created as illustrated in Table 2. One or two individuals were removed per table for all types of cancer cell lines, except for the prostate cancer line (which only had two individuals). This specific pattern would reflect a study in which a lot of rows were missing.

To compare the MFA configurations, one incomplete dataset was chosen from a large range of possibilities of incomplete datasets (6×10^{14}) according to the scenario of missingness illustrated in Table 2. We then computed our MI-MFA method on this incomplete dataset by using $M = 50$ imputed datasets. As with the liver toxicity data, RI-MFA and MVI-MFA were also performed. We chose $M = 50$ in order to achieve stable results. Figure 7 shows the divergence of the MI-, RI- and MVI-MFA configurations from the true configuration. For this specific example, the MI-MFA configuration was closest to the true configuration (Fig. 7, top-right) with a RV coefficient of 0.97, whereas the configurations obtained with RI-MFA and MVI-MFA were more distorted with RV coefficients of 0.94 and 0.87 respectively (Fig. 7, bottom).

To broaden the scope of assessment to more than a single case, 100 possible cases of incomplete datasets were chosen at random among the 6×10^{14} possibilities according to the specific missingness scenario shown in Table 2.

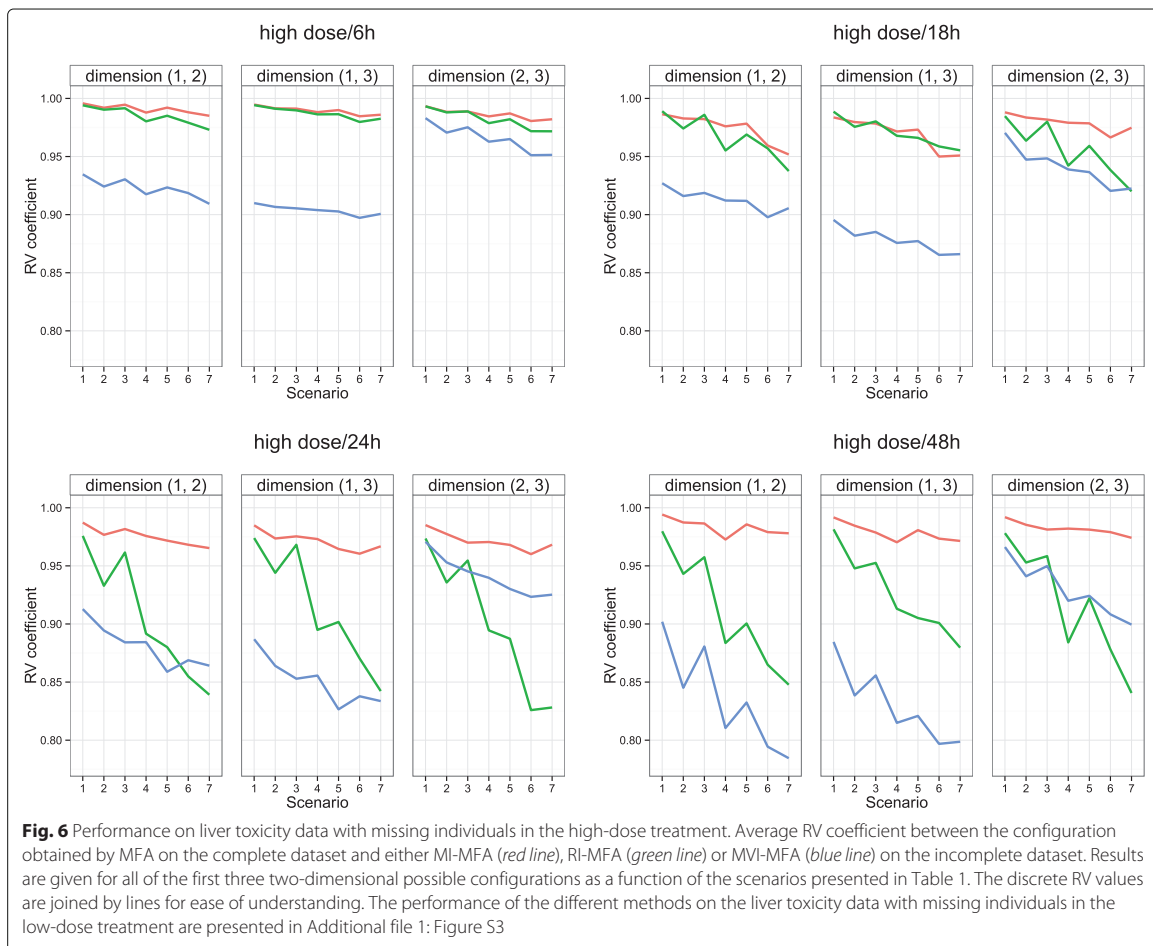


Table 2 Scenarios of missingness for the NCI-60 data analysis

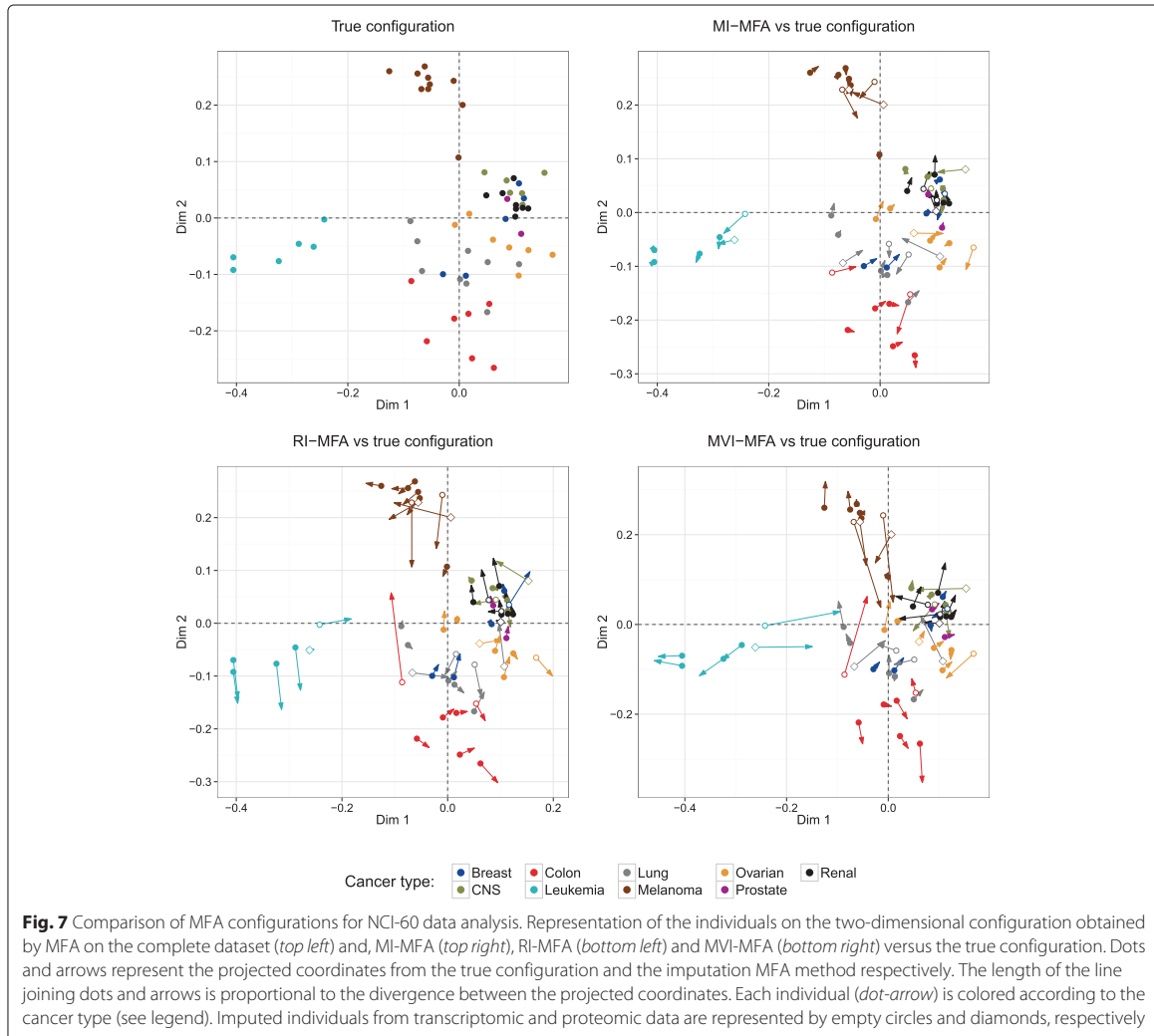
Cell line	Number of missing rows		# cases
	Transcriptome	Proteome	
Breast	1	0	5
CNS	1	1	60
Colon	2	0	42
Lung	2	2	3024
Leukemia	1	1	60
Melanoma	2	2	5040
Ovarian	1	1	84
Prostate	0	0	//
Renal	2	1	504

Number of missing rows inserted in each stratum (cell line type) of the original dataset, including both transcriptomic and proteomic data, for incomplete data creation. The # cases indicate the number of possibilities of incomplete cases per stratum

For each incomplete dataset, MI-MFA was performed using $M = 50$ imputed datasets. RI-MFA and MVI-MFA were also computed. Figure 8 shows the RV coefficient for the two-dimensional configurations as a function of each case. The results obtained with MI-MFA and RI-MFA were similar, with RV coefficients of approx. 0.97, whereas the results obtained with MVI-MFA were much further from the true configuration. Thus, even with complex patterns of missingness, the MI-MFA approach still provided satisfactory results, as did RI-MFA in this case. However, unlike RI-MFA, MI-MFA took into account the variability of missing row values, as demonstrated in the following section.

Why is it essential to evaluate uncertainty?

This question was addressed through an example using the NCI-60 dataset. A specific pattern of missing values was created. The missing rows were randomly introduced for four melanoma and two leukemia cancer lines in



the transcriptomic table. The six inserted missing rows represented 10 % of the total number of individuals. The missing rows were inserted for specific groups of individuals that contributed substantially to the construction of the first two dimensions of the MFA on the original dataset (see Fig. 7, top left). MI-MFA was performed on the incomplete dataset using $M = 50$ imputed datasets. Confidence ellipses and convex hulls were then computed from the 50 configurations projected on the compromise configuration. Figure 9 (top) shows the uncertainty due to missing rows around individuals on the compromise configuration. The use of different imputed individuals in each dataset implied slightly different configurations. Consequently, since the configurations changed, the positions of all the individuals also changed and thus all the individuals had confidence areas, even individuals

not being imputed (Fig. 9, individuals represented by filled circles). However, the greatest uncertainty occurred around the imputed individuals (Fig. 9, empty circles).

In order to highlight the importance of the uncertainty of MI-MFA configurations induced by missing rows, additional rows were removed from the transcriptomic table resulting in 30 % missing rows. We then carried out MI-MFA on the incomplete dataset using $M = 50$ imputed datasets. Figure 9 (bottom) shows the impact of the missing rows around individuals on the compromise configuration. As expected, the size of the ellipses (and convex hulls) of the additional missing individuals was increased. However, the size of the ellipses and convex hulls was not excessive even when 30 % of the rows were missing from the transcriptomic table.



Fig. 8 Performance on NCI-60 data. RV coefficient between the configuration obtained by MFA on the complete dataset and either MI-MFA (red line), RI-MFA (green line) or MVI-MFA (blue line) on the incomplete dataset. The results shown are the mean RV coefficients for the 50 two-dimensional configurations as a function of the cases according to the scenario represented in Table 2. The discrete RV values are joined by lines for ease of understanding

Discussion

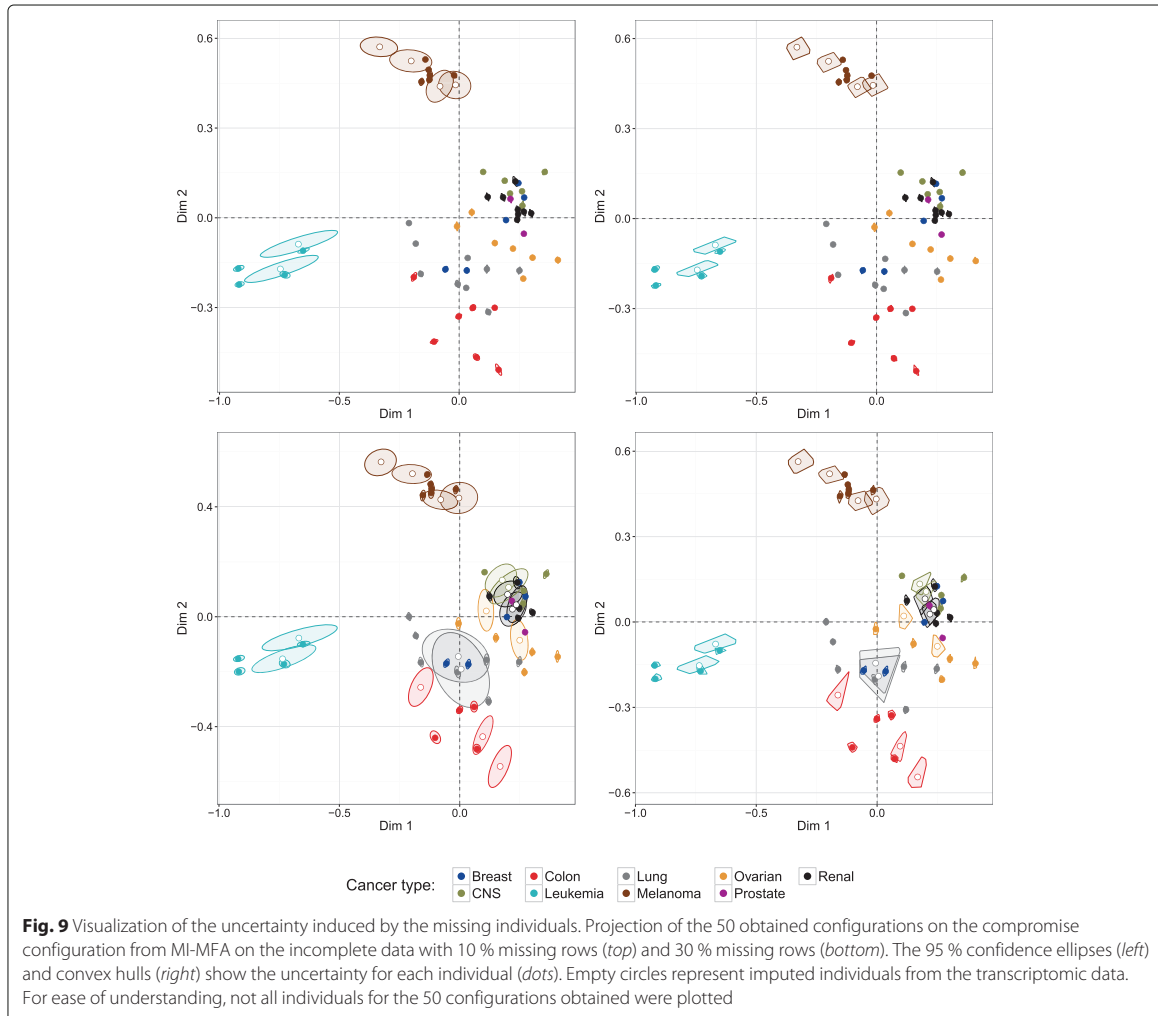
In the present paper, we propose a MI approach to handle missing row values in an MFA framework and therefore resolve one of the major issues associated with multiple omics data tables. The proposed method, which we called MI-MFA, provides point estimates of MFA configurations with a notion of uncertainty due to the missing values. The aim of our method was not to obtain the best possible estimates for the missing values, but to replace them so as to be able to estimate MFA configurations.

The MI-MFA method generates M imputed datasets from an MFA model, where multiple hot-deck imputation is used to fill in the missing values. The hot-deck approach resolves the most important limitation of other model-based techniques (such as JM and FCS) in that it can be applied to large datasets. Furthermore, other major advantages of this method are that: (1) it is not necessary to define an explicit model for the distribution of the missing values, (2) imputations tend to be realistic since they are based on observed values, and (3) it is flexible in the sense that it can preserve complex within-unit and between-variable associations. However, a weakness is that it requires good donor-recipient matches that reflect the available covariate information. Finding such good matches is not an easy task and is beyond the scope of this article [18, 19]. In an ideal framework of stratified multiple omics data tables, donor

pools would consist of the available individuals belonging to the same stratum and the same omics table as the recipient. A potential shortcoming of the method is that the donor pools might contain too few donor observations and thus introduce a risk of bias on the MFA results. Likewise, if the overall sample size is very small, then typically there are also too few potential donors. However, all imputation techniques are challenged by small sample sizes since these reduce the availability of information required to create suitable conditional statements [19].

An important aspect of our strategy is the choice of the number M of imputed datasets. As shown in the two case studies, we think that this number should be a good compromise between the need to obtain stable estimates and to avoid computation bottlenecks.

The STATIS method was proposed to combine the results of MI-MFA. As the core of MFA is a PCA, combining the results from MFA is the same as combining the results from PCA. Several procedures have previously been proposed to combine results from PCA such as the Mean Varimax Method (MVM) or Mean Correlation Matrix (MCM) approaches (as discussed in [35]). However, Van Ginkel and Kroonenberg [10, 35] demonstrated that Generalized Procrustes Analysis (GPA) was more suited to this purpose. GPA fits the PCA configurations obtained from the imputed datasets



to a single fixed reference configuration to produce the final solution, the centroid configuration (the mean of all transformed solutions). One advantage of STATIS, as compared to MVM and MCM, is that it automatically corrects for possible reflection, dilation or rotation of the different configurations. Additionally, and contrary to the GPA procedure, the STATIS algorithm is very computer-time efficient since it is a non-iterative process. Another appealing feature of STATIS is its robust properties. As this algorithm includes weights proportional to the agreement between configurations, the results do not seem to be affected by the presence of large outliers.

Two approaches have been proposed to visualize the uncertainty of the estimated MFA configurations due to missing row values: confidence ellipses and convex hulls. These graphical representations provide scientists

with considerable guidance when interpreting the significance of MFA results in a missing data framework. Indeed, ellipses and convex hull areas offer great assistance by either supporting the MFA results if they are small or suggesting that caution be exercised otherwise. It should be noted that RI-MFA (or MVI-MFA) also provides a configuration of individuals whatever the missingness pattern; however, there is no way of knowing if the results obtained are plausible and if the user can interpret the results without making any mistakes.

We have illustrated our approach by applying it to two real case studies using the liver toxicity and NCI-60 datasets. Incomplete artificial datasets with different patterns of missingness were created within these datasets. The configurations resulting from MI-, RI- and MVI-MFA were compared with the MFA configurations

of the corresponding original population (the true configuration). Performance of the methods was assessed by considering the RV coefficient with respect to the true configuration.

In the liver toxicity study, the performances of the methods were compared in two different missingness settings. First the number of missing rows in each stratum of the transcriptomic table was chosen to be low (1 row), medium (2) or high (3). Secondly, seven scenarios were created by inserting missing rows in the original dataset which included both transcriptomic and clinical tables. The overall results showed that MI-MFA clearly outperformed the RI-MFA and MVI-MFA approaches in nearly all settings.

In the NCI-60 study, we illustrated the performance of our method on complex patterns of missingness where substantial numbers of rows were missing from both tables of the NCI-60 dataset. As previously, this study showed that MI-MFA clearly performed better than MVI-MFA. We also demonstrated that RI-MFA performed better than MVI-MFA. The differences between MI-MFA and RI-MFA were small, but on average slightly in favor of our method. As the purpose of this study was also to illustrate the uncertainty of MI-MFA configurations induced by missing rows, we demonstrated how the areas of the confidence ellipses and convex hulls got larger as the number of missing rows increased.

Conclusion

We propose here a new method, MI-MFA, an extension of MFA, designed to deal with multiple tables with missing row values. MI-MFA is a useful and attractive method to estimate the coordinates of individuals for MFA configurations despite the missing rows. The study cases showed that the other proposed methods either encountered serious problems or were unable to adequately assess the accuracy due to missing data. The configurations obtained with our method were closer to the true configuration even when a significant number of individuals were missing, and thus provided better results. Moreover, the uncertainty due to the missing rows could be visualized on the compromise configuration. The software for our MI-MFA method is available in an easy-to-use code for the R statistical environment.

Additional files

Additional file 1: Supplementary figures. Figures S1–S3. (PDF 225 kb)

Additional file 2: Calculation of the total number of possible imputations in MI-MFA. (PDF 133 kb)

Additional file 3: R code implementing the MI-MFA method. R (>=3.2) is required. (PDF 224 kb)

Abbreviations

FCS: Fully conditional specification; JM: Joint modeling; MAR: Missing at random; MCAR: Missing completely at random; MFA: Multiple factor analysis; MI: Multiple imputation; MICE: Multivariate imputation by chained equations; MI-MFA: Multiple imputation multiple factor analysis; MVI-MFA: Mean variable imputation multiple factor analysis; PCA: Principal component analysis; RI-MFA: Regularized iterative multiple factor analysis; STATIS: Structuration des tableaux à trois indices de la statistique (in French)

Acknowledgements

WV is a PhD fellow supported by the INRA GA (Génétique Animale), the INRA PHASE (Physiologie Animale et Systèmes d'Élevage) and the région Languedoc-Roussillon Midi-Pyrénées. The authors would like to thank Helen Mundutéguy-Hutchings and Diana Goodfellow for the English revision. The authors thank the reviewers and the editors for their useful comments and suggestions that helped improve the quality of this paper.

Funding

WV is a PhD fellow supported by the INRA GA (Généétique Animale), the INRA PHASE (Physiologie Animale et Systèmes d'Élevage) and the Région Languedoc-Roussillon Midi-Pyrénées.

Availability of data and materials

The liver toxicity data can be accessed from the *mixOmics* R package [30] (<https://cran.r-project.org/web/packages/mixOmics>). The NCI-60 data can be downloaded from the Cellminer [33] at <https://discover.nci.nih.gov/cellminer/loadDownload.do>. The methods introduced here are implemented using the R computing environment [24]. The code, reference manual and examples are freely available as Additional file 3.

Authors' contributions

IG supervised the study. MSC, LL and PB contributed important ideas to the study. IG and WV implemented the method, carried out the simulation studies and data analysis. IG and WV drafted the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable. The liver toxicity data set were already published in [29]. The NCI-60 (National Cancer Institute) transcriptome data were published in [31] and the NCI-60 proteome data were published in [32].

Author details

¹Université de Toulouse, INRA, INPT, INP-ENVT, UMR1388, GenPhySE, F-31326 Castanet-Tolosan, France. ²Université de Toulouse INSA, UMR5219 Institut de Mathématiques, F-31077 Toulouse, France. ³INRA, UR875 Mathématiques et Informatiques Appliquées, F-31326 Castanet-Tolosan, France.

Received: 15 March 2016 Accepted: 21 September 2016

Published online: 03 October 2016

References

- Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkschlager M, Gisel A, Ballestar E, Bongcam-Rudloff E, Conesa A, Tegner J. Data integration in the era of omics: Current and future challenges. *BMC Syst Biol*. 2014;8(Suppl 2):1.
- Pigott TD. A review of methods for missing data. *Educ Res Eval*. 2001;7(4):353–83.
- Rubin DB. *Multiple Imputation for Non-Response in Surveys*. Hoboken: Wiley-Interscience; 2004.
- Nakagawa S, Freckleton RP. Missing inaction: the dangers of ignoring missing data. *Trends Ecol Evol*. 2008;23:592–6.
- Little RJA, Rubin DB. *Statistical Analysis with Missing Data*, 2nd edn. Hoboken: Wiley; 2002.

6. van de Velden M, Bijmolt THA. Generalized canonical correlation analysis of matrices with missing rows: a simulation study. *Psychometrika*. 2006;71(2):323–31.
7. González I, Déjean S, Martin PGP, Gonçalves O, Besse P, Baccini A. Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *J Biol Syst*. 2009;17(02):173–99.
8. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *Eur J Oper Res*. 2014;238(2):391–403.
9. Husson F, Josse J. Handling missing values in multiple factor analysis. *Food Qual Prefer*. 2013;30(2):77–85.
10. Josse J, Pagès J, Husson F. Multiple imputation in principal component analysis. *Adv Data Anal Classif*. 2011;5(3):231–46.
11. Josse J, Husson F. Missing values in exploratory multivariate data analysis methods. *Journal de la SFdS*. 2012;153(2):79–99.
12. Schafer JL. *Analysis of Incomplete Multivariate Data*, 1st edn. Chapman & Hall: CRC Press, Taylor & Francis Group; 1997.
13. van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. *J Stat Comput Simul*. 2006;76(12):1049–64.
14. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res*. 2007;16:219–42.
15. Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581–92.
16. Escoufier B, Pagès J. Multiple factor analysis (AFMULT package). *Comput Stat Data Anal*. 1994;18(1):121–40.
17. Kalton G, Kasprzyk D. The treatment of missing survey data. *Survey Methodol*. 1986;12:1–16.
18. Andridge RR, Little RJA. A review of hot deck imputation for survey non-response. *Int Stat Rev*. 2010;78(1):40–64.
19. Cranmer SJ, Gill J. We have to be discrete about this: A non-parametric imputation technique for missing categorical data. *British J Polit Sci*. 2013;43(02):425–49.
20. Reilly M. Data analysis using hot deck multiple imputation. *J Royal Stat Soc*. 1993;42(3):307–13.
21. Milan L, Whittaker J. Application of the parametric bootstrap to models that incorporate a singular value decomposition. *J Royal Stat Soc*. 44(1): 31–49. 1995.
22. Lavit C, Escoufier Y, Sabatier R, Traissac P. The ACT (STATIS method). *Comput Stat Data Anal*. 1994;18(1):97–119.
23. Robert P, Escoufier Y. A unifying tool for linear multivariate statistical methods: The RV coefficient. *J Royal Stat Soc*. 1976;25(3):257–65.
24. Team RC. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2015. R Foundation for Statistical Computing.
25. Lê S, Josse J, Husson F. FactoMineR: An R package for multivariate analysis. *J Stat Softw*. 2008;25(1):1–18.
26. Dray S, Dufour AB, Chessel D. The ade4 package-II: Two-table and K-table methods. *R News*. 2007;7(2):47–52.
27. Husson F, Josse J. *missMDA: Handling Missing Values With/in Multivariate Data Analysis (Principal Component Methods)*. 2014. R package version 1.7.3. <https://CRAN.R-project.org/web/packages/missMDA/missMDA.pdf>.
28. Goodall C. Procrustes methods in the statistical analysis of shape. *J Royal Stat Soc Series B (Methodol)*. 1991;53(2):285–339.
29. Bushel PR, Wolfinger RD, Gibson G. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Syst Biol*. 2007;1(15).
30. Lê Cao KA, González I, Déjean S, Rohart F, Benoit Gautier B, Monget P, Coquery J, Yao F, Lique B. *mixOmics: Omics Data Integration Project*. 2015. R package version 5.0-4. <http://CRAN.R-project.org/package=mixOmics>.
31. Liu H, D'Andrade P, Fulmer-Smentek S, Lorenzi P, Kohn KW, Weinstein JN, Pommier Y, Reinhold WC. mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Mol Cancer Ther*. 2010;9(5): 1080–91.
32. Nishizuka S, Charboneau L, Young L, Major S, Reinhold WC, Waltham M, Kourou-Mehr H, Bussey KJ, Lee JK, Espina V, Munson PJ, Petricoin E, Liotta LA, Weinstein JN. Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proc Natl Acad Sci USA*. 2003;100(24):14229–34.
33. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, Doroshow J, Pommier Y. CellMiner: A web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res*. 2012;72(14):3499–511.
34. Meng C, Kuster B, Culhane A, Gholami AM. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinforma*. 2014;15(1):162.
35. van Ginkel JR, Kroonenberg PM. Using generalized procrustes analysis for multiple imputation in principal component analysis. *J Classif*. 2014;31(2): 242–69.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit



4.2.3.3 Illustration de la méthode MI-MFA

Dans cette partie, la méthode MI-MFA va être utilisée avec nos données provenant du projet ANR Porcinet. J'ai choisi d'intégrer quatre transcriptomes issus de quatre tissus différents : muscle, foie, surrénales et sang du cordon. Toutes ces données transcriptomiques sont issues de la biopuce Agilent 60K (voir Voillet *et al.* (2014)). Les tissus muscle et foie ont été sélectionnés pour leur rôle dans l'accumulation de réserves énergétiques comme le glycogène ou les lipides (Herpin *et al.*, 2002a; Voillet *et al.*, 2014). Ces deux tissus sont sous l'influence d'hormones sécrétées par les surrénales. Le sang du cordon a quant à lui été choisi car c'est un tissu plus facilement accessible et donc intéressant pour l'obtention de possibles biomarqueurs de maturité à la naissance. Ainsi, l'objectif de cette partie est d'observer sur un même plan ces quatre transcriptomes. Pour ce faire, nous avons choisi l'AFM, capable d'analyser plus de deux groupes de variables simultanément, et en particulier notre extension, la MI-MFA, afin d'illustrer son utilité avec des données réelles et de l'associer à notre thème de recherche.

En premier lieu, comme le montre la Figure 4.6A, comparativement aux 64 initiaux, seulement 50 individus étaient présents dans les quatre transcriptomes. Les 64 individus initiaux étaient les mêmes dans tous les transcriptomes ; cependant, suite à des problèmes (d'échantillonnage ou de technique), quelques individus ont été perdus. J'ai choisi d'imputer les individus qui étaient présents dans au moins deux transcriptomes afin d'apporter le moins de variabilité non-souhaitée lors de l'estimation des composantes de l'AFM. Après normalisation des quatre tissus, les variables utilisées pour l'analyse ont été sélectionnées. Dans les quatre tissus, le nombre de variables est différent. Ceci est notamment dû au contrôle qualité des données, à la normalisation des données et à la différence d'expression entre tissus. Il est en effet possible que des variables soient exprimées dans un tissu mais pas dans un autre. Globalement, le nombre de variables présentes dans les quatre transcriptomes étant grand (Figure 4.6B), nous avons choisi de sélectionner les variables différentielles pour l'interaction entre l'âge gestationnel et le génotype fœtal afin d'éviter des problèmes computationnels lors de l'utilisation de la MI-MFA. De plus, l'étude de l'interaction entre l'âge gestationnel et le génotype fœtal est particulièrement intéressant pour étudier les différences d'expression génique entre les génotypes au cours du processus de maturation. Pour ce faire, nous avons ajusté le

modèle suivant (déjà présenté dans Voillet *et al.* (2014)) :

$$y_{ijk} = \mu + A_i + FG_j + A.FG_{ij} + S_k + \epsilon_{ijk} \quad (1)$$

où $i \in \{d90, d110\}$, $j \in \{LW, MS, LWMS, MSLW\}$, $k = 1, \dots, 18$, $S_k \sim N(0, \sigma_S^2)$ est indépendamment et identiquement distribuée (iid). S_k et ϵ_{ijk} sont mutuellement indépendants. y_{ijk} est l'expression d'une sonde étudiée, μ est la moyenne et ϵ_{ijk} est le résidu. Ce modèle contient deux effets fixes et leur interaction : A_i est l'effet de l'âge gestationnel, FG_j est l'effet du génotype foetal et $A.FG_{ij}$ est l'interaction entre ces deux effets. S_k représente l'effet aléatoire de la truie k .

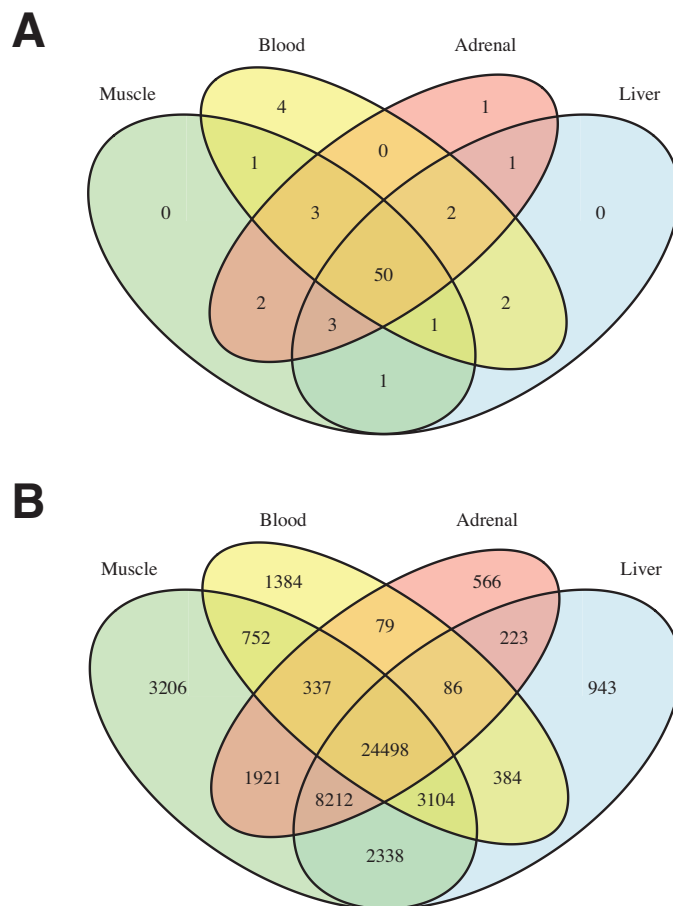


Figure 4.6 – Diagrammes de Venn des individus (A) et des variables (B) entre les quatre transcriptomes étudiés. Le transcriptome musculaire est composé de 61 individus et 44 368 sondes. Le transcriptome des surrénales est composé de 62 individus et 35 922 sondes. Le transcriptome du foie est composé de 60 individus et 39 788 sondes. Le transcriptome du sang est composé de 63 individus et 30 624 sondes.

Afin d'identifier les variables différentielles pour l'interaction entre l'âge gestationnel et le génotype fœtal, nous avons effectué un test de Fisher entre le modèle (1) et le modèle additif sans interaction $y_{ijk} = \mu + A_i + FG_j + S_k + \epsilon_{ijk}$. Une correction pour tests multiples a également été appliquée. Le tableau 4.1 montre les résultats obtenus selon les corrections utilisées. Nous observons clairement que le transcriptome musculaire comporte plus de sondes différentielles pour l'interaction entre l'âge gestationnel et le génotype fœtal que les autres tissus. Par ailleurs, les corrections de Bonferroni étaient, comme attendu, très stringentes. Dans les transcriptomes du foie et du sang, très peu de sondes ont été déclarées différentielles avec cette correction.

Ainsi, pour l'intégration avec la méthode MI-MFA, nous avons sélectionné les variables étant différentielles pour un BH à 5%. Ainsi, nous obtenons finalement 4845 variables pour le transcriptome du muscle, 258 variables pour le transcriptome du foie, 801 variables pour le transcriptome des surrénales et 47 variables pour le transcriptome du sang du cordon. Parmi les sondes sélectionnées, aucune n'est présente dans les quatre transcriptomes (Figure 4.7).

	Muscle	Foie	Surrénales	Sang
Bonferroni - 1%	173	9	32	4
Bonferroni - 5%	269	13	51	5
Benjamini-Hochberg - 1%	1890	42	178	5
Benjamini-Hochberg - 5%	4845	258	801	47

Table 4.1 – **Nombre de sondes déclarées différentielles par tissu après correction pour la multiplicité des tests.**

Nous avons donc utilisé notre méthode MI-MFA avec ces données utilisant le code R disponible en fichier additionnel de l'Article 4. Pour ce faire, nous avons choisi d'effectuer $m = 50$ imputations, nombre raisonnable pour obtenir une bonne estimation et éviter tout risque de goulot d'étranglement computationnel (voir Article 4). Les résultats obtenus sont représentés dans la Figure 4.8. Nous retrouvons sur l'axe 1 la séparation des fœtus selon l'âge gestationnel (90 et 110 jours de gestation), alors que nous observons la séparation des génotypes fœtaux à 90 jours et la séparation des génotypes fœtaux à 110 jours de gestation selon les axes 2 et 3 respectivement.

Nous pouvons également facilement observer que l'incertitude autour des individus imputés ne semble pas trop importante, ce qui est rassurant et sécurisant concernant l'analyse et l'interprétation des résultats.

Les sorties de l'AFM permettent également d'observer les projections des différents jeux de données sur la projection globale finale. Les figures 4.9 et 4.10 montrent ces projections des quatre transcriptomes (muscle, foie, surrénales et sang respectivement) par rapport à la projection finale de la MI-MFA. Nous pouvons constater que l'axe 1 sépare les deux âges gestationnels pour les quatre tables de données, alors que les axes 2 et 3 séparent les génotypes à 90 jours et 110 jours de gestation respectivement. Les transcriptomes ont plus ou moins d'influence sur la projection de ces axes. Par exemple, on peut observer que le transcriptome du sang semble participer principalement à la discrimination des fœtus purs et croisés à 90 jours de gestation (les MS sont très différents des LW à 90 jours du point de vue de ce transcriptome, Figure 4.10B), alors que le transcriptome du foie semble séparer les fœtus purs et croisés à 110 jours de gestation (les MS sont différents des LW à 110 jours du point de vue de ce transcriptome, Figure 4.9B). L'AFM permet donc d'observer les structures communes entre les différents jeux de données, en particulier le fort effet de l'âge gestationnel présent dans les quatre transcriptomes.

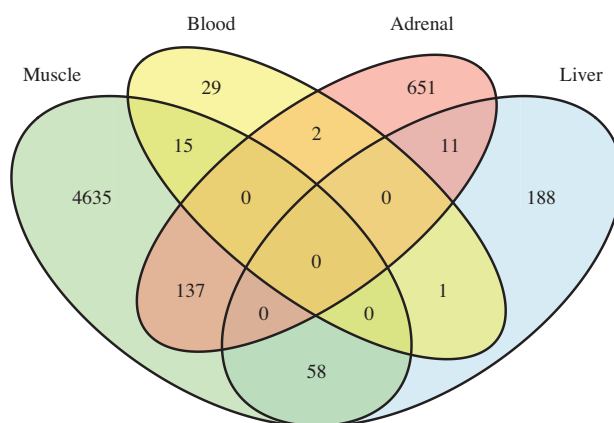


Figure 4.7 – **Diagrammes de Venn des variables sélectionnées entre les quatre transcriptomes.** Les variables sont différentielles pour l'interaction entre l'âge gestationnel et le génotype foetal avec une correction pour tests multiples (Benjamini-Hochberg 5%).

Ainsi, notre méthode permet l'imputation d'individus sans pour autant déformer les résultats, comme nous l'avons illustré dans cette partie avec les données Porcinet.

De plus, les ellipses et *convex hulls* démontrent que l'imputation des individus manquants est plutôt stable. Cette stratégie d'imputation est prometteuse car elle permet d'avoir, dans certains cas, des plans d'expériences plus équilibrés grâce à l'imputation d'individus manquants à partir des données. Une fois l'imputation de données effectuée, il serait intéressant d'effectuer des analyses d'intégration plus poussées sur ces données imputées. Par ailleurs, une des suites logiques de cette analyse pourrait être d'améliorer la sélection de variables grâce à l'imputation d'individus. Nous discuterons plus en détails les perspectives de cette méthode dans la partie discussion et perspectives.

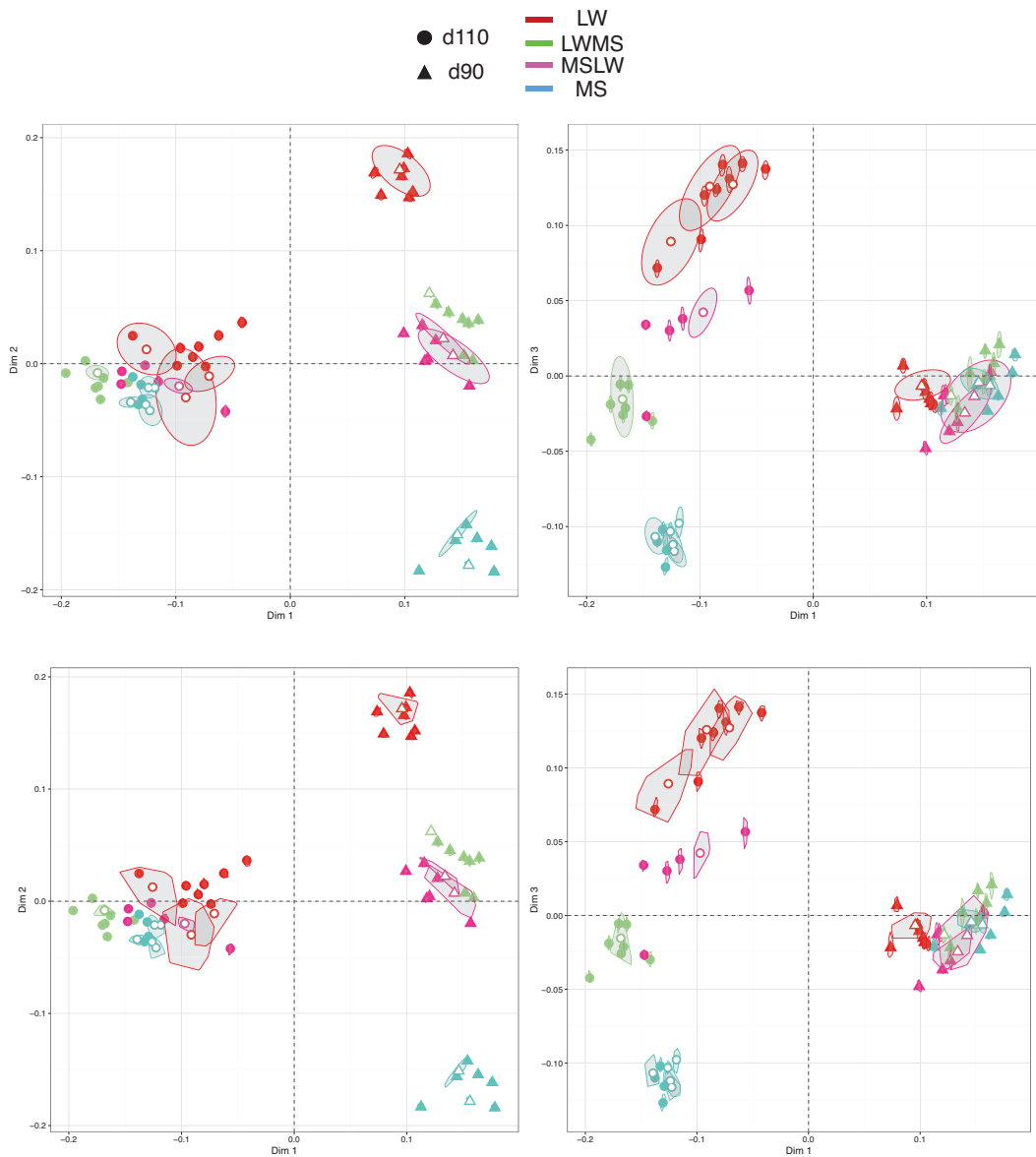


Figure 4.8 – **Visualisation de l'incertitude des individus manquants.** Projection des 50 configurations obtenues sur l'espace compromis de la MI-MFA. Les ellipses à 95% (haut) et *convex hulls* (ou aire d'incertitude) (bas) représentent l'incertitude autour de chaque individu. Les projections des individus selon l'axe 1 et 2 ou selon l'axe 1 et 3 sont respectivement représentés à gauche et à droite. Les cercles ou triangles blancs représentent les individus absents et imputés dans au moins un des quatre transcriptomes.

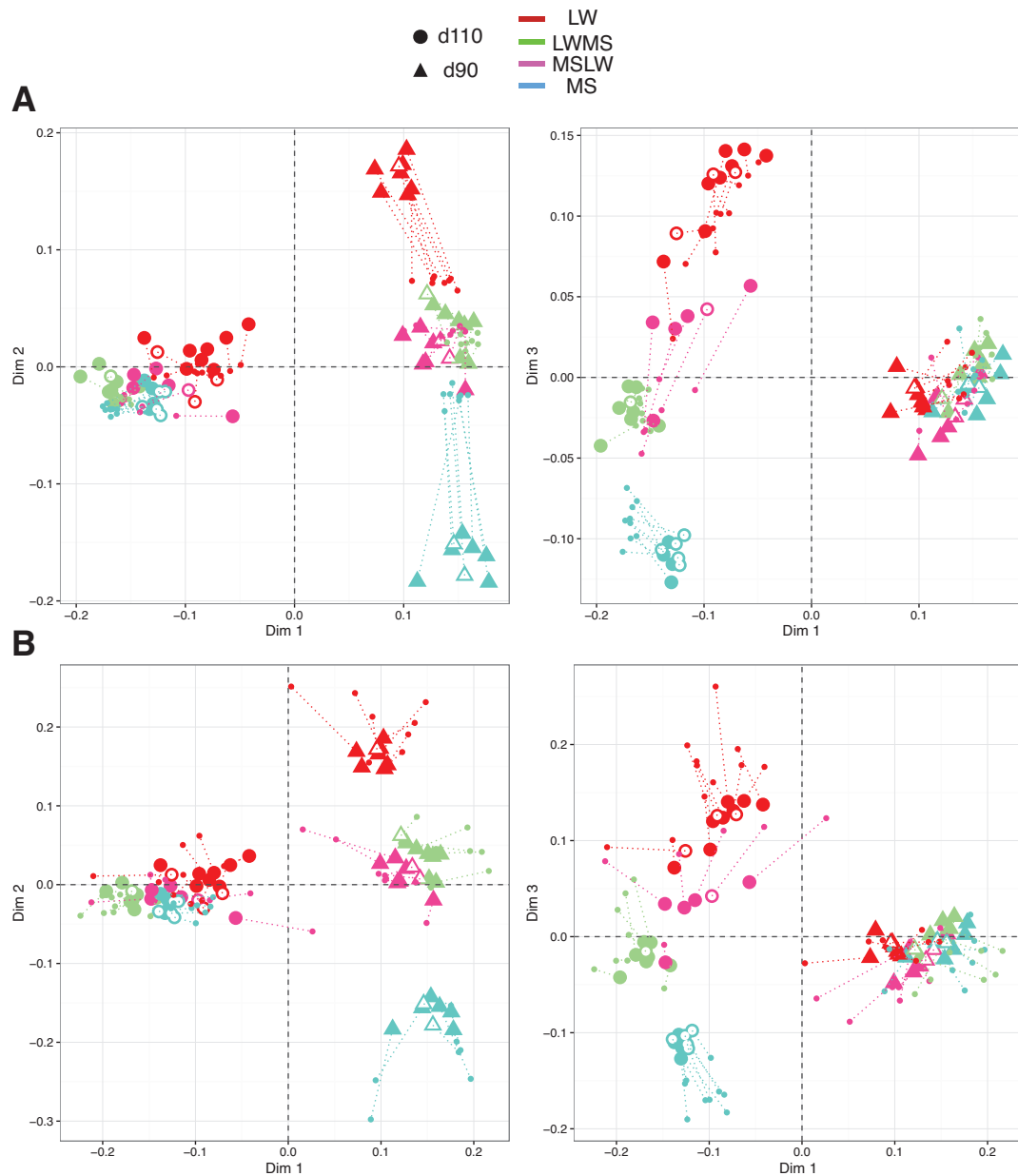


Figure 4.9 – Visualisation de la projection globale MI-MFA avec la projection partielle des données musculaire (A) ou hépatique (B). Projection globale de la MI-MFA selon l’axe 1 et 2 ou selon l’axe 1 et 3. Les projections partielles des données musculaires (A) ou hépatiques (B) sont également représentées (petites points ; reliées aux valeurs compromises). Les cercles ou triangles blancs représentent les individus absents et imputés dans au moins un des quatre transcriptomes.

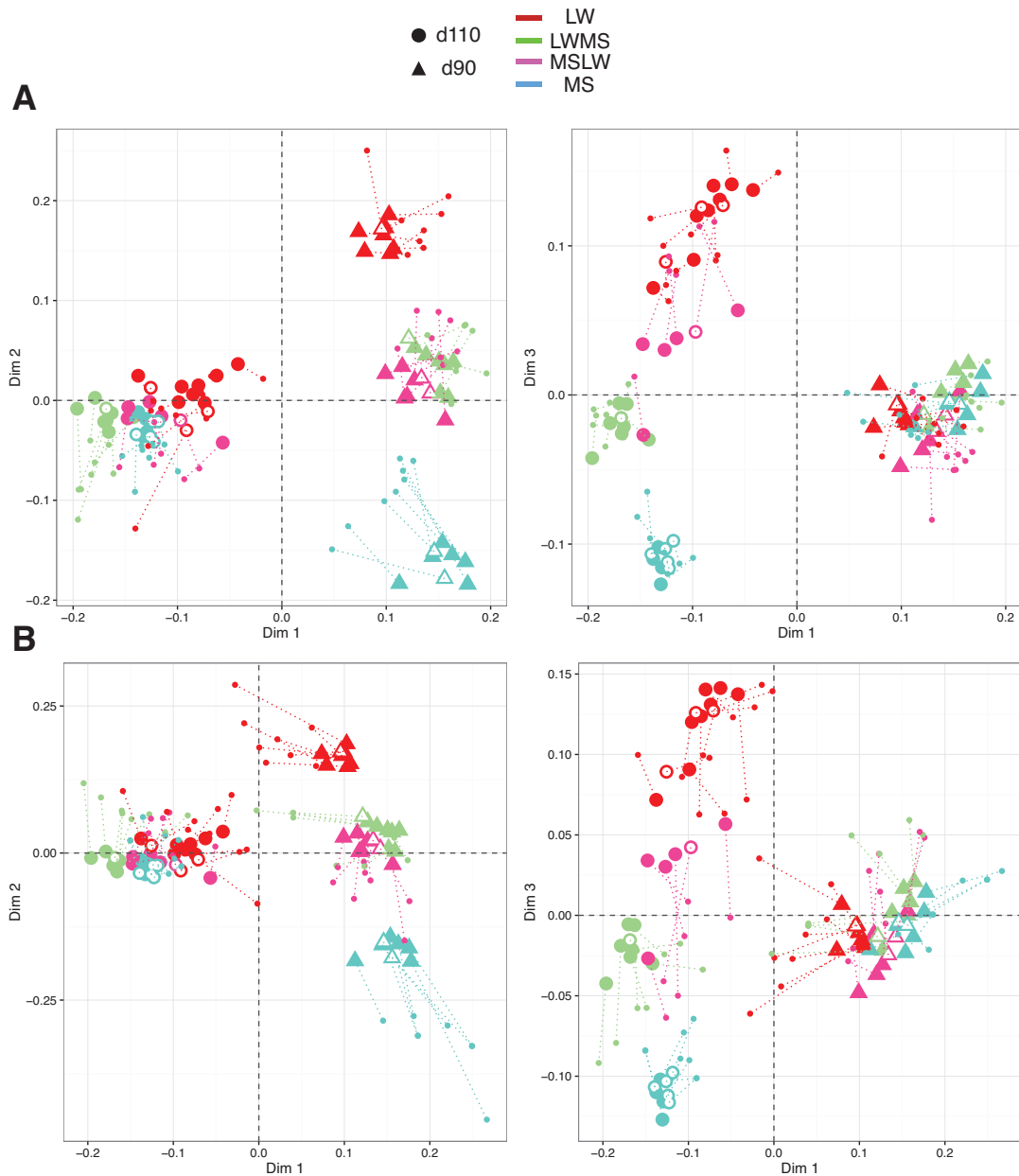


Figure 4.10 – Visualisation de la projection globale MI-MFA avec la projection partielle des données des surrénales (A) ou du sang (B). Projection globale de la MI-MFA selon l’axe 1 et 2 ou selon l’axe 1 et 3. Les projections partielles des données des surrénales (A) ou du sang (B) sont également représentées (petites points ; reliées aux valeurs compromises). Les cercles ou triangles blancs représentent les individus absents et imputés dans au moins un des quatre transcriptomes.

Chapitre 5

Discussion - Perspectives

Les objectifs de cette thèse étaient de combiner différents types de données omiques afin d'identifier les bases moléculaires et cellulaires intervenant dans la mise en place du processus de maturation au cours du dernier tiers de gestation chez le porc, et d'identifier de possibles leviers génétiques pouvant être proposés afin d'augmenter les chances de survie des porcelets à la naissance. Les chapitres 2, 3 et 4 ont fait l'objet de ce travail et la discussion biologique a pour but d'en dresser le bilan. Nous discuterons des résultats obtenus pour aller au-delà des discussions spécifiques présentées dans les publications précédentes des chapitres 2, 3 et 4. Cette thèse a également contribué à des aspects méthodologiques avec l'utilisation de différentes méthodes statistiques et le développement d'une stratégie permettant l'imputation de lignes manquantes dans le cadre de l'analyse factorielle multiple (AFM). La deuxième partie du chapitre 4 fut l'objet de ce travail. Nous allons discuter, dans la partie discussion méthodologique, des statistiques utilisées afin de répondre à notre question scientifique, mais aussi de notre méthode MI-MFA développée. Enfin, nous aborderons au cours de ces discussions quelques perspectives possibles à ces travaux.

5.1 Discussion biologique

Au cours des dernières décennies, le secteur porcin s'est principalement focalisé sur une sélection menant à plus de prolificité et de viande maigre. Cette forte sélection s'est malheureusement accompagnée d'un accroissement de la mortinatalité, associé à des problèmes économiques et éthiques. L'augmentation de la prolificité s'est adjoint d'une augmentation de l'hétérogénéité des poids des porcelets de la portée (Tribout *et al.*, 2003) et d'une diminution du poids moyen à la naissance. Cette augmentation de la variabilité semble être liée à l'amélioration de la prolificité. L'hétérogénéité de poids semble s'accroître durant le dernier tiers de gestation où les trois quarts du gain de poids de la portée *in utero* s'effectue (McPherson *et al.*, 2004), engendrant aussi une hausse de la consommation de la truie durant cette période (Samuel *et al.*, 2012). Or, le risque de mortalité d'un porcelet avant sevrage semble dépendre notamment de son poids à la naissance, mais également de la variabilité de poids dans la portée (Milligan *et al.*, 2002). Le poids est donc encore très souvent utilisé comme un critère de survie à la naissance, les porcelets de faible poids ayant peu de réserves corporelles à la naissance et étant plus sensibles au froid (Herpin *et al.*, 2002a) et au risque d'hypoxie (Herpin *et al.*, 1996). Cependant le poids à la naissance ne peut pas être considéré comme le seul critère pouvant expliquer la mortinatalité (Canario, 2006). Par exemple, les fœtus MS qui présentent très peu de mortalité à la naissance ont pourtant de plus faibles poids et sont décrits comme étant plus matures à la naissance que les fœtus LW de poids plus élevé. La bonne mise en place de plusieurs fonctions physiologiques musculaires, comme la thermorégulation ou la capacité à se déplacer, sont nécessaires pour assurer une meilleure survie du porcelet à la naissance (van der Lende *et al.*, 2001). Nous nous sommes donc intéressés à l'étude du muscle en fin de gestation (à partir de 90 jours de gestation jusqu'à la naissance) afin de comprendre et proposer de nouveaux critères de maturité. Nos travaux ont démontré que des mécanismes importants se déroulaient en fin de gestation, notamment la mise en place de métabolismes régulant les apports d'énergie pour une bonne survie à la naissance.

5.1.1 Remaniement de l'expression durant la fin de gestation

Comme cela a déjà été décrit dans la littérature (Picard *et al.*, 2002), chez le porc, le développement musculaire est clairement un phénomène biphasique dans lequel la mise en place des deux générations successives de fibres se termine vers 90 jours de gestation. Le nombre total de fibres est donc définitivement fixé vers 90 jours de gestation. Entre 90 jours de gestation et la naissance, nous avons montré que différents mécanismes métaboliques se mettaient en place. En effet, un nombre important de gènes (environ 12 000 gènes) sont impactés par l'effet de l'âge gestationnel. Nous observons ainsi un remaniement impressionnant de l'expression d'un quart des gènes exprimés dans le muscle en fin de gestation (Figure 5.1). Au travers de l'analyse plus précise des 1 120 gènes uniques impactés par l'interaction entre le stade gestationnel et le génotype, nous mettons en évidence un «switch» entre la fin du développement musculaire et la mise en place de mécanismes métaboliques, notamment le métabolisme oxydatif, ou encore le métabolisme du glycogène. Il est intéressant de noter que ce switch d'expression a été observé au niveau des données protéomiques également. Par ailleurs, nous avons aussi mis en avant que ce remaniement d'expression entre la fin de développement (cycle cellulaire et matrice extracellulaire) et la mise en place des métabolismes (oxydatif (mitochondrie), lipides et glycogène) était présent chez d'autres espèces d'intérêt agronomique (bovin et mouton), ce qui est en accord avec la littérature (Sudre *et al.*, 2003; Byrne *et al.*, 2010). En comparaison avec la vie intra-utérine, où la température du fœtus dépend entièrement de la truie, le porcelet doit être autonome pour assurer sa thermorégulation après la naissance et permettre sa survie (Herpin *et al.*, 2002a). Ainsi, les réserves énergétiques, comme le glycogène, sont à leur maximum avant la naissance (environ 10% du poids du muscle, contre 1% chez l'adulte). Chez le porc, le glycogène est stocké dans le muscle et dans le foie, ce stockage est d'autant plus important physiologiquement que le tissu adipeux brun est décrit comme étant absent chez le porcelet à la naissance, contrairement à l'homme (Trayhurn *et al.*, 1989).

En outre, il a été démontré, grâce à des expériences impliquant des transferts d'embryons, que le génotype fœtal aurait une forte importance durant le dernier tiers de gestation (Wilson *et al.*, 1998; Biensen *et al.*, 1998, 1999). En effet, il a été

observé que le poids et la surface placentaire dépendent du génotype fœtal à partir de 90 jours de gestation. Ces deux mesures étaient plus grandes lorsque les fœtus MS ou YS (Yorkshire - une lignée européenne) étaient dans un environnement utérin YS plutôt qu'un environnement utérin MS jusqu'à 90 jours de gestation. A 110 jours de gestation, la surface placentaire des fœtus YS était plus grande que celle des fœtus MS quel que soit l'environnement utérin. Il a également été démontré que la vascularisation placentaire durant la fin de gestation augmentait chez les MS, mais restait constante chez les YS. Typiquement, ces données semblent indiquer que la taille et la vascularisation du placenta sont largement déterminés par l'environnement utérin jusqu'à environ 90 jours de gestation. Après 90 jours, c'est plutôt le génotype fœtal qui influencerait le développement placentaire afin d'optimiser sa propre croissance. Ainsi, le fœtus posséderait son propre programme génétique qui déterminerait son développement, alors que l'environnement utérin empêcherait les excès de poids.

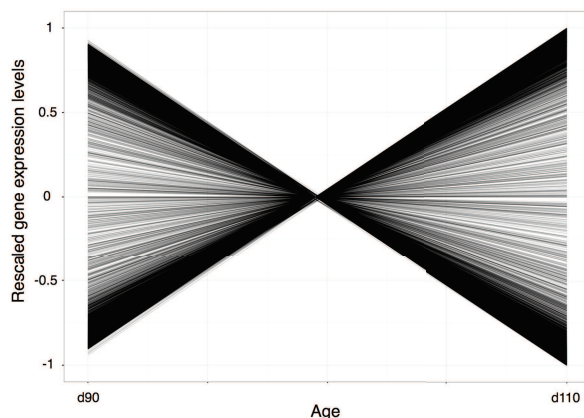


Figure 5.1 – **Remaniement de l'expression des 12 326 sondes (Bonferroni 1%) à 90 et 110 jours de gestation.** Les données d'expression des variables ont été centrées-réduites afin de pouvoir les comparer. Pour chaque sondes différentielles, les moyennes pour les quatre génotypes (MS, LW, MSLW et LWMS) ont été calculées à 90 et 110 jours de gestation.

Toutes ces constatations font également écho à la notion de conflit parental (Haig, 2000). Grâce au design expérimental de notre étude (composé de fœtus de génotype croisé), nous avons mis en évidence des gènes possiblement impactés par les génomes parentaux durant le processus de maturation : 106 gènes ont été identifiés comme étant influencés par le génome maternel, contre 366 gènes par le génome paternel (Voillet *et al.*, 2014). De Koning *et al.* (2000) ont déjà souligné que des QTLs (affectant la croissance, la composition de carcasse, la reproduction ou l'épaisseur

de lard dorsal) étaient soumis à empreinte chez le porc. Les génomes parentaux ont ainsi un impact direct sur l'expression des gènes au niveau foetal. La composante génétique pour la survie du porcelet à la naissance est une association entre la composante génétique maternelle (génotype de la mère) et la composante génétique paternelle (génotype du père). Dans notre contexte, les gènes à expression paternel sont sur-exprimés durant la fin de gestation afin de permettre au foetus d'exprimer son propre potentiel de croissance. Par ailleurs, Bischoff *et al.* (2009) ont noté, avec l'utilisation de modèles uni-parentaux, que les gènes sous influence du génome paternel ne semblaient pas essentiels à l'initiation du développement foetal, mais que leurs rôles devenaient plus importants en fin de gestation.

5.1.2 Influence de la sélection génétique sur le métabolisme énergétique musculaire

Notre étude intègre des facteurs tels que le type génétique des embryons. Ainsi, il nous a été possible d'observer les différences entre génotypes extrêmes et croisés durant le processus de maturation.

Il existe deux grands types de métabolismes énergétiques musculaires : oxydatif et glycolytique. Ils sont généralement caractérisés par la mesure d'activités enzymatiques spécifiques : (i) le métabolisme glycolytique étant mesuré par l'activité enzymatique de la lactate déshydrogénase (LDH) et (ii) le métabolisme oxydatif par l'activité enzymatique de la citrate synthase (CS, une enzyme du cycle de Krebs) ou de la béta-hydroxy-acyl-CoA déshydrogénase (Acyl-CoA, un marqueur de la béta-oxydation des acides gras) (Figure 5.2). Ces deux métabolismes permettent de classer les différents types de muscles et de fibres les composant en plusieurs catégories : les muscles de type glycolytique, les muscle de type oxydatif et les muscles de type oxydo-glycolytique (Hocquette *et al.*, 2000; Picard *et al.*, 2002). Par ailleurs, les fibres musculaires se distinguent également par leurs caractéristiques contractiles avec les fibres lentes-oxydatives (de type I), les fibres rapides oxydo-glycolytiques (de type IIa et IIx) et les fibres rapides glycolytiques (de type IIb) (Figure 5.3). Dans le contexte de cette étude, il a été observé que pour tous les génotypes, les myosines embryonnaires et périnatales diminuaient en fin de gestation, alors que les myosines rapides (IIa + IIx) et lentes (I et α -cardiaque) augmentaient. De plus, comme précédemment souligné (Lefaucheur *et al.*, 2004), nous observons

aussi une possible régulation transcriptionnelle des chaînes lourdes de myosine. Il est particulièrement intéressant d'observer que les myosines rapides sont significativement plus fortement exprimés chez les MS que chez les LW en fin de gestation, ce qui en font des marqueurs de maturité *a priori* pertinents. En outre, nos résultats ont clairement montré que des gènes et protéines, impliqués dans le métabolisme énergétique oxydatif, sont aussi plus fortement exprimés chez les MS juste avant la naissance. Ces résultats semblent donc être en accord avec ceux, plus anciens, de Bonneau *et al.* (1990), qui ont souligné des différences significatives entre les LW et MS au niveau des voies enzymatiques oxydatives, avec des valeurs plus élevées chez les MS à la naissance. Nous pouvons donc faire l'hypothèse que la sélection a un effet direct sur la mise en place des types de fibres. En effet, la comparaison entre les MS et les LW suggère que la sélection intensive pour plus de viande maigre induit un métabolisme musculaire plus glycolytique et moins oxydatif. Il est possible d'imaginer que la sélection pour plus de fibres glycolytiques ait influencé des facteurs limitants au niveau de la voie oxydative, voire déterminants pour la bonne survie à la naissance.

Dans un second temps, nous pensons également que la sélection pourrait affecter aussi la capacité musculaire à accumuler du glycogène avant la naissance. En effet, nous avons clairement observé que le glycogène était significativement plus élevé chez les MS par rapport aux LW à 110 jours de gestation. Ces résultats sont en accord avec des résultats précédents de Leenhouders *et al.* (2002), qui ont comparé des génotypes ayant des valeurs génétiques différentes pour la survie à la naissance et observé que le taux de glycogène avant la naissance était plus élevé chez les fœtus ayant des valeurs génétiques élevées pour la survie. En cohérence avec ces résultats, nous avons observé une plus forte expression de *PCK2* chez les MS par rapport aux LW. Sachant que le glycogène est un facteur déterminant pour la bonne survie à la naissance via la thermorégulation sans frisson, nous pouvons donc faire l'hypothèse que la sélection ait aussi affecté des gènes importants pour le bon fonctionnement des métabolismes impliqués dans l'accumulation et/ou la mobilisation du glycogène musculaire afin de produire de l'ATP. L'objectif de la sélection étant d'augmenter la quantité de muscle (et donc plus de fibres glycolytiques), nous aurions pu imaginer que le taux de glycogène soit plus élevé chez les LW par rapport aux MS à la naissance, les fibres glycolytiques étant plus riches en glycogène que les fibres oxydatives. En ce sens, il a d'ailleurs été observé qu'à l'abattage, le taux de glycogène est signi-

ficativement plus élevé chez les Piétrain (race hautement sélectionnée comme les LW) que les MS (Müller *et al.*, 2002). Le nombre total de fibres musculaires à 90 jours de gestation est plus faible chez les MS par rapport aux LW (Bonneau *et al.*, 1990), ceci contribue à expliquer la capacité de croissance musculaire postnatale supérieure chez les porcs LW par rapport aux porcs MS, en particulier pour les muscles de type glycolytique (plus riches en glycogène). En outre, il est important de noter que les taux de glycogène musculaire avant et après la naissance sont difficilement comparables, le taux de glycogène musculaire étant d'environ 10% à la naissance contre 1% à l'abattage. De plus, nous avons observé des différences significatives entre les LW et MS pour des gènes et protéines impliqués dans le métabolisme oxydatif mitochondrial comme ATP5A1, CKMT2 ou la navette du glycérol phosphate GPD1 (voir Figure 5.4), avec une plus forte expression chez les MS par rapport aux LW. Ainsi, une plus forte capacité à oxyder le glycogène musculaire et produire de l'ATP serait donc présente à la naissance chez les MS en comparaison aux LW.

L'oxydation des lipides produit également de l'énergie. Les acides gras à chaînes longues (sous forme de triglycérides ou sous forme libre (ou non estérifiée)) sont estérifiés pour donner des triglycérides de réserve dans le cytosol ou sont transportés vers les sites d'oxydation (mitochondries) afin d'y être catabolisés par la bêta-oxydation, le cycle de Krebs et la chaîne respiratoire (Figure 5.2). L'entrée des acides gras dans les mitochondries est sous le contrôle d'une enzyme spécifique : la carnitine palmitoyl-transférase 1 (*CPT1*). Dans le réseau de l'Article 1 (Voillet *et al.*, 2014), nous avons mis en avant toute une communauté de gènes impliqués dans le métabolisme des lipides (notamment la bêta-oxydation des acides gras). Il est intéressant d'observer que ces gènes étaient significativement plus exprimés chez les MS par rapport aux LW à 110 jours de gestation. La sélection pourrait également avoir eu un effet sur ces gènes, avec une plus faible expression chez les LW que chez les MS. Une sélection en faveur du développement musculaire, et donc d'une diminution des lipides intramusculaires, aurait eu pour conséquence d'affecter les gènes impliqués dans le métabolisme lipidique. Il semblerait donc que les MS aient un métabolisme musculaire plus oxydatif dès la fin de gestation, avec une plus forte capacité à oxyder les lipides, en complément du glycogène musculaire pour la production d'énergie nécessaire à la survie néonatale.

Cette thèse a donc permis de mieux caractériser la fin de gestation durant laquelle des

métabolismes fondamentaux se mettent en place afin de permettre une bonne survie à la naissance. Le métabolisme énergétique musculaire peut être proposé comme marqueur de maturité à la naissance. En effet, la sélection, la composante génétique et l'environnement semblent avoir des effets directs sur l'état physiologique du muscle squelettique à la naissance. Comme principale perspective, il serait intéressant d'étudier la plus forte hétérogénéité des LW pour ces métabolismes afin de proposer de nouvelles stratégies de sélection au sein de ce génotype. Il serait également opportun d'intégrer les données musculaires avec les données sanguines afin de mettre en évidence les liens possibles entre ces deux compartiments et proposer des marqueurs de maturité plus facilement mesurables (dans le sang). Typiquement, les méthodes de statistiques multidimensionnelles pourraient être utilisées, notamment notre stratégie MI-MFA.

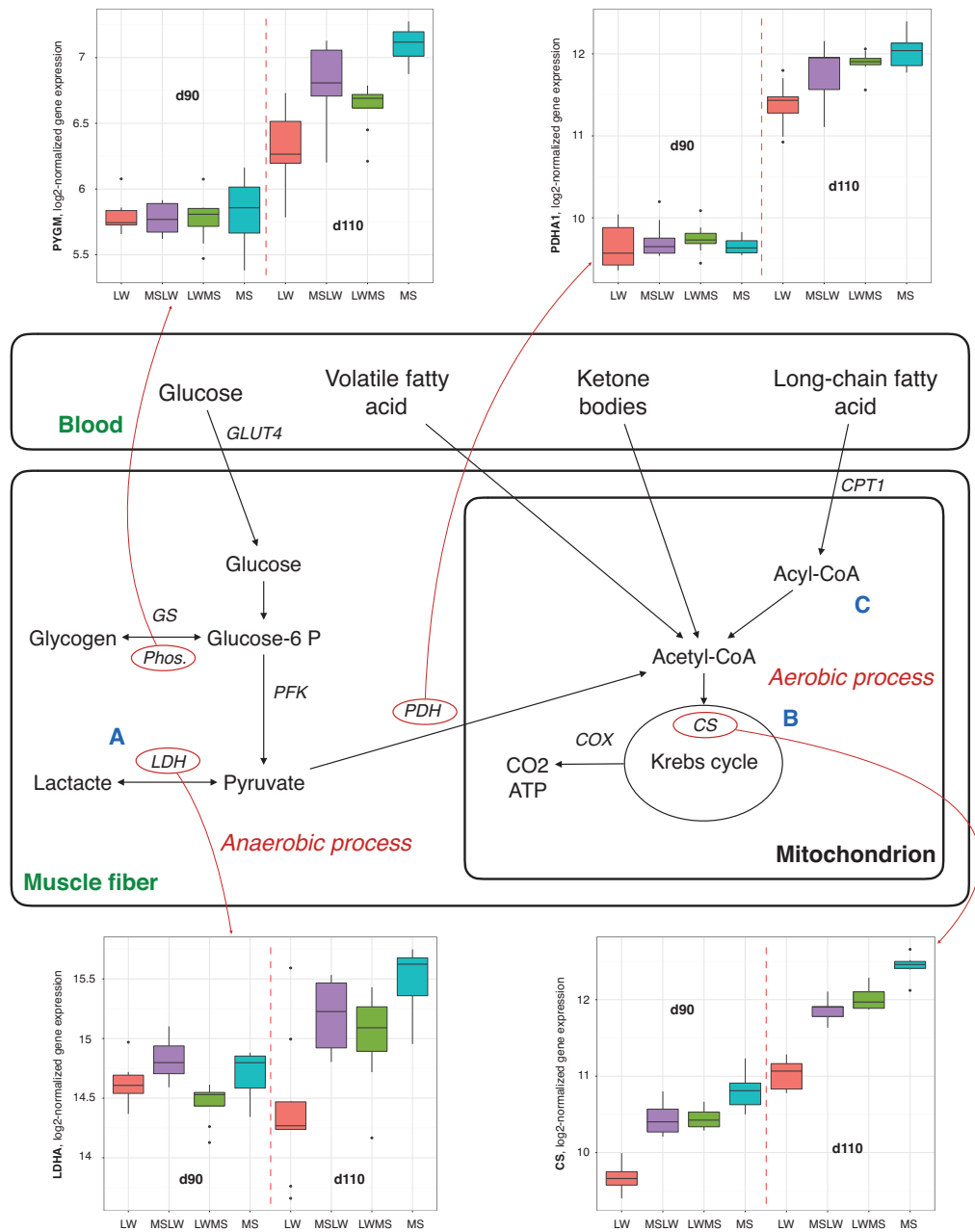


Figure 5.2 – Représentation schématique des voies des métabolismes énergétiques musculaires (inspiré de Hocquette *et al.* (2000)). Il est possible de distinguer deux voies métaboliques principales : le métabolisme glycolytique (voies A) et le métabolisme oxydatif (voies B et C). GLUT4 : Glucose transporter type 4 ; GS : Glycogène synthase ; Phos. : Phosphorylase ; PFK : Phosphofruktokinase ; LDH : Lactate déshydrogénase ; PDH : Pyruvate déshydrogénase ; COX : cytochrome c oxydase ; CS : Citrate synthase.

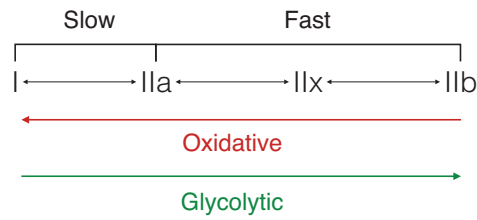


Figure 5.3 – **Caractérisation des fibres musculaires.** Deux types de fibres musculaires caractérisées par leurs caractéristiques contractiles et énergétique : glycolytique (rapide) ou oxydative (lente).

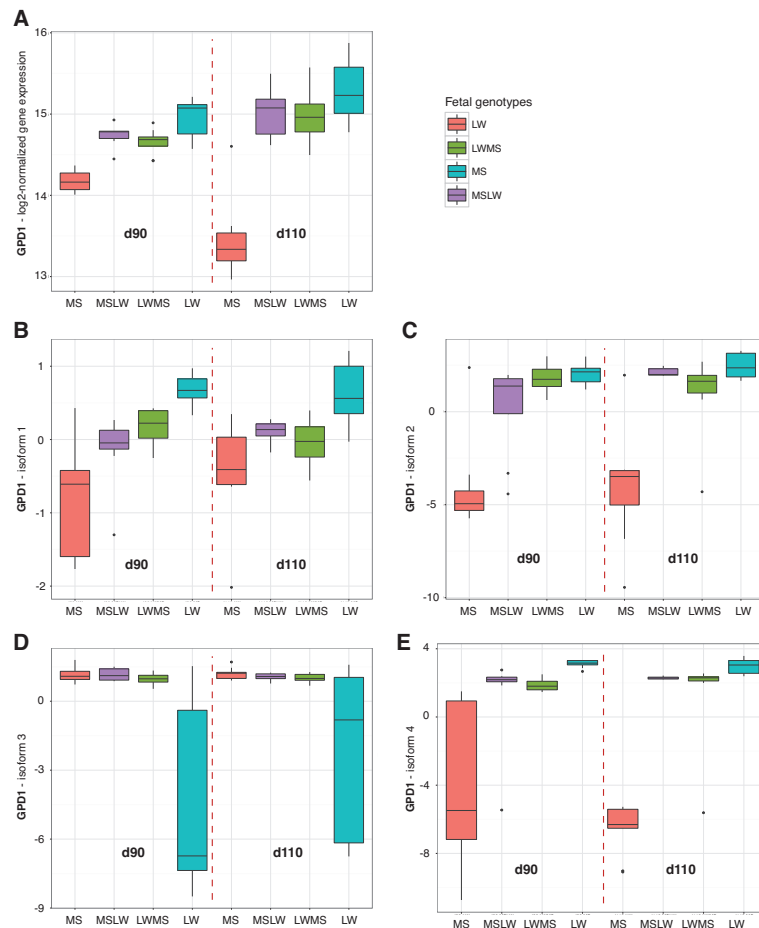


Figure 5.4 – **Expression (A, niveau génique ; B-E, niveau protéique) de GPD1.** A) Expression génique de GPD1. B) Expression protéique de l'isoforme 1 de GPD1. C) Expression protéique de l'isoforme 2 de GPD1. D) Expression protéique de l'isoforme 3 de GPD1. E) Expression protéique de l'isoforme 4 de GPD1.

5.2 Discussion méthodologique

5.2.1 Le choix de la méthode d'inférence des réseaux protéiques

Lors de nos travaux d'intégration des données protéiques, transcriptomiques et phénotypiques, l'inférence des réseaux protéiques fut une étape déterminante en vue des analyses qui ont suivi. En effet, le choix de la méthode d'inférence dépendait totalement de ce que nous souhaitions observer pour modéliser les interactions protéiques. Dans ce travail, nous avons finalement choisi d'inférer uniquement les liens directs conditionnels entre protéines, avec l'utilisation de l'algorithme PCIT (Reverter & Chan, 2008), afin d'explorer les possibles liens de régulations entre protéines. Comme décrit dans la section 3.2.3, le principe de cette méthode est d'appliquer à une matrice de corrélation un coefficient de corrélation partielle combiné à la théorie de l'information afin d'identifier les arêtes significatives. Cette stratégie PCIT a déjà été utilisée pour des données transcriptomiques (Hudson *et al.*, 2009; Pérez-Montarelo *et al.*, 2012). Ici, nous avons choisi de l'utiliser pour nos données protéiques.

Les méthodes de type *Gaussian Graphical Models* (GGM), décrites dans la section 3.2.2.2, auraient également pu être utilisées dans le but d'observer uniquement les liens directs. Cependant, ces stratégies sont difficilement réalisables lorsque (i) le nombre de variables p est beaucoup plus grand que le nombre d'individus n , ou (ii) lorsque les variables sont très corrélées entre elles. Dans nos données, de forts effets de l'âge gestationnel et du génotype foetal ont été observés, donnant ainsi des variables fortement corrélées entre elles. Typiquement, comme le montre la Figure 5.5, nous pouvons facilement constater des corrélations relativement élevées (positives ou négatives - distribution bimodale) entre les spots protéiques (113 spots conservés) à 110 jours de gestation. Les méthodes GGM sont adaptées aux cas où il y a une seule distribution normale ($y_i \sim \mathcal{N}(\mu, \Sigma)$), contrairement au cas, comme notre étude, où il y a plusieurs conditions expérimentales ($y_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ où i est une condition expérimentale). Dans cette étude, nous avons observé un très fort effet de l'âge gestationnel, mais aussi un assez fort effet du génotype foetal. En utilisant les GGM, nous avons donc plus de risques de capter le différentiel entre ces deux stades (Figure 5.6). Ainsi, pour éviter ces forts effets, nous avons choisi

d'inférer des réseaux dans les deux âges gestationnels respectivement. Nous aurions également pu comparer des réseaux inférés pour chaque génotype dans chaque stade gestationnel, mais le nombre d'observations était trop faible (environ 8 individus par conditions). Nous avons d'ailleurs essayé d'utiliser ces méthodes GGM sur nos données protéiques, mais il était difficile d'obtenir une bonne estimation des corrélations partielles, en particulier lorsque l'on étudiait les corrélations avec les phénotypes d'intérêt. L'algorithme PCIT est, quant à lui, probablement moins sensible au fait que les données ne soient pas normales mais bimodales. En revanche, si les données ne sont pas normales, l'absence de lien entre deux variables sous-entend une dé-corrélation conditionnelle et non une indépendance entre ces variables.

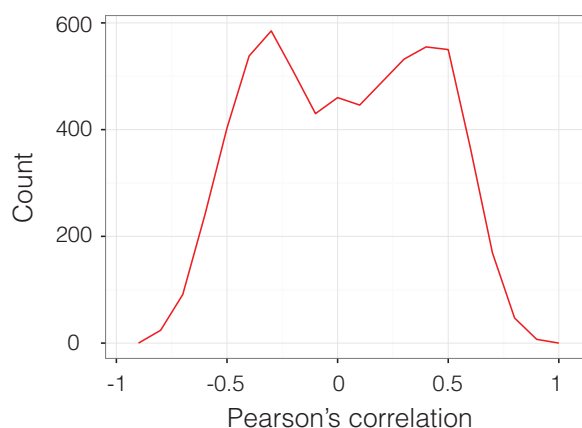


Figure 5.5 – **Distribution des corrélations entre variables protéiques à 110 jours de gestation.** 113 spots protéiques ont été conservés (voir Voillet *et al.* (2016b)).

Les outils de statistique spatiale ont déjà été utilisés pour analyser les relations entre des réseaux et des phénotypes (Villa-Vialaneix *et al.*, 2013). Cependant, de ce cas, les données étaient composées d'individus appartenant à une seule et même condition. L'estimation de la corrélation par GGM était donc adaptée à ces données. Dans notre cas, comme précédemment souligné, l'estimation de la matrice de corrélation partielle n'était pas bonne, ceci entraînant des soucis lors de l'analyse spatiale. En effet, les outils de statistique spatiale nécessitent l'utilisation d'une matrice de corrélation ou de corrélation partielle en fonction du réseau. C'est donc aussi pour cette raison que nous avons choisi d'utiliser l'algorithme PCIT afin d'inférer nos réseaux protéiques.

En outre, l'un des autres principaux avantages de la méthode PCIT est l'utilisation de seuils locaux, et non d'un seuil global unique, pour déterminer les arêtes significatives (voir section 3.2.3). Le choix d'un seuil approprié, au-dessus duquel les interactions entre variables sont considérées comme pertinentes, demeure un défi majeur pour la plupart des méthodes d'inférence de réseaux. La qualification d'une arête en arête significative peut être arbitraire (par exemple, avec un seuil choisi sur la matrice de similarité), ou dépendre d'une statistique associée (Davidson *et al.*, 2001; Carter *et al.*, 2004; Zhang & Horvath, 2005). Ici, avec la méthode PCIT, nous avons choisi d'utiliser des seuils locaux plutôt qu'un seuil global, comme nous l'avons fait dans l'Article 1 (Voillet *et al.*, 2014). En effet, le réseau décrit dans cet article a été obtenu par utilisation d'un seuil global très élevé sur la matrice de corrélation ($|r| \geq 0.97$ - l'effet de l'âge gestationnel étant très fort), afin d'avoir un réseau final facilement analysable. L'algorithme PCIT fournit donc des seuils locaux pour chaque triplet de variables via le calcul de la moyenne des ratios entre les corrélations conditionnelles (partielles) et les corrélations directes entre ces variables. Ces seuils locaux sont utilisés afin de déterminer la significativité des arêtes (pour prendre en compte les estimations inexactes des différences entre les deux valeurs de corrélations proches). L'utilisation de ces seuils locaux a plusieurs avantages : (i) ces seuils sont imputés directement et intrinsèquement à partir des données (et donc sans interventions extérieures) et (ii) ces seuils prennent en compte les différences de corrélations entre toutes les paires de variables, plutôt que d'être appliqués à l'ensemble des données.

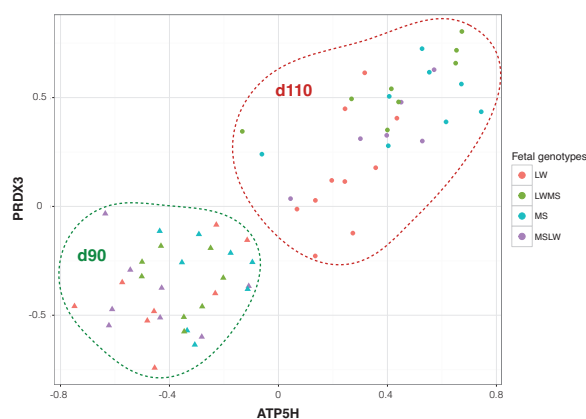


Figure 5.6 – **Exemple de corrélation entre deux variables protéiques.** Les spots protéiques ATP5H et PRDX3 proviennent des 113 spots protéiques conservés (voir Voillet *et al.* (2016b)).

5.2.2 Intégration des données protéomiques et transcriptomiques : autres stratégies

A l'heure actuelle, l'intégration des données protéomiques et transcriptomiques reste un véritable challenge à cause de caractéristiques différentes entre ces deux types de données, telles que les annotations incomplètes ou encore les différences entre isoformes (Haider & Pal, 2013). En outre, l'expression des gènes est très rarement corrélée avec l'abondance en protéines. En effet, des modifications post-transcriptionnelles, comme l'efficacité de la traduction en protéine, l'épissage alternatif ou le repliement des protéines, peuvent se dérouler. Dans notre étude, nous avons choisi d'intégrer les données en plusieurs étapes : (i) inférence et clustering de réseaux protéiques, (ii) association de phénotypes d'intérêts avec ces sous-réseaux (clusters) protéiques et (iii) étude de la corrélation entre expression génique et expression protéique. Ainsi, les réseaux protéiques sont considérés comme le centre de notre analyse auxquels nous avons intégré d'autres types d'information (phénotype et transcriptome). L'utilisation des réseaux est assez courante en biologie et génétique ; elle permet d'observer les liens d'interactions entre les variables et de mettre en avant des clusters (ou variables) importants. L'étude des liens entre les sous-réseaux protéiques et les phénotypes d'intérêts a été effectuée grâce à l'utilisation d'outils de la statistique spatiale. Il existe d'autres stratégies pour déterminer le lien entre une communauté et un phénotype d'intérêts, comme la corrélation entre la première composante principale de la matrice d'expression du cluster (considérée comme une représentation moyenne de l'expression du cluster) avec le phénotype (Langfelder & Horvath, 2008). Cependant, les outils de statistiques spatiales ont l'avantage de prendre en compte la structure du réseau (et donc l'importance de certaines protéines (*hubs*)) pour l'étude de la corrélation avec un phénotype.

L'intégration avec les données transcriptomiques a été plus complexe. En effet, le choix de la liste de gènes à utiliser pour l'analyse (à partir des 44 368 sondes) fut fastidieux. Quelques méthodes comme la sPLS ou la sCCA ont été utilisées afin d'obtenir une liste de départ de 100 à 200 gènes, mais ceux-ci étaient trop corrélés entre eux. Nous avons donc choisi de nous limiter à l'étude des gènes codant pour les protéines identifiées. Par ailleurs, une autre approche réseau a été développée au cours de cette thèse (non-présentée dans ce manuscrit) correspondant à la stratégie

développée par Montastier *et al.* (2015). L'idée de cette méthode est d'intégrer des réseaux inférés à partir des différents types de données (ici, protéome, transcriptome et phénotypes), et des réseaux bipartites obtenus à partir de méthodes multivariées (comme l'ACC). Cette méthode a l'avantage de représenter tous les liens possibles entre les différentes strates d'expression. Les résultats obtenus étaient cependant difficilement interprétables : le clustering ne permettant pas d'identifier des fonctions biologiques particulières associées aux communautés. Par exemple, la Figure 5.7 montre le réseau obtenu à 110 jours de gestation. Nous pouvons observer que les protéines et les gènes partageaient peu de liens (quelques nœuds englobaient la majorité des liens), le clustering ne permettait donc pas toujours d'obtenir des clusters avec des protéines, des gènes et des phénotypes, ce qui était l'objectif de cette méthode. Nous pouvons également discuter de l'étude de la corrélation partielle entre les différents types de variables (gènes et protéines). Cette stratégie a été essayée (non représentée), mais très peu de liens entre les différents types de données étaient présents. Ceci peut être dû à une plus grande similitude des profils transcriptomiques entre eux qu'avec les profils protéiques.

5.2.3 Perspectives de la MI-MFA

L'émergence des biotechnologies a notamment permis d'obtenir des données multiples pour un même ensemble d'individus, offrant la possibilité d'effectuer des études tout génome. Aujourd'hui, le problème de lignes manquantes (typiquement, l'absence d'un individu pour tout un groupe de variables) dans les méthodes multidimensionnelles est encore très peu étudié. Nous avons pu observer que notre méthode, appelée MI-MFA, était très prometteuse, avec une bonne estimation des composantes de l'AFM pour les individus ayant des données manquantes pour des groupes de variables (voir section 4.2.3.3). En premier lieu, notre stratégie génère m jeux de données imputés grâce à la procédure d'imputation multiple de type *hot-deck* (les valeurs manquantes sont imputées par des valeurs dites «semblables» choisies au hasard à partir d'un pool) (Andridge & Little, 2010). L'avantage de cette méthode d'imputation est sa capacité à être appliquée à des données de grandes dimensions. Cependant, elle nécessite de fortes similitudes entre les valeurs manquantes et les valeurs remplaçantes. Dans le cadre de données ayant plusieurs conditions expérimentales (comme Porcinet avec 8 conditions : deux âges gestationnels associés à quatre génotypes), le substitut peut logiquement faire partie de la même

condition que l'individu manquant. Toutefois, une des limitations de l'algorithme est le nombre d'échantillons disponibles par condition. En effet, un nombre trop faible de valeurs de substitution pour l'imputation peut engendrer des biais au niveau des résultats de l'AFM. En outre, si le nombre global d'échantillons est faible, alors il a potentiellement peu de valeurs remplaçantes. Toutes les méthodes d'imputation font face au défi des données ayant peu d'échantillons, ceci réduisant la quantité d'information nécessaire afin de construire un pool de valeurs remplaçantes convenables. Ainsi, dans notre fonction R fournie à l'utilisation, nous avons choisi d'avoir toujours au maximum le même nombre de valeurs présentes et de valeurs absentes pour chaque condition. Ensuite, m AFM sont effectuées sur ces m jeux de données imputées différents. Pour finir, notre stratégie utilise la méthode STATIS (Structuration des tableaux à trois indices de la statistique) afin de combiner les m composantes des m AFM obtenues (Lavit *et al.*, 1994). En outre, nous avons également proposé une façon d'observer l'incertitude autour des individus imputés par des représentations graphiques fournissant aux utilisateurs de la méthode des orientations considérables lors de l'interprétation des résultats de l'AFM dans le cadre de données manquantes. En effet, ces ellipses ou *convex hulls* sont d'une grande aide, soit en confirmant les résultats de l'AFM si elles sont petites, soit en préconisant la prudence si elles sont grandes. Dans les analyses effectuées (avec les données Porcinet, liver toxicity et NCI-60), quelques secondes étaient nécessaires afin d'obtenir nos résultats. Toutefois, les performances computationnelles (temps de calcul) de notre algorithme n'ont pas encore été testées. La fonction R va être parallélisée afin d'améliorer ses performances (parallélisation des AFM sur les données imputées).

Pour le moment, notre stratégie impute uniquement les composantes de l'AFM. Il serait donc pertinent d'imputer les valeurs manquantes et de les fournir aux utilisateurs, l'imputation étant une étape préliminaire importante avant l'analyse et l'intégration des données. En partant des valeurs de projection imputées, une approche itérative, comme développée dans la RI-MFA de Husson & Josse (2013), pourrait être proposée. Une autre perspective serait d'exploiter notre méthodologie (imputation multiple et combinaison des résultats avec STATIS) à d'autres méthodes d'analyses multidimensionnelles comme la rGCCA. La rGCCA est une version généralisée de l'ACC permettant l'analyse de plus de deux jeux de données (Tenenhaus *et al.*, 2014). Cette stratégie permet d'étudier les relations entre différents jeux

de données, mais aussi d'identifier des sous-groupes de variables pour chaque jeu de données ayant des relations avec d'autres sous-groupes de variables. Comme pour l'ACC et la PLS, une version *sparse* a aussi été développée afin de permettre la sélection de variables (Tenenhaus *et al.*, 2014). Nous avons choisi de ne pas décrire cette méthode dans la section 4.2.2 car elle n'a pas été employée au cours de cette thèse. Toutefois, elle constitue une perspective intéressante à notre travail. Cette méthode est présentée dans le document de Lê Cao (2014). Une fois cette stratégie adaptée, comme évoquée dans la section 4.2.3.3, il serait important d'observer l'effet de l'imputation des lignes manquantes sur la sélection de variables, par exemple sur la sPLS. Nous pourrions imaginer que l'imputation améliorerait la sélection de variables, rendant ainsi les résultats possiblement plus fiables.

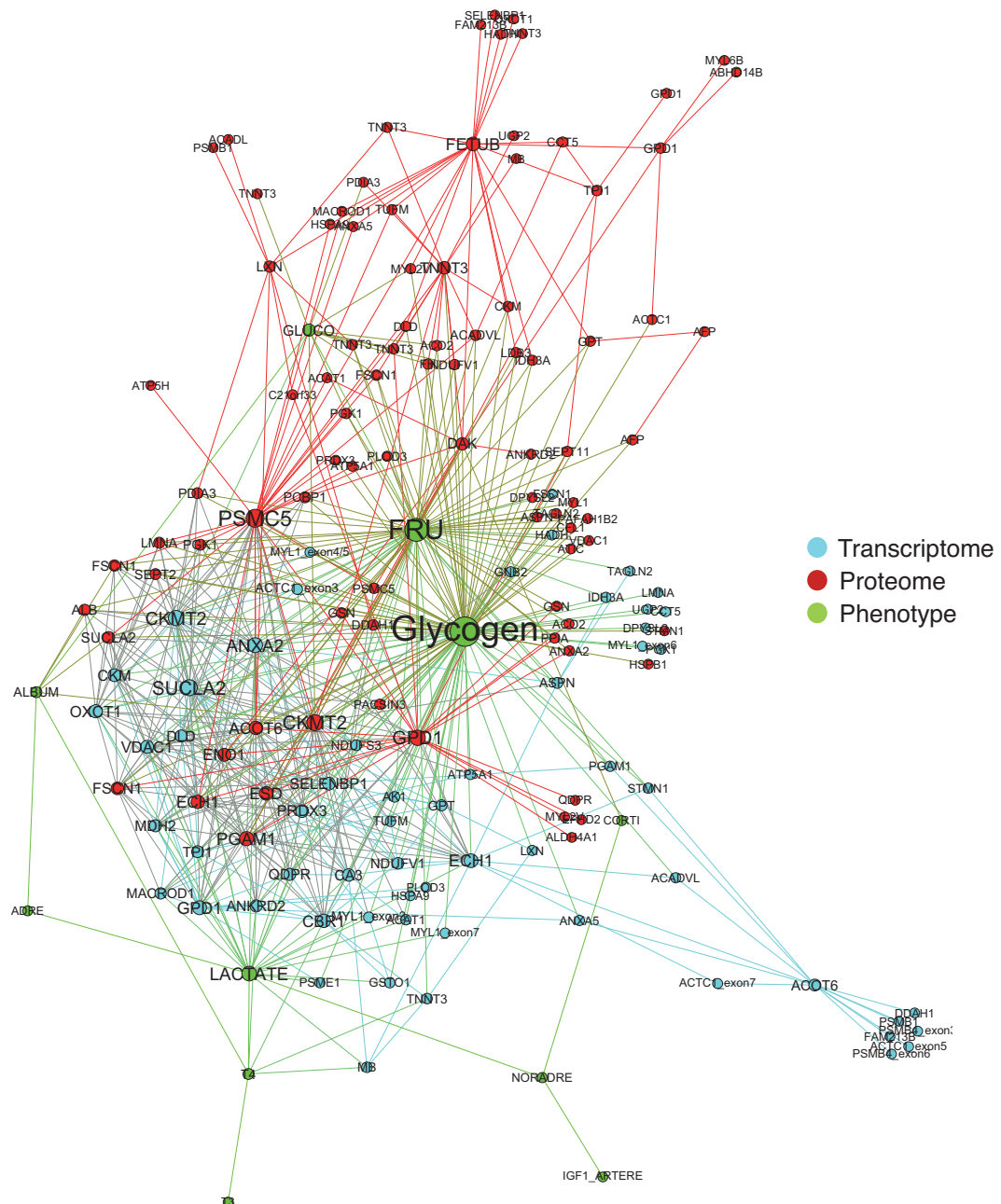


Figure 5.7 – Exemple de réseau global (protéome, transcriptome et phénotypes) à 110 jours de gestation. La stratégie publiée par Montastier *et al.* (2015) a été utilisée avec nos données. Tous les liens entre les différentes strates d'expression sont présents. Des réseaux sont inférés à partir des différents types de données (ici, protéome, transcriptome et phénotypes), et des réseaux bipartites sont aussi inférés à partir de méthodes multivariées (comme l'ACC). Tous ces réseaux sont ensuite fusionnés en un seul global.

Liste des articles et communications

Articles publiées

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

V. Voillet, P. Besse, L. Liaubet, M. San Cristobal and I. González. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics* **17**:402 (2016).

Articles soumis ou en préparation

V. Voillet, M. San Cristobal, L. Liaubet and L. Lefaucheur. Integrated network analysis of proteomic, transcriptomic and phenotypic data highlights muscle late fetal maturation process. *In preparation* (2016).

V. Voillet, M. San Cristobal, L. Liaubet and B.P Dalrymple. Comparative transcriptomic analysis of fetal muscle development in cattle, sheep and pigs. *In preparation* (2016).

Communications

Orales

V. Voillet, M. San Cristobal, L. Lefaucheur and L. Liaubet. Integrated network multi-omics approach highlights muscle late fetal maturation process. 35th International Society for Animal Genetics Conference. July 23 – 27th, 2016. Salt Lake City, USA.

V. Voillet, M. San Cristobal, L. Liaubet, P. Besse and I. González. Recovering missing individual block information in multiblock multiple factor analysis. The 1st missData Conference. June 18 – 19th, 2015. Rennes, France.

V. Voillet. Systems biology of piglet maturity. Séminaire des Thésards du Département Génétique Animale. May 21st – 22nd, 2015. La Rochelle, France.

V. Voillet, M. San Cristobal, Y. Lippi, P.G.P Martin, N. Iannuccelli, Y. Billon, L. Canario and L. Liaubet. Muscle transcriptomic investigation of late fetal development and determinism of maturity at birth in two extreme breeds: meishan and large white. 10th World Congress of Genetics Applied to Livestock Production. August 17th – 22nd, 2014. Vancouver, Canada.

V. Voillet, L. Lefaucheur, L. Canario, M.C. Père, A. Paris, Y. Billon, C. Canlet, N. Iannuccelli, Y. Lippi, P.G.P. Martin, H. Quesnel, M. San Cristobal and L. Liaubet. Systems biology of piglet maturity with focus on muscle metabolism. 34th International Society for Animal Genetics Conference. July 28th - August 1st, 2014. X'ian, China.

Posters

V. Voillet, M. San Cristobal, L. Lefaucheur and L. Liaubet. Integrated network multi-omics approach highlights muscle late fetal maturation process. 35th International Society for Animal Genetics Conference. July 23 – 27th, 2016. Salt Lake City, USA.

V. Voillet, P. Besse, L. Liaubet, M. San Cristobal and I. González. Handling missing rows in multi-omics data integration: multiple imputation

in multiple factor analysis framework. 5th Rencontres R. June 22 – 24th, 2016. Toulouse, France.

V. Voillet, M. San Cristobal, L. Liaubet, P. Besse and I. González. Recovering missing individual block information in multiblock multiple factor analysis. The 1st missData Conference. June 18 – 19th, 2015. Rennes, France.

V. Voillet, M. San Cristobal, M.C. Père, Y. Billon, L. Canario and L. Liaubet. Integrated muscle analysis of transcriptome and proteome to explain piglet maturity at birth. 23rd International Plant & Animal Genome. January 10 – 14th, 2015. San Diego, USA.

V. Voillet, M. San Cristobal, P.G.P Martin, Y. Lippi, L. Lefaucheur and L. Liaubet. Integrative approach to define biomarkers of piglet maturity. 13th European Conference on Computational Biology. September 7–10th, 2014. Strasbourg, France.

V. Voillet, M. San Cristobal, and L. Liaubet. Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity. Bioinformatics & Biostatistics regional workshop. June 13th, 2014. Toulouse, France.

V. Voillet. Integrative biology to identify early biomarkers for neonatal survival. Séminaire des Thésards du Département Génétique Animale. April 23 – 24th, 2014. Jouy-en-Josas, France.

Prix et bourses

ISAG bursary to attend the 35th International Society for Animal Genetics Conference (Salt Lake City, July 23 – 27th, 2016) for an oral presentation and one poster: Integrated network multi-omics approach highlights muscle late fetal maturation process.

Mobility Grant from INP Toulouse: 4 months in CSIRO in the lab of B.P. Dalrymple (Brisbane, Australia) from August to December 2015.

Références

- AITTOKALLIO, T. & SCHWIKOWSKI, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, **7**, 243–255.
- ALTELAAR, A.F., MUNOZ, J. & HECK, A.J. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, **14**, 35–48.
- ANDRIDGE, R.R. & LITTLE, R.J.A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, **78**, 40–64.
- BARABÁSI, A.L. & OLTVAI, Z.N. (2004). Network biology: understanding the cell’s functional organization. *Nature Reviews Genetics*, **5**, 101–113.
- BARABÁSI, A.L., GULBAHCE, N. & LOSCALZO, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, **12**, 56–68.
- BAXTER, E.M., JARVIS, S., D’EATH, R.B., ROSS, D.W., ROBSON, S.K., FARISH, M., NEVISON, I.M., LAWRENCE, A.B. & EDWARDS, S.A. (2008). Investigating the behavioural and physiological indicators of neonatal survival in pigs. *Theriogenology*, **69**, 773–783.
- BEN HUR, A. & NOBLE, W.S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21**, 38–46.
- BIENSEN, N.J., WILSON, M.E. & FORD, S.P. (1998). The impact of either a Meishan or Yorkshire uterus on Meishan or Yorkshire fetal and placental development to days 70, 90, and 110 of gestation. *Journal of Animal Science*, **76**, 2169–2176.
- BIENSEN, N.J., WILSON, M.E. & FORD, S.P. (1999). The impacts of uterine environment and fetal genotype on conceptus size and placental vascularity during late gestation in pigs. *Journal of Animal Science*, **77**, 954–959.

- BISCHOFF, S.R., TSAI, S., HARDISON, N., MOTSINGER-REIF, A.A., FREKING, B.A., NONNEMAN, D., ROHRER, G. & PIEDRAHITA, J.A. (2009). Characterization of conserved and nonconserved imprinted genes in swine. *Biology Reproduction*, **81**, 906–920.
- BLACKSTOCK, W.P. & WEIR, M.P. (1999). Proteomics: quantitative and physical mapping of cellular proteins. *Trends in Biotechnology*, **17**, 121–127.
- BOLSTAD, B.M., IRIZARRY, R., ÅSTRAND, M. & SPEED, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
- BONNEAU, M., MOUROT, J., NOBLET, L., LEFAUCHEUR, L. & BIDANEL, J.P. (1990). Tissue development in Meishan pigs: Muscle and fat development and metabolism and growth regulation by somatotropic hormone. *Chinese Pig Symp*, 202–213.
- BONNETA, L. (2008). Epigenomics: The new tool in studying complex diseases. *Nature Education*, **1**, 178.
- BREKER, M. & SCHULDINER, M. (2014). The emergence of proteome-wide technologies: systematic analysis of proteins comes of age. *Nature Reviews Molecular Cell Biology*, **15**, 453–464.
- BUTTE, A., TAMAYO, P., SLONIM, D., GOLUB, T. & KOHANE, I. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, **97**, 12182.
- BYRNE, K., VUOCOLO, T., GONDRO, C., WHITE, J.D., COCKET, N.E., HADFIELD, T., BIDWELL, C.A., WADDELL, J.N. & TELLAM, R.L. (2010). A gene network switch enhances the oxidative capacity of ovine skeletal muscle during late fetal development. *BMC Genomics*, **11**, 378.
- CANARIO, L. (2006). *Genetic aspects of piglet mortality at birth and in early suckling period: relationships with sow maternal abilities and piglet vitality*. Ph.D. thesis, Institut National Agronomique Paris-Grignon, Paris.

- CARTER, S., BRECHBÜHLER, C., GRIFFIN, M. & BOND, A. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, **20**, 2242–2250.
- CHEN, J. & YUAN, B. (2006). Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, **22**, 2283–2290.
- CIVELEK, M. & LUSIS, A.J. (2014). Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, **15**, 34–48.
- CULHANE, A.C., PERRIÈRE, G. & HIGGINS, D.G. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, **4**, 59.
- DAVIDSON, G., WYLIE, B. & BOYACK, K. (2001). Cluster stability and the use of noise in interpretation of clustering. In *IEEE Information Visualization*, no. 23-30 in 1.
- DE KONING, D.J., RATTINK, A.P., HARLIZIUS, B., VAN ARENDONK, J.A.M., BRASCAMP, E.W. & GROENEN, M.A.M. (2000). Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proceedings of the National Academy of Sciences*, **97**, 7947–7950.
- DE TAYRAC, M., LÊ, S., AUBRY, M., MOSSER, J. & HUSSON, F. (2009). Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple factor analysis approach. *BMC Genomics*, **10**, 32.
- DEMPSTER, A. (1972). Covariance selection. *Biometrics*, **1**, 157–175.
- DUMAS, M.E., CANLET, C., DEBRAUWER, L., MARTIN, P. & PARIS, A. (2005). Selection of biomarkers by a multivariate statistical processing of composite metabolomic data sets using multiple factor analysis. *Journal of Proteome Research*, **4**, 1485–1492.
- DYKSTRA, R. (1970). Establishing the positive definiteness of the sample covariance matrix. *The Annals of Mathematical Statistics*, **41**, 2153–2154.
- EDWARDS, S. (2011). Knowledge synthesis: animal health and welfare in organic pig production. Tech. rep., Newcastle University, UK.

- ESCOFIER, B. & PAGÈS, J. (1988-1998). *Analyses factorielles simples et multiples ; objectifs, méthodes et interprétation*. Dunod, Paris.
- FAGAN, A., CULHANE, A.C. & HIGGINS, D.G. (2007). A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics*, **7**, 2162–2171.
- FLINTOFT, L. (2010). Complex disease: Epigenomics gets personal. *Nature Reviews Genetics*, **11**, 746–747.
- FORTUNATO, S. (2010). Community detection in graphs. *Physics Reports*, **486**, 75–174.
- FOWDEN, A.L., COMLINE, R.S. & SILVER, M. (1985). The effects of cortisol on the concentration of glycogen in different tissues in the chronically catheterized fetal pig. *Quarterly Journal of Experimental Physiology*, **70**, 23–35.
- FOXCROFT, G.R., DIXON, W.T., NOVAK, S., PUTMAN, C.T., TOWN, S.C. & VINSKY, M.D. (2006). The biological basis for prenatal programming of postnatal performance in pigs. *Journal of Animal Science*, **84**, 105–112.
- FRIEDMAN, N., LINIAL, M., NACHMAN, I. & PE'ER, D. (2000). Using bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.
- FRIEDMAN, N., HASTIE, T. & TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.
- GONZÁLEZ, I., DÉJEAN, S., MARTIN, P.G., GONÇALVES, O., BESSE, P. & BACCINI, A. (2008). Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems*, **17**, 173–199.
- GONZÁLEZ, I., LÊ CAO, K., DAVIS, M. & DÉJEAN, S. (2012). Visualising associations between paired omics data sets. *BioData Mining*, **5**, 19.
- GRIFFITHS, W.J. & WANG, Y. (2008). Mass spectrometry: from proteomics to metabolomics and lipidomics. *Chemical Society Reviews*, **38**, 1882–1896.

- GU, Q., NAGARAJ, S.H., HUDSON, N.J., DALRYMPLE, B.P. & REVERTER, A. (2011). Genome-wide patterns of promoter sharing and co-expression in bovine skeletal muscle. *BMC Genomics*, **12**, 23.
- GUIMERÀ, R. & AMARAL, L.A.N. (2005). Functional cartography of complex metabolic networks. *Nature*, **433**, 895–900.
- GUO, B., GREENWOOD, P.L., CAFE, L.M., ZHOU, G., ZHANG, W. & DALRYMPLE, B.P. (2015). Transcriptome analysis of cattle muscle identifies potential markers for skeletal muscle growth rate and major cell types. *BMC Genomics*, **16**, 177.
- GYGI, S.P., ROCHON, Y., FRANZA, B.R. & AEBERSOLD, R. (1999). Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, **19**, 1720–1730.
- HAIDER, S. & PAL, R. (2013). Integrated analysis of transcriptomic and proteomic data. *Current Genomics*, **14**, 91–110.
- HAIG, D. (2000). The kinship theory of genomic imprinting. *Annual Review of Ecology, Evolution, and Systematics*, **31**, 9–32.
- HAMID, J.S., HU, P., ROSLIN, N.M., LING, V., GREENWOOD, C.M.T. & BEYENE, J. (2009). Data integration in genetics and genomics: methods and challenges. *Human Genomics and Proteomics*, 869093.
- HAWKINS, R.D., HON, G.C. & REN, B. (2010). Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, **11**, 476–486.
- HERPIN, P., LE DIVIDICH, J. & 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.
- HERPIN, P., LE DIVIDICH, J., HULIN, J.C., FILLAUT, M., DE MARCO, F. & BERTIN, R. (1996). Effects of the level of asphyxia during delivery on viability at birth and early postnatal vitality of newborn pigs. *Journal of Animal Science*, **74**, 2067–2075.
- HERPIN, P., DAMON, M. & LE DIVIDICH, J. (2002a). Development of thermoregulation and neonatal survival in pigs. *Livestock Production Science*, **78**, 20.

- HERPIN, P., LOSSEC, G., SCHMIDT, I., COHEN-ADAD, F., DUCHAMP, C., LEFAUCHEUR, L., GOGLIA, F. & LANNI, A. (2002b). Effect of age and cold exposure on morphofunctional characteristics of skeletal muscle in neonatal pigs. *European Journal of Physiology*, **444**, 610–618.
- HOCQUETTE, J.F., ORTIGUES-MARTY, I., DAMON, M., HERPIN, P. & GEAY, Y. (2000). Métabolisme énergétique des muscles squelettiques chez les animaux producteurs de viandes. *INRA Production Animale*, **13**, 185–200.
- HOLZINGER, E.R. & RITCHIE, M.D. (2012). Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics*, **13**, 213–222.
- HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika*, **28**, 321–377.
- HUDSON, N., REVERTER, A., WANG, Y., GREENWOOD, P. & DALRYMPLE, B.P. (2009). Inferring the transcriptional landscape of bovine skeletal muscle by integrating co-expression networks. *PLOS One*, **4**, e7249.
- HUSSON, F. & JOSSE, J. (2013). Handling missing values in multiple factor analysis. *Food Quality and Preference*, **30**, 77–85.
- JOHNSTONE, I.M. & TITTERINGTON, D.M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society*, **367**, 4237–4253.
- JOSSE, J. & HUSSON, F. (2012). Missing values in exploratory multivariate data analysis methods. *Journal de la SFdS*, **153**, 79–99.
- KADARMIDEEN, H.N., VON ROHR, P. & JANSSE, L.L. (2006). From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. *Mammalian Genome*, **17**, 548–564.
- KIRKDEN, R.D., BROOM, D.M. & ANDERSON, I.L. (2013). Piglet mortality: Management solutions. *American Society of Animal Science*, **91**, 3361–3389.
- KOKETSU, Y., TAKENOBU, S. & NAKAMURA, R. (2006). Prewaning mortality risks and recorded causes of death associated with production factors in swine

- breeding herds in Japan. *The Journal of Veterinary Medical Science*, **68**, 821–826.
- LANCKRIET, G.R.G., DE BIE, T., CRISTIANINI, N., JORDAN, M.I. & NOBLE, S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, **20**, 2626–2635.
- LANGFELDER, P. & HORVATH, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
- LAVIT, C., ESCOUFIER, Y., SABATIER, R. & TRAISSAC, P. (1994). The ACT (STATIS method). *Computational Statistics & Data Analysis*, **18**, 97–119.
- LAWN, J.E., COUSENS, S. & ZUPAN, J. (2005). 4 million neonatal deaths: When? where? why? *The Lancet*, **365**, 891–900.
- LÊ CAO, K.A., ROSSOUW, D., ROBERT-GRANIÉ, C. & BESSE, P. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, **7**, Article 35.
- LÊ CAO, K.A., MARTIN, P.G., ROBERT-GRANIÉ, C. & BESSE, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, **10**, 34.
- LEBRET, B., LEFAUCHEUR, L. & MOUROT, J. (1999). La qualité de la viande de porc. *INRA Production Animale*, **12**, 11–28.
- LEENHOUWERS, J.I., DE ALMEIDA JUNIOR, C.A., KNOL, E.F. & VAN DER LENDE, T. (2001). Progress of farrowing and early postnatal pig behavior in relation to genetic merit for pig survival. *Journal of Animal Science*, **79**, 1416–1422.
- LEENHOUWERS, J.I., KNOL, E.F., DE GROOT, P.N., VOS, H. & 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.
- LEFAUCHEUR, L., ECOLAN, P., LOSSEC, G., GABILLARD, J.C., BUTLER-BROWNE, G.S. & HERPIN, P. (2001). Influence of early postnatal cold exposure on myofiber maturation in pig skeletal muscle. *Journal of Muscle Research and Cell Motility*, **22**, 439–452.

- LEFAUCHEUR, L., MILAN, D., ECOLAN, P. & LE CALLENNEC, C. (2004). Myosin heavy chain composition of different skeletal muscles in Large White and Meishan pigs. *Journal of Animal Science*, **82**, 1931–1941.
- LIOLIOS, K., TAVERNARAKIS, N., HUGENHOLTZ, P. & KYRPIDES, N.C. (2006). The genomes on line database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Research*, **34**, D332–D334.
- MCPHERSON, R.L., JI, F., WU, G., BLANTON, J.R. & KIM, S.W. (2004). Growth and compositional changes of fetal tissues in pigs. *Journal of Animal Science*, **82**, 2534–2540.
- MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High dimensional graphs and variable selection with lasso. *Annals of Statistics*, **34**, 1436–1462.
- MELLOR, D.J. & COCKBURN, F. (1986). A comparison of energy metabolism in the new-born infant, piglet and lamb. *Quarterly Journal of Experimental Physiology*, **71**, 361–379.
- MENG, C., KUSTER, B., CULHANE, A.C. & GHOLAMI, A.M. (2014). A multi-variate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, **15**, 162.
- METZKER, M.L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, **11**, 31–46.
- MILLER, D.R., BLACHE, D., JACKSON, R.B., DOWNIE, E.F. & ROCHE, J.R. (2010). Metabolic maturity at birth and neonate lamb survival: association among maternal factors, litter size, lamb birth weight, and plasma metabolic and endocrine factors on survival and behavior. *Journal of Animal Science*, **88**, 581–593.
- MILLIGAN, B.N., FRASER, D. & KRAMER, D.L. (2002). Within-litter birth weight variation in the domestic pig and its relation to preweaning death, weight gain, and variation in weaning weights. *Livestock Production Science*, **76**, 181–191.
- MONTASTIER, E., VILLA-VIALANEIX, N., CASPAR-BAUGUIL, S., HLAVATY, P., TVRZICKA, E., GONZALEZ, I., SARIS, W.H.M., LANGIN, D., KUNESOVA, M. & VIGUERIE, N. (2015). System model network for adipose tissue signatures

- related to weight changes in response to calorie restriction and subsequent weight maintenance. *PLOS Computational Biology*, **11**, e1004047.
- MÜLLER, E., RUTTEN, L., MOSER, G., REINER, G., BARTENSCHLAGER, H. & GELDERMANN, H. (2002). Fibre structure and metabolites in *M. longissimus dorsi* of Wild Boar, Pietrain and Meishan pigs as well as their crossbred generations. *Journal of Animal Breeding and Genetics*, **119**, 125–137.
- NEWMAN, M. & GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, **69**, 026113.
- PAGÈS, J. (2002). Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. *Revue de Statistique appliquée*, **L**, 5–37.
- PANZARDI, A., BERNARDI, M.L., MELLAGI, A.P., BIERHALS, T., BORTOLOZZO, F.P. & WENTZ, I. (2013). Newborn piglet traits associated with survival and growth performance until weaning. *Preventive Veterinary Medicine*, **110**, 206–213.
- PENG, J., WANG, P., ZHOU, N. & ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, **104**, 735–746.
- PÉREZ-MONTARELO, D., HUDSON, N., FERNÁNDEZ, A., RAMAYO-CALDAS, Y., DALRYMPLE, B.P. & REVERTER, A. (2012). Porcine tissue-specific regulatory networks derived from meta-analysis of the transcriptome. *PLOS One*, **7**, e46159.
- PICARD, B., LEFAUCHEUR, L., BERRI, C. & DUCLOS, J. (2002). Muscle fibre ontogenesis in farm animal species. *Reproduction Nutrition Development*, **42**, 415–431.
- QUESNEL, H., BROSSARD, L., VALANCOGNE, A. & QUINIOU, N. (2008). Influence of some sow characteristics on within-litter variation of piglet birth weight. *Animal*, **2**, 1842–1849.
- REIF, D.M., WHITE, B.C. & MOORE, J.H. (2004). Integrated analysis of genetic, genomic and proteomic data. *Expert Review of Proteomics*, **1**, 67–75.

- REVERTER, A. & CHAN, E.K.F. (2008). Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*, **24**, 2491–2497.
- RITCHIE, M.D., HOLZINGER, E.R., LI, R., PENDERGRASS, S.A. & KIM, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, **16**, 85–97.
- RUBAKHIN, S.S., ROMANOVA, E.V., NEMES, P. & SWEEDLER, J.V. (2011). Profiling metabolites and peptides in single cells. *Nature Methods Supplement*, **8**, S20–S29.
- SABIDÓ, E., SELEVSEK, N. & AEBERSOLD, R. (2012). Mass spectrometry-based proteomics for systems biology. *Current Opinion in Biotechnology*, **23**, 591–597.
- SAMUEL, R.S., MOEHN, S., PENCHARZ, P.B. & BALL, R.O. (2012). Dietary lysine requirement of sows increases in late gestation. *Journal of Animal Science*, **90**, 4896–4904.
- SCHÄFER, J. & STRIMMER, K. (2005a). An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–764.
- SCHÄFER, J. & STRIMMER, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implication for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**, 1–32.
- SCUTARI, M. (2010). Learning bayesian networks with the nb-learn R package. *Journal of Statistical Software*, **35**, 1–22.
- SERENIUS, T. & STALDER, K.J. (2007). Length of productive life of crossbred sows is affected by farm management, leg conformation, sow’s own prolificacy, sow’s origin parity and genetics. *Animal*, **1**, 745–750.
- SHULAEV, V. (2006). Metabolomics technology and bioinformatics. *Briefings in Bioinformatics*, **7**, 128–139.
- SIEBERTS, S.K. & SCHADT, E.E. (2007). Moving toward a system genetics view of disease. *Mammalian Genome*, **18**, 389–401.

- SINGH, A., GAUTIER, B., SHANNON, C.P., VACHER, M., ROHART, F., TEBBUTT, S.J. & LÊ CAO, K.A. (2016). Diablo - an integrative, multi-omics, multivariate method for multi-group classification. *Submitted*.
- SPIRIN, V. & MIRNY, L.A. (2003). Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, **100**, 12123–12128.
- STRANGE, T., ASK, B. & NIELSEN, B. (2013). Genetic parameters of the piglet mortality traits stillborn, weak at birth, starvation, crushing, and miscellaneous in crossbred pigs. *Journal of Animal Science*, **91**, 1562–1569.
- STUART, J.M., SEGAL, E., KOLLER, D. & KIM, S.K. (2003). A gene-coexpression network for global discovery of conserved genetics modules. *Science*, **302**, 249–255.
- SUDRE, K., LEROUS, C., PIÉTU, G., CASSAR-MALEK, I., PETIT, E., LISTRAT, A., AUFRAY, C., PICARD, B., MARTIN, P. & HOCQUETTE, J.F. (2003). Transcriptome analysis of two bovine muscles during ontogenesis. *Journal of Biochemistry*, **133**, 745–756.
- TENENHAUS, A., PHILIPPE, C., GUILLEMOT, V., LÊ CAO, K.A., GRILL, J. & FROUIN, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*, kxu001.
- THIELE, I. & PALSSON, B.O. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocol*, **5**, 93–121.
- TRAYHURN, P., TEMPLE, N.J. & VAN AERDE, J. (1989). Evidence from immunoblotting studies on uncoupling protein that brown adipose tissue is not present in the domestic pig. *Canadian Journal of Physiology and Pharmacology*, **67**, 1480–1485.
- TRIBOUT, T., CARITEZ, J.C., GOGUÉ, J., GRUANT, J., BILLON, Y., BOUFAUD, M., LAGANT, H., LE DIVIDICH, J., THOMAS, F., QUESNEL, H., GUÉBLEZ, R. & BIDANEL, J.P. (2003). Estimation, par utilisation de semence congelée, du progrès génétique réalisé en France entre 1977 et 1998 dans la race porcine Large White : résultats pour quelques caractères de reproduction femelle. *Journée de la Recherche Porcine*, **35**, 285–292.

- TUCHSCHERER, M., PUPPE, B., TUCHSCHERER, A. & TIEMANN, U. (2000). Early identification of neonates at risk: traits of newborn piglets with respect to survival. *Theriogenology*, **54**, 371–388.
- UNICEF (2014). Levels & trends in child mortality: report 2014. Tech. rep., UNICEF.
- VAN DE VELDEN, M. & BIJMOLT, T.H.A. (2006). Generalized canonical correlation analysis of matrices with missing rows: A simulation study. *Psychometrika*, **71**, 323–331.
- VAN DE VELDEN, M. & TAKANE, Y. (2011). Generalized canonical correlation analysis with missing values. *Computational Statistics*, **27**, 551–571.
- VAN DER LENDE, T., KNOL, E.F. & LEENHOUWERS, J.I. (2001). Prenatal development as a predisposing factor for perinatal losses in pigs. *Reproduction Supplement*, **58**, 247–261.
- VAN IERSEL, M.P., SOKOLOVIĆ, M., LENAERTS, K., KUTMON, M., BOUWMAN, F.G., LAMERS, W.H., MARIMAN, E.C.M. & EVELO, C.T. (2014). Integrated visualization of a multi-omics study of starvation in mouse intestine. *Journal of Integrative Bioinformatics*, **11**, 235.
- VERZELEN, N. (2010). Adaptive estimation of covariance matrices via Cholesky decomposition. *Electronic Journal of Statistics*, **4**, 1113–1150.
- VILLA-VIALANEIX, N., LIAUBET, L., LAURENT, T., CHEREL, P., GAMOT, A. & SAN CRISTOBAL, M. (2013). The structure of a gene co-expression network reveals biological functions underlying eqtls. *PLOS One*, **8**, 1–13.
- VOILLET, V., SAN CRISTOBAL, M., LIPPI, Y., MARTIN, P.G., IANNUCELLI, N., LASCOR, C., VIGNOLES, F., BILLON, Y., CANARIO, L. & LIAUBET, L. (2014). Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity. *BMC Genomics*, **15**, 797.
- VOILLET, V., BESSE, P., LIAUBET, L., SAN CRISTOBAL, M. & GONZÀLEZ, I. (2016a). Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics*, **17**, 402.

- VOILLET, V., SAN CRISTOBAL, M., PÈRE, M.C., LIAUBET, L. & LEFAUCHEUR, L. (2016b). Integrated network analysis of proteomic and transcriptomic data highlights late fetal muscle maturation process. *Molecular and Cellular Proteomics*, **In preparation**.
- WANG, Z., GERSTEIN, M. & SNYDER, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63.
- WATSON-HAIGH, N.S., KADARMIDEEN, H.N. & REVERTER, A. (2010). PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches. *Bioinformatics*, **26**, 411–413.
- WICKHAM, H. (2014). Tidy data. *Journal of Statistical Software*, **59**, 10.
- WILSON, M.E., BIENSEN, N.J., YOUNGS, C.R. & FORD, S.P. (1998). Development of Meishan and Yorkshire littermate conceptuses in either a Meishan or Yorkshire uterine environment to day 90 of gestation and to term. *Biology Reproduction*, **58**, 905–910.
- WOLD, B. & MYERS, R.M. (2008). Sequence census methods for functional genomics. *Nature Methods*, **5**, 19–21.
- WOLD, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, **1**, 391–420.
- WOLD, S., SJÖSTRÖM, M. & ERIKSSON, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, **58**, 109–130.
- YAUK, C.L. & BERNDT, M.L. (2007). Review of the literature examining the correlation among DNA microarray technologies. *Environmental and molecular mutagenesis*, **48**, 380–394.
- ZAMPIERI, N., SORANZO, N. & ALTAFINI, C. (2008). Discerning static and causal interactions in genome-wide reverse engineering problems. *Bioinformatics*, **24**, 1510–1515.
- ZENOBI, R. (2013). Single-cell metabolomics: analytical and biological perspectives. *Science*, **342**, 1201–1211.

- ZHANG, B. & HORVATH, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, **4**, 1128.