Biologie intégrative des réponses au stress et robustesse chez le porc

Valérie Sautron

Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Institut National Polytechnique de Toulouse (INP Toulouse)*

Présentée et soutenue le *27/10/2016* **par :**

VALÉRIE SAUTRON

Biologie intégrative des réponses au stress et robustesse chez le porc

---

JURY

PHILIPPE BESSE	Professeur des Universités	Président du Jury
DARYA BAZOVKINA	Researcher	Examinatrice
SANDRINE LAGARRIGUE	Professeur des Universités	Rapporteur
ROBERT SABATIER	Professeur des Universités	Rapporteur
ELENA TERENINA	Chargée de Recherche	Co-directrice de thèse
NATHALIE VILLA-VIALANEIX	Chargée de Recherche	Co-directrice de thèse

---

École doctorale et spécialité :
SEVAB : Pathologie, Toxicologie, Génétique et Nutrition

Unité de Recherche :
UMR 1388 - GenPhySE

Directeur(s) de Thèse :
Elena TERENINA et *Nathalie VILLA-VIALANEIX*

Rapporteurs :
Sandrine LAGARRIGUE et *Robert SABATIER*

Table des matières

Liste des figures

Liste des tableaux

RÉSUMÉ Le travail de cette thèse s'inscrit dans le cadre du projet ANR SUSoSTRESS qui a pour objectif la compréhension des mécanismes moléculaires et génétiques sous-jacents à la variabilité individuelle de réponses de stress et a collecté des données longitudinales à plusieurs niveaux biologiques sur une population d'étude porcine (race Large White). La thèse est organisé en deux partie.
La première partie s'articule autour de l'analyse de données cliniques et transcriptomiques collectées à plusieurs pas de temps avant et après application de deux types de stress : injection d'ACTH et de LPS. Dans cette partie, on cherche à développer d'un modèle fonctionnel permettant de décrire et d'intégrer au mieux l'ensemble des sources de variation génétique du fonctionnement de l'axe corticotrope et plus généralement des réponses de stress dans notre population d'étude. Plus précisément, il s'agit d'élaborer un modèle (au sens biologique du terme) décrivant les différentes réponses biologiques de stress et l'influence des variations génétiques (simples et en interaction), dans le but de prédire les leviers les plus efficaces en fonction de l'objectif de sélection. Ce travail a mis en évidence une liste de 65 gènes différentiellement exprimé au cours des réponses au stress, dont un ensemble de 8 gènes liés au au cortisol (l'hormone principale du stress) par *NR3C1*, le récepteur aux glucocorticoides. Ces gènes sont des biomarqueurs potentiels pouvant être fournis aux éleveurs en tant que leviers de sélection permettant un meilleur équilibre entre amélioration des caractères de production et des caractères de robustesse.
La deuxième partie de ce travail s'articule autour du développement d'un outil d'analyse statistiques adapté à l'intégration de données 'omiques longitudinales avec une variable cible d'intérêt. Nous proposons la « multiway-SIR », qui étend la méthode dual-STATIS, une méthode d'analyse de données cubiques non supervisée, au cadre de la SIR, une méthode de régression semi-paramétrique pouvant être utilisée à des fins exploratoires. Cette méthode est appliquée sur les données cliniques de l'expérience d'ACTH et permet d'y explorer l'influence de la variabilité de la réponse du cortisol à une injection d'ACTH.

MOTS CLEFS : stress, axe corticotrope, cortisol, évolution temporelle, biologie intégrative

ABSTRACT : This PhD thesis is part of the SUSoSTRESS project. This ANR funded project aims at improving the knowledge about molecular and genetic mechanisms underlying inter-individual variability in stress responses. Longitudinal data were collected at several biological levels on a porcine population (Large White). This work is structured in 2 parts.
The first part is built around clinical and transcriptomic longitudinal data analyses collected before and after 2 types of stress factors : ACTH and LPS injection. The aim of this contribution is to develop a functional model describing all sources of genetic variation in the HPA axis activity and in stress responses in our study population. More precisely, it aims at defining a model describing the different biological stress responses and the influence of genetic variations in order to identify the most efficient selection levers according to selection goals. This work allowed for the identification of 65 differentially expressed genes during stress responses. Among them, 8 genes were highly linked to cortisol (the main stress hormone) through *NR3C1* (glucocorticoid receptor (GR)). These genes are potential biomarkers and can be communicated to breeders as selection levers for a better trade-off between production and robustness traits in farm animals.
The second part is built around the development of a statistical tool suited for the data integration of repeated omic measurements with a real target variable. We introduce the "multiway-SIR" approach which extends the dual-STATIS (an approach to study 3-way datasets) method to the SIR framework (a semi-parametric regression model that can be used in an exploratory way). This method is illustrated on clinical data from the ACTH experiment. It allows for the exploration of the link between clinical variable response over time and inter-individual variability in the cortisol response to an ACTH injection.

KEYWORDS : Stress, Hypothalamic-pituitary-adrenal (HPA) axis, Cortisol, Time-course, Systems biology

Remerciements

Je remercie en premier lieu mes directeurs de thèses : Elena TERENINA, Nathalie VILLA-VIALANEIX et Pierre MORMEDE. Merci à vous pour la confiance que vous m'avez accordée et les précieux conseils que vous m'avez prodigués durant ces 3 années. Je remercie tout particulièrement Nathalie pour avoir eu la patience et l'écoute pendant toutes mes phases de doute et de déprime : je n'ai probablement pas été la doctorante la plus simple du monde à gérer, mais tu m'as apporté beaucoup et grâce à toi, je pense avoir gagné en confiance en moi et être un peu mieux préparée à affronter le monde des adultes. Un grand merci à toi.

Je remercie également la région Midi-Pyrénées et l'Agence Nationale de la Recherche (ANR) pour le financement de cette thèse.

Je remercie également les membres de mon comité de thèse : Pascal MARTIN, Dominique ROCHA et Magali SAN-CRISTOBAL. Merci à vous pour vos conseils éclairés qui m'ont permis d'avancer sur le bon chemin.

Je remercie Sandrine LAGARRIGUE et Robert SABATIER pour avoir accepté de rapporter cette thèse. Merci à vous pour votre lecture minutieuse et vos remarques pertinentes qui permettront l'amélioration de ce travail. Je remercie également Philippe BESSE et Darya BAZOVKYNA pour leur participation à mon jury de thèse.

Je remercie également toute l'équipe du laboratoire GenPhySE dans lequel j'ai réalisé mon doctorat. Merci à vous pour ces 3 merveilleuses années. Je remercie particulièrement Laurence LIAUBET, sans qui je ne serais probablement pas là aujourd'hui : merci de m'avoir suggéré les contacts de Pierre, Elena et Nathalie, ainsi que pour tous tes conseils et ton aide. Tu as été comme une troisième maman pour moi(la deuxième étant Nathalie bien sûr !). Je remercie également tout particulièrement les doctorants du labo. Merci à vous pour tous ces bons moments : les repas cohésion du jeudi, les soirées Quizz du mercredi et surtout les soirées Bachelor du lundi (#MerciStagiaire). Le petit animal sauvage qui vit en moi a été difficile à

apprivoiser, mais maintenant qu'il l'est, vous pouvez comptez sur lui pour toujours avoir une petite pensée émue pour vous ! Merci aussi à Valentin VOILLET pour avoir été le meilleur co-bureau du monde. Les 3 années passés en ta compagnie ont été super malgré la jalousie qui me rongeait devant tes ACP si parfaites ! Je remercie également Caroline YDIER pour son aide sur les analyses de données LPS qui m'a permis de gagner un temps précieux sur ma dernière année.

Je remercie également les membres du laboratoire MIAT, dans lequel j'ai pu réaliser la plus petite mobilité du monde. Merci à vous pour votre accueil chaleureux. Je remercie également Nathalie VIGUERIE et Marie CHAVENT pour leur aide et leur collaboration sur le développement méthodologique réalisé au cours de cette thèse.

Je remercie également mes amis et ma famille pour leur soutien. Merci à Stéphane pour m'avoir accueillie à Toulouse et avoir été le super grand frère toujours prêt à m'aider. Merci à Edwige pour toutes nos soirées de rigolade et pour m'avoir écoutée quand je n'en pouvais plus de cette thèse. Merci également à Julien, Edouard et Poring pour m'avoir apporté bonne humeur, écoute et fun sur la dernière ligne droite. Vous avez rendu mon ultime traversée du désert supportable et si j'ai pu arriver au bout de ce travail aujourd'hui, c'est aussi un peu grâce à vous.

Enfin, je tiens à remercier tout particulièrement Léo, le bel inconnu devenu meilleur des copains puis meilleur des maris : cette thèse, je te la dédie entièrement. Sans toi, je n'aurais pas eu la démarche de me lancer dans un tel projet. Sans toi, je n'aurais pas eu le courage de le finir. Tu as été le pilier de ces trois années et tu les as rendues plus belles que toutes les autres années de ma vie. Si je suis docteur aujourd'hui, c'est à la fois pour et grâce à toi. Merci d'être arrivé dans ma vie, d'y être resté et d'avoir accepté de la finir avec moi.

Note

Cette thèse est rédigée en français, à l'exception du chapitre 2 correspondant aux articles publiés ou en cours de publication dans une revue internationale à comité de relecture.

— Le chapitre 1 est rédigé en français.

— Le chapitre 2 comporte 2 articles. Le premier article a fait l'objet d'une publication [SAUTRON et collab., 2015]. Il a également fait l'objet d'une communication orale lors de la journée des doctorants de l'école doctorale SEVAB en mars 2015 et d'un poster présenté à *35th International Society for Animal Genetics Conference (2015)* à Salt Lake City en juillet 2016.
Le second article est en cours de soumission. Ce travail a fait l'objet d'un projet tutoré pour le parcours GMM (4ème année) de l'INSA et du stage de Caroline Ydier, encadré en collaboration avec Nathalie Villa-Vialaneix et Laurence Liaubet.

— Le chapitre 3 est rédigé en français. Ce travail a fait l'objet d'un poster présenté à *Statistical Methods for Post Genomic Data (SMPGD) 2016* à Lille en février 2016 et à *22nd International Conference on Computational Statistics* à Oviedo en août 2016. Il fait également l'objet d'un article en cours de rédaction et son implémentation est en cours dans un package R : le package **SirStatis**.

— Le chapitre 4 est rédigé en français.

Chapitre 1

Introduction

Sommaire

1.1 Impact de la sélection sur les élevages porcins

L'élevage des animaux de production est apparu il y a environ 10000 ans et a permis la conversion de matières non digestibles par les humains (herbe,...) en sources de protéines, lipides et nutriments utiles à leur survie et à leur développement. Des processus de sélection se sont alors mis en place afin de sélectionner les animaux possédant des caractères jugés attractifs par les éleveurs (quantité de lait produite, taille des œufs...) [HAYES et collab., 2013]. Cette sélection a d'abord été le fruit de méthodes empiriques mais depuis l'avènement de la génétique, la science s'est invitée dans les processus de sélection. Les objectifs de sélection chez les animaux d'élevage incluent le plus souvent des caractères de production, de qualité, de robustesse et d'adaptation des animaux. Ils se sont compliqués avec le temps, pour prendre en compte des demandes de plus en plus variées [RAUW et collab., 1998].

Chez le porc, ces caractères incluent par exemple :

— vitesse de croissance

— efficacité alimentaire

— ratio gras/maigre dans les carcasses

— taille des portées

— etc.

En particulier, le coût le plus important pour les éleveurs étant celui lié à l'alimentation des animaux, les programmes de sélection ont grandement été dirigés en faveur d'une amélioration de l'efficacité alimentaire [LUITING, 1990]. Dans son papier de 1998, RAUW et collab. [1998] ont montré une augmentation de la vitesse de croissance et de l'efficacité alimentaire chez des porcs hollandais et norvégiens entre 1960 et 1996. Selon Rauw, ces caractères disposant d'une héritabilité haute à modérée et l'accent étant mis sur les caractères de production dans les programmes d'élevage, l'augmentation des scores obtenus pour ces caractères seraient principalement d'origine génétique. En France, TRIBOUT et collab. [2003] ont réalisé une étude comparant différents caractères de production entre des populations de porcs issus de semences congelées datant de 1977 et de 1998. Ils montrent que la sélection génétique effectuée depuis 1977 a entrainé une diminution significative de la composition adipeuse des carcasses (différence de $-5,2$ mm pour l'épaisseur de lard dorsal à 20 semaines d'âge entre 1977 et 1998) et une augmentation de la teneur en viande maigre (différence de +8,6 points) et de la vitesse de croissance des porcs (différence de 75 g/j entre le sevrage et l'abattage). De même, selon les chiffres de l'Institut de la Filière Porcine (IFIP), on a pu observer une augmentation

quasi-constante de l'efficacité alimentaire sur les 20 dernières années (voir figure 1.1). Cependant, la sélection génétique a entrainé une contre sélec-

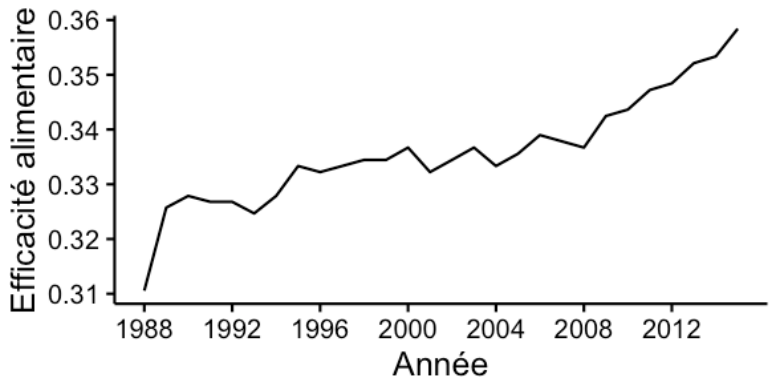


FIGURE 1.1 – Évolution de l'efficacité alimentaire dans les élevages français de porcs entre 1988 et 2015 [Institut de la Filière Porcine, 2016].

tion pour certains caractères pourtant importants pour les éleveurs, soit parce que ces caractères sont difficiles à mesurer, soit parce que leur amélioration a été largement ignorée jusque-là [HAYES et collab., 2013]. Ce dernier cas est notamment vrai chez le porc d'élevage pour les caractères liés à la robustesse, pour lesquels l'impact de la sélection génétique a été négatif. La robustesse est définie par Knapp comme étant la capacité pour les animaux à combiner un haut potentiel de production à une grande résistance aux conditions environnementales afin de maintenir ce haut potentiel de production dans une large variété d'environnements [KNAP, 2005]. Il s'agit d'une mesure globale de la sensibilité de l'animal aux facteurs d'environnement, comme par exemple les températures élevées, le microbisme ou des conditions d'élevage sub-optimales. Ce concept inclut également quelques caractères spécifiques, regroupés sous le terme de caractères fonctionnels, qui se manifestent particulièrement lorsque les conditions d'environnement sont dégradées [KNAP et collab., 2008] :

— qualité des aplombs
— survie des nouveau-nés
— résistance aux maladies
— longévité productive
— etc.

Parmi les exemples d'impact négatif sur les caractères de production, on peut citer SATHER [1987] qui a montré qu'une plus grande faiblesse des pattes était observée chez des porcs Lacombe sélectionnés pour leur forte quantité de viande maigre et leur fort taux de croissance par rapport à

des porcs contrôles non sélectionnés. De même, des études menées dans plusieurs races de porcs (Yorkshire, Landrace, Duroc, Large White) ont montré des corrélations génétiques négatives entre les scores de taux de croissance et de quantité de viande maigre avec les scores d'épaisseur de gras dorsal et de faiblesse des pattes [Huang et collab., 1995; Lundeheim, 1987; Webb et collab., 1983]. D'autre part, il a été montré que les races rustiques de porc présentaient une quantité plus importante de fibres musculaires de type I (fibres musculaires « lentes »), particulièrement sollicitées pour le maintien de la posture dans les muscles dorsaux, que chez les races sélectionnées [Essén-Gustavsson et Lindholm, 1984; Rahelic et Puac, 1981]. Les fibres musculaires chez les porcs rustiques étaient également plus vascularisées et présentaient un plus haut niveau d'activité enzymatique [Karlström, 1995]. Rauw suggère que ces résultats montrent que la sélection a entrainé une diminution de la résistance des porcs aux stress environnementaux, ce qui limiterait donc l'expression de leur plein potentiel de production en conditions environnementales non optimales. D'après Schinckel [2010a], les porcs élevés dans les conditions commerciales expriment moins de 80% de leur potentiel génétique.

Selon Knap [2009], « les objectifs de sélection des animaux pour un élevage durable » combinent les objectifs de robustesse et de production de telle sorte que la sélection balance les modifications génétiques du potentiel de production avec les modifications génétiques de sensibilité à l'environnement. Les objectifs de robustesse prennent donc une part de plus en plus importante dans les processus de sélection et ce afin d'obtenir des animaux dont le niveau de production puisse se maintenir dans une large gamme de conditions climatiques et de systèmes d'élevage, tout en optimisant le bien-être des animaux [Mormède et Foury, 2009]. Cependant, comme cité plus haut, certains caractères autres que ceux de production ont été complètement ignorés par la sélection alors qu'ils peuvent présenter un intérêt pour les éleveurs. De plus, certains de ces caractères sont difficiles à mesurer. Ainsi, bien que les stratégies de sélection actuelles aient permis d'améliorer les caractères de production de façon substantielle, il est nécessaire de leur faire prendre une nouvelle direction de manière à répondre aux problématiques spécifiques au contexte économique et sociétal émergeant et de définir de nouveaux objectifs de sélection pour assurer la durabilité des élevages [Hayes et collab., 2013].

Plusieurs méthodes peuvent être mises en place pour améliorer les caractères de robustesse en élevage [Mormède et collab., 2011]. Premièrement, il est possible d'intégrer des caractères fonctionnels comme la qualité des aplombs ou la mortalité à diverses périodes de la vie en tant

que points spécifiques de programmes de sélection [KNAP et collab., 2009]. Cette approche demande cependant beaucoup de temps pour arriver à un résultat observable et certains caractères peuvent se révéler difficiles à mesurer, la rendant compliquée à mettre en place. Elle est cependant utilisée avec succès depuis une quinzaine d'années [KNAP et collab., 2009]. Deuxièmement, il est possible de mesurer la sensibilité globale à l'environnement en comparant des mesures effectuées chez des animaux ayant un génotype identique mais placés dans des environnements différents. Sa mise en œuvre est cependant difficile compte tenu de la faible héritabilité de ce caractère de résistance globale aux effets de l'environnement [KNAP et SU, 2008]. Enfin, la troisième stratégie, sur laquelle nous allons nous concentrer, se focalise plus sur la génétique des réponses neuroendocriniennes au stress, et notamment sur l'axe corticotrope qui occupe une place importante dans ce processus.

L'intérêt pour les questions de stress dans les élevages existe depuis les années 60 [DANTZER et MORMÈDE, 1983]. Le terme de stress était alors utilisé pour désigner les effets de l'environnement sur les animaux et permettait d'expliquer les pertes exceptionnelles en animaux et en productivité. On l'utilisait également pour désigner les réactions exceptionnellement prononcées à des stimulus communs [CASSENS et collab., 1975]. L'utilisation du terme « stress » était alors peu rigoureuse et il convient, pour une meilleure compréhension des mécanismes en jeux, de revenir sur sa définition.

1.2 Concept de stress

Le concept de stress a été popularisé par Hans Selye en 1936 [SELYE et collab., 1936]. Il décrivait alors un syndrome apparaissant chez des rats ayant reçu des dommages d'origines diverses (exposition au froid, activité physique excessive, intoxications avec des doses non létales d'agents médicamenteux...) et dont les symptômes étaient indépendants de la nature des agressions subies ou du type pharmacologique du médicament employé. Ces symptômes, communs à tous les stimulus, comprenaient l'élargissement des glandes surrénales, le rétrécissement du thymus et des organes lymphoïdes et des ulcères au niveau du système digestif. Si les stimulus étaient maintenus dans le temps, les rats entraient dans une phase de résistance à ceux-ci, et s'ils étaient maintenus pendant une plus grande période de temps, ils entraient à nouveau dans une phase délétère qui débouchait sur la mort. Selye a appelé cette réponse le syndrome général d'adaptation. Par la suite, SELYE [1956] précise le concept en définissant le *stress* comme la réponse non spécifique des organismes à toutes sortes

de contraintes, appelées « stresseurs » (ou facteurs de stress), qui peuvent être de toute nature (environnementale, infectieuse...). Cette réponse se caractérise principalement par l'activation des glandes surrénales et donc la libération de cortisol et repose sur le principe qu'elle est exactement la même quelque soit le stresseur [SELYE, 1973].
Depuis, ce modèle de réponse au stress s'est peu à peu compliqué. Au début des années 70, MASON [1971] mettait en évidence l'importance de la dimension psychologique comme premier médiateur de la cascade d'activation de l'axe corticotrope. En soumettant des singes à un jeûne en présence d'animaux contrôles ayant un régime alimentaire normal, il mesurait une augmentation de la sécrétion de cortisol dans les urines des singes jeûnant. Cependant, cette modification de l'activité de l'axe corticotrope n'était pas observée lorsque les singes jeûnant étaient isolés des singes contrôles ou lorsqu'ils étaient nourris en même quantité que les contrôles mais avec de la nourriture moins calorique. En d'autres termes, la réponse biologique observée n'était pas provoquée par les conséquences physiologiques du stresseur, mais par la perception que les animaux avaient de la situation et si eux la trouvaient stressante.
Plus tard, DANTZER et MORMÈDE [1983] définirent les réponses au stress comme étant à la fois physiologiques et comportementales et l'interaction entre ces deux composantes comme essentielles à la compréhension des réponses au stress. Selon eux, les réponses physiologiques sont en effet non-spécifiques au stresseur tandis que la composante comportementale, qui peut être définie par les modes d'adaptation active (fuite/lutte) ou passive (inhibition) est spécifique au stresseur et vise à contrôler le facteur menaçant . Cette composante psychique serait alors dépendante du patrimoine génétique des individus et de leurs expériences passées [KOOLHAAS et collab., 1999].

Depuis, le concept de stress a été à de nombreuses reprises exploré et mis en relation avec les notions d'évaluation cognitive et de nombreux travaux ont été réalisés pour mieux comprendre les mécanismes sous-jacents aux réponses au stress, que ce soit chez l'homme ou l'animal, y compris chez l'animal d'élevage [CHROUSOS et GOLD, 1992; KOOLHAAS et collab., 1999; LAZARUS, 1993; MOISAN et LE MOAL, 2012]. Chez l'humain par exemple, une réponse émotionnelle peut être induite ou modifiée par la nouveauté d'un événement, le plaisir qu'il peut procurer, son importance ou encore les possibilités de contrôle sur la situation engendrée et son importance aux yeux des autres [SCHERER, 2001]. La pertinence de tels éléments dans le contexte des animaux d'élevage est encore à explorer [BOISSY et collab., 2007]. Cependant, des études menées chez différentes espèces animales tendent à montrer que la possibilité d'avoir du contrôle sur une

situation engendrée par un stresseur (pouvoir interagir avec ses nouveaux partenaires lors d'un changement d'environnement, ne pas être restreint lors de la survenue d'un stress...) pouvait engendrer une hyperactivité ou une apathie complète chez des veaux et des truies, suggérant que cet élément soit aussi d'importance chez les animaux d'élevage [BOISSY et collab., 2001; BROOM, 1987; LADEWIG et SMIDT, 1989; VEISSIER et collab., 2001]. Enfin, DÉSIRÉ et collab. [2006] ont montré dans leurs travaux que les moutons présentaient des réponses physiologiques et comportementales différentes en fonction de la manière dont sont présentés des stimulus : un stimulus soudain entraine une augmentation temporaire du rythme cardiaque et un sursaut, tandis qu'un nouveau stimulus entraine une modification de l'orientation de la tête et des variations dépendantes des individus pour le rythme cardiaque. Tous ces travaux suggèrent donc que les réponses comportementales, aussi bien que les réponses physiologiques, sont spécifiques des individus, y compris chez les animaux d'élevage, et que le modèle de réponse au stress proposé par Selye est probablement bien plus complexe en réalité. La figure 1.2, tirée de [VEISSIER et BOISSY, 2007] illustre ce modèle.

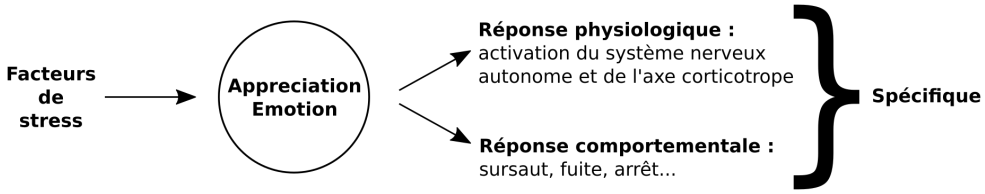


FIGURE 1.2 – Modèle des réponses au stress proposé par VEISSIER et BOISSY [2007]. Ce modèle considère que l'expérience, les sentiments ou les émotions *per se*, mais aussi les réponses physiologiques et comportementales, sont le résultat de l'appréciation que chaque individu fait d'un facteur de stress donné.

D'après ce modèle, il est donc possible d'étudier les réponses au stress à deux niveaux :

— d'un point de vue psychologique pour avoir une meilleur compréhension des mécanismes en jeu dans l'appréciation que font les animaux des facteurs de stress et du choix de la méthode d'adaptation pour y faire face ;

— d'un point de vue biologique pour une meilleure compréhension des mécanismes en jeu lors de la mise en place des réponses physiologiques par l'axe corticotrope et le système nerveux autonome. Bien que ces réponses physiologiques soient non-spécifiques au sens de Selye, notamment pour la réponse de l'axe corticotrope, il existe une variabilité individuelle dans l'intensité de ces réponses.

1.3 Description de l'axe corticotrope

L'axe corticotrope est l'un des systèmes clefs, avec le système nerveux autonome, de la réponse au stress. Localisé dans l'hypothalamus, l'hypophyse et les glandes surrénales, il est à l'interface du système nerveux central et des fonctions endocrines.

1.3.1 Hypothalamus

L'hypothalamus, situé sur la face ventrale de l'encéphale, est à la fois une structure du système nerveux central et une glande endocrine. Il répond à divers stimulus tels que la lumière, les odeurs, les informations nerveuses provenant du cœur, de l'estomac, des organes reproducteurs ou encore du système nerveux autonome et à diverses hormones. En réponse, l'hypothalamus sécrète des hormones qui passent directement à l'hypohyse par le système porte hypothalamo-hypophysaire, réseau de capillaires qui assure la circulation des hormones hypothalamiques vers l'antéhypophyse. Ces hormones sont impliquées dans différents mécanismes physiologiques allant de la reproduction à la croissance ou encore dans le mécanisme spécifique de réponse aux stress.

Parmi les hormones impliquées dans ce dernier, on compte les hormones les plus impliquées dans la réponse au stress : CRH (corticotropin-releasing hormone) et vasopressine. Il est à noter que la vasopressine impliquée dans le contrôle de la sécrétion d'ACTH par l'hypophyse (en concomitance avec la CRH) a une localisation cellulaire très spécifique (cellules parvocellulaires du sous-noyau paraventriculaire) différente de la vasopressine impliquée dans la régulation du métabolisme de l'eau et des électrolytes (localisée dans les noyaux magnocellulaires).

1.3.2 Hypophyse

L'hypophyse est une glande complexe reliée à l'hypothalamus par la tige pituitaire. Elle peut être divisée en deux lobes : l'anté-hypophyse (ou adénohypophyse) et la post-hypophyse (ou neurohypophyse). L'antéhypohyse, sous contrôle de l'hypothalamus, est responsable de la production de nombreuses hormones impliquées dans la croissance et les métabolismes (hormone de croissance ou GH pour *growth hormone*, somatostatine, TSH(*thyroid-stimulating hormone*)), la reproduction (LH (*luteinizing hormone*), FSH (*follicle-stimulating hormone*), prolactine) et le stress (ACTH (*adrenocorticotropic hormone*)).
La post-hypophyse est en réalité un prolongement de l'hypothalamus. Elle est en effet constituée des parties terminales des axones des neurones

constituant l'hypothalamus. Ces neurones sont en contact avec des capillaires sanguins entrant par l'artère hypophysaire inférieure et sortant par la veine hypophysaire. Les axones sécrètent dans ces capillaires sanguins les neurohormones contenues dans les granules de sécrétion, vasopressine et ocytocine.

1.3.3 Glandes surrénales

Les glandes surrénales sont les dernières composantes de l'axe corticotrope. Elles sont situées au-dessus des reins et peuvent être divisées en deux parties, le cortex et la médulla. La médulla se situe au cœur des glandes surrénales. C'est un noyau de neurones ne possédant pas d'axones, mais spécialisés dans la production de catécholamines (dopamine, noradrénaline, adrénaline). Les catécholamines sont des neuro-transmetteurs qui induisent des modifications physiologiques de l'organisme : augmentation de la fréquence cardiaque, de la pression artérielle et du taux de glucose dans le sang.
Le cortex est situé entre la capsule (paroi externe des glandes surrénales) et la médulla. Il est divisé en trois couches de cellules (la zona glomerulosa, la zona fasciculata et la zona reticularis) qui sont principalement sous le contrôle de l'ACTH et sécrètent des hormones telles que les glucocorticoïdes et minéralocorticoïdes, des androgènes et des œstrogènes.

1.3.4 Synthèse et métabolisme des glucocorticoïdes

Les glucocorticoïdes sont des hormones stéroïdiennes (au même titre que les androgènes et les œstrogènes) synthétisées à partir de cholestérol. Les principaux glucocorticoïdes sont le cortisol et la corticostérone, dont l'importance varie selon les espèces : cortisol chez le porc, l'homme, les ruminants, les poissons, corticostérone chez les volailles et les animaux de laboratoire. Ces deux hormones sont très proches et ont la même activité physiologique.

Ils sont synthétisés par le cortex des glandes surrénales et plus particulièrement par la zona fasciculata et la zona reticularis qui possèdent l'enzyme 17-alpha-hydroxylase. La zona glomerulosa étant dépourvue de cette enzyme, elle ne peut participer à ce processus, mais participe en revanche à la synthèse d'aldostérone (une autre hormone stéroïdienne dite minéralocorticoïde en raison de son action sur le métabolisme hydrosodique) par l'oxydation de la corticostérone. La figure 1.3 illustre les voies de synthèses des hormones stéroïdiennes, dont les glucocorticoïdes.

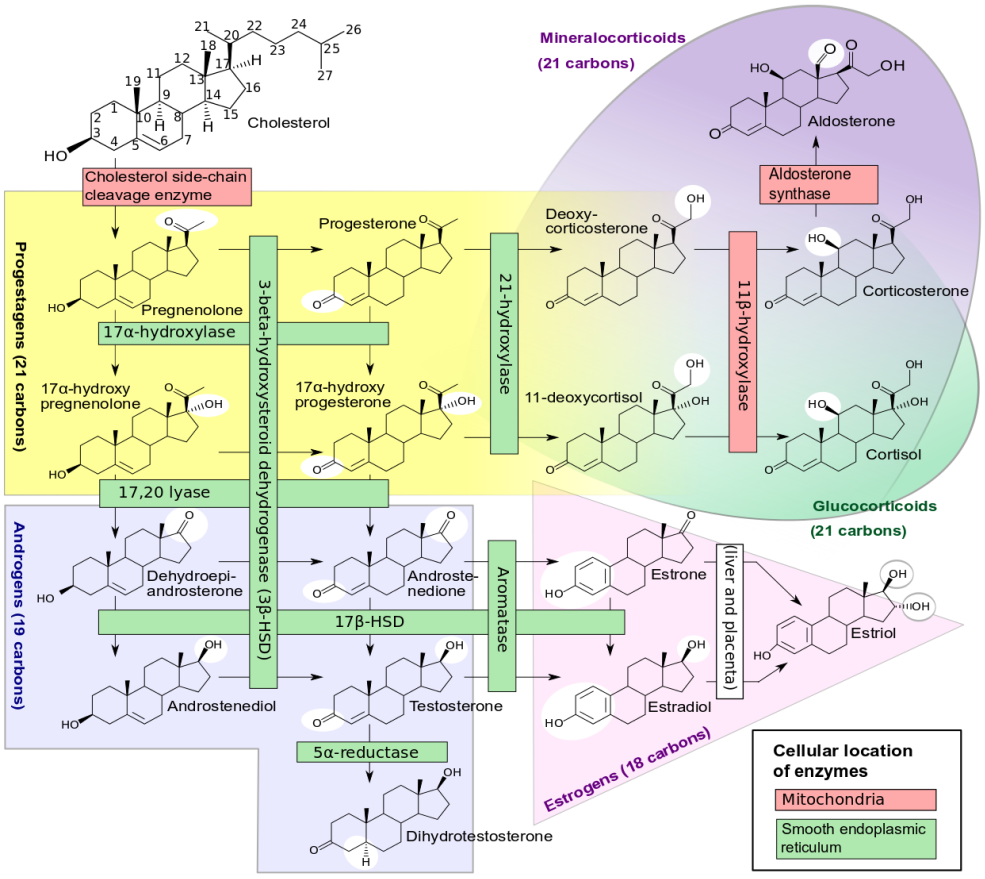


FIGURE 1.3 – Voie de synthèse des hormones stéroïdiennes. Häggström M, Richfield D (2014). "Diagram of the pathways of human steroidogenesis". Wikiversity Journal of Medicine 1 (1)

Une fois libérés dans la circulation sanguine, on peut trouver les glucocorticoïdes sous trois formes :

— 75% du cortisol est lié à la CBG (*corticosteroid-binding globulin*) ou encore transcortine, une alpha-globuline produite par le foie et qui lie les glucocorticoïdes de façon très spécifique avec une affinité élevée ;

— 15% est lié à l'albumine par une liaison peu spécifique et dont l'affinité est faible ;

— 10% est sous forme libre dans le sang.

Le cortisol étant peu soluble dans l'eau, il est transporté lié à l'albumine et à la CBG dans la circulation sanguine. En revanche, sa grande liposolubilité lui permet de passer facilement à travers les membranes cellulaires pour aller jouer son rôle de facteu transcription dans ses tissus cibles. Le cortisol libre possède une durée de demi-vie de 60 à 90 minutes dans la circulation sanguine et il est métabolisé dans le foie. La majeure partie du cortisol est réduite en dihydrocortisol, puis en tetrahydrocortisol et enfin dérivé en tetrahydrocortisol glucuronide qui, soluble dans l'eau, sera éliminé par les urines. Une partie du cortisol est transformée en cortisone inactive qui, de la même manière que le cortisol pourra être réduite et conjuguée pour former le tetrahydrocortisone glucoronide qui sera lui aussi évacué par les urines (voir figure 1.4).

1.3.5 Mode d'action des glucocorticoïdes

Les glucocorticoïdes ont pour cible un grand nombre de tissus sur lesquels ils interagissent grâce aux récepteurs aux glucocorticoïdes (GR ou type II dans le cerveau) et minéralocorticoïdes (MR ou type I dans le cerveau). Les glucocorticoïdes agissent dans l'organisme par l'intermédiaire des récepteurs aux glucocorticoïdes (GR). Les GR sont des récepteurs exprimés par tous les tissus de l'organisme et sont habituellement localisés dans le cytoplasme des cellules [SCHONEVELD et collab., 2004]. Les GR possèdent 6 domaines. Parmi eux, 4 sont des domaines structuraux : 2 constituent la partie N-terminale, 1 constitue la partie C-terminale et 1 forme un domaine charnière. Les 2 autres domaines sont des domaines fonctionnels et sont responsables respectivement de la liaison à un ligand et de la liaison à l'ADN. En situation normale, les GR sont localisés dans le cytoplasme à une densité allant de 2000 à 30000 molécules par cellule en fonction du type cellulaire [ADCOCK, 2000]. En l'absence de glucocorticoïdes, ils sont liés à des protéines chaperons telles que les protéines de choc thermique *hsp90* et les protéines *FKBP*. Ces protéines chaperons permettent aux GR de rester dans le cytoplasme en dissimulant les sites de liaisons nécessaires

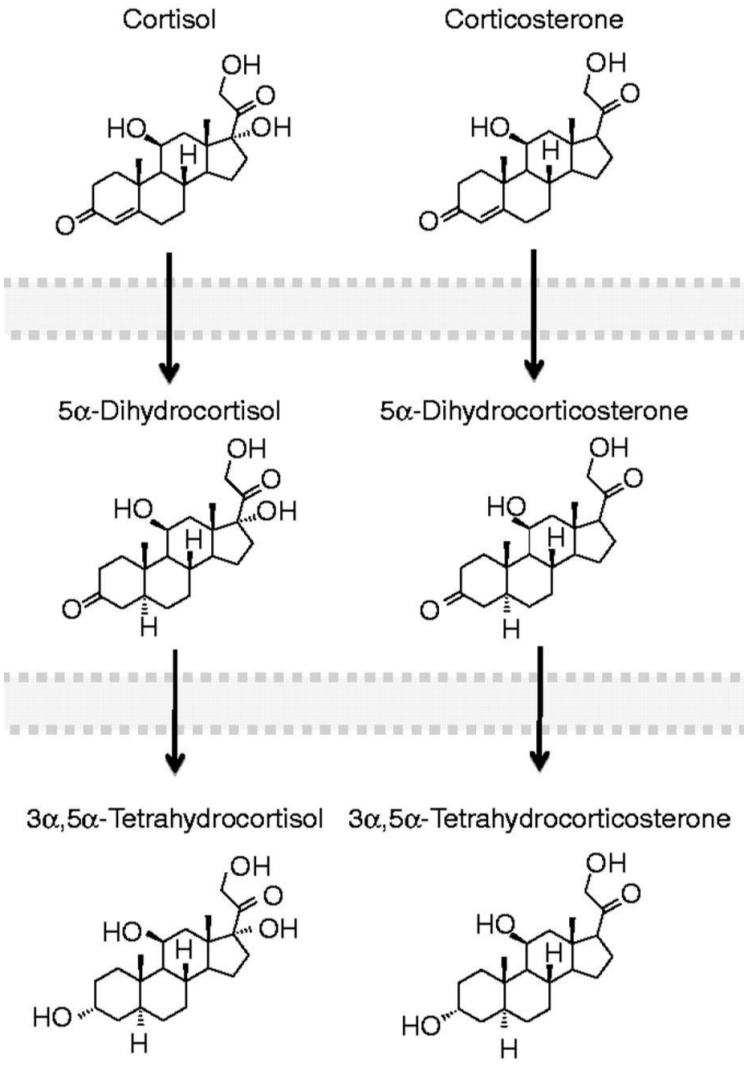


FIGURE 1.4 – Métabolisme du cortisol

au transport des GR vers le noyau [WU et collab., 2004]. Lorsque les glucocorticoïdes se fixent à leur récepteur, celui-ci change de conformation et se sépare des protéines chaperons. Le GR est alors transporté au travers de la membrane nucléaire et peut se fixer à l'ADN pour agir comme facteur de transcription et ainsi moduler l'expression de certains gènes [BARNES, 2006]. Le nombre de gènes directement régulés par les GR sous forme de dimères peut aller de 10 à 100, mais un très grand nombre de gènes est régulé de façon indirecte par l'interaction entre les GR et d'autres facteurs de transcription et autres cofacteurs [HAYASHI et collab., 2004]. Grâce à la nature ubiquitaire de leurs récepteurs, les glucocorticoïdes peuvent activer et inhiber une grande diversité de gènes impliqués dans un nombre critique

de voies métaboliques et inflammatoires [SCHONEVELD et collab., 2004].

1.3.6 Actions physiologiques des glucocorticoïdes

La grande liposolubilité du cortisol libre lui permet de se répandre dans une multitude de tissus et de cellules où il agit par l'intermédiaire des récepteurs glucocorticoïdes. Parmi les effets du cortisol on compte :

1. **Effets sur le métabolisme**. Le cortisol présente un rôle catabolique sur les tissus périphériques en favorisant les voies de la protéolyse et de la lipolyse dans plusieurs types de tissus tels que les muscles, les tissus adipeux, les tissus conjonctifs ou encore les tissus lymphoïdes. Le cortisol possède aussi un rôle permissif dans le sens où sa présence est indispensable pour permettre l'action d'autres hormones. Ainsi, la présence de cortisol est requise pour permettre aux catécholamines de développer leurs effets calorigéniques et lipolytiques. Le cortisol a aussi un effet anabolique sur le foie en favorisant les voies de la gluconéogenèse (production de glucose) et de la glycogénogénèse (production de glycogène à partir de glucose). Cette hormone participe donc, avec les catécholamines et l'insuline, à la régulation de la glycémie.
2. **Inhibition de l'ACTH**. Le cortisol libre exerce une action de rétrocontrole négatif (*feedback*) sur la sécrétion d'ACTH. Ce contrôle s'effectue à la fois au niveau de l'hypothalamus et de l'hypophyse (voir Fonctionnement de l'axe corticotrope dans la réponse au stress). Plus la concentration en cortisol est élevée et plus la sécrétion d'ACTH est inhibée.
3. **Système cardio-vasculaire**. Le cortisol participe à la régulation de la pression sanguine en maintenant la réactivité des muscles de la paroi vasculaire aux catécholamines.
4. **Effets sur le système immunitaire**. Le cortisol (et tous les glucocorticoïdes en général) inhibe les réponses inflammatoires. En effet, le cortisol supprime la synthèse et ralentit la production de l'acide arachidonique qui est un précurseur de nombreux acteurs de la réponse inflammatoire. Le cortisol ralentit aussi la multiplication des mastocytes et stabilise les lysosomes et ralentit la production de facteurs activateurs de plaquettes et d'oxide nitrique, ce qui supprime la réponse inflammatoire locale. Le cortisol supprime également les réponses immunitaires en réduisant le nombre de lymphocytes T_4, la production d'interleukines et d'interférons.
5. **Effets sur le système nerveux central**. Le cortisol est capable de moduler directement la réponse neuronale en se fixant sur les récepteurs aux glucocorticoïdes de type I et II qui sont particulièrement

exprimés dans le système limbique et l'hippocampe. Il aurait également un rôle dans la diminution du volume de l'hippocampe au cours du vieillissement et dans la mémoire [LUPIEN et collab., 1998].

6. **Effets sur le stress**. Le cortisol est connu comme étant une hormone clef des réponses au stress. Cet effet particulier du cortisol est détaillé dans le paragraphe suivant.

1.3.7 Fonctionnement de l'axe corticotrope lors de la réponse au stress

Lorsqu'un individu est confronté à un facteur de stress (situation nouvelle, menaçante,...), une cascade de réactions physiologiques va se mettre en place, et ce, de façon inconsciente. Deux régions clefs du cerveau, l'amygdale (ou complexe amygdalien) et le locus cœruleus vont orchestrer cette réponse en activant de façon concomitante le système nerveux autonome et l'axe corticotrope. Ce dernier va sécréter la CRH ainsi que la vasopressine qui vont être transportés par le système porte hypothalamo-hypophysaire jusqu'à l'hypophyse pour y déclencher la sécrétion de l'ACTH dans la circulation sanguine. L'ACTH va à son tour provoquer la sécrétion des glucocorticoïdes par le cortex des glandes surrénales. En périphérie, les glucocorticoïdes, dont le cortisol, vont agir de concert avec les catécholamines (libérées par le système nerveux autonome) pour augmenter le tonus vasculaire, la pression artérielle et la fréquence respiratoire. Ils vont également avoir une action catabolique sur les lipides et protéines et mobiliser les facteurs énergétiques pour les diriger vers les muscles et le cerveau afin de subvenir aux besoins des réponses comportementales. Enfin, les glucocorticoïdes vont provoquer un boost temporaire du système immunitaire et une inhibition temporaire de fonctions biologiques coûteuses telles que la digestion, la croissance et la reproduction. Lorsque le système fonctionne correctement, ces modifications sont limitées dans le temps. En effet, le cortisol (de concert avec les catécholamines) va aller exercer son rétrocontrôle négatif sur le complexe hypothalamo-hypophysaire pour inhiber la production d'ACTH et stopper la réponse au facteur de stress [MOISAN et LE MOAL, 2012]. Le mécanisme physiologique de la réponse au stress impliquant l'axe corticotrope est illustré en figure 1.5.

Il existe une variabilité génétique de l'axe corticotrope à tous ses niveaux [MORMEDE et TERENINA, 2012], avec une bonne héritabilité [LARZUL et collab., 2015]. Son étude constitue donc un atout pour parvenir aux objectifs de sélection décrits précédemment pour des animaux plus robustes et des élevages plus durables.

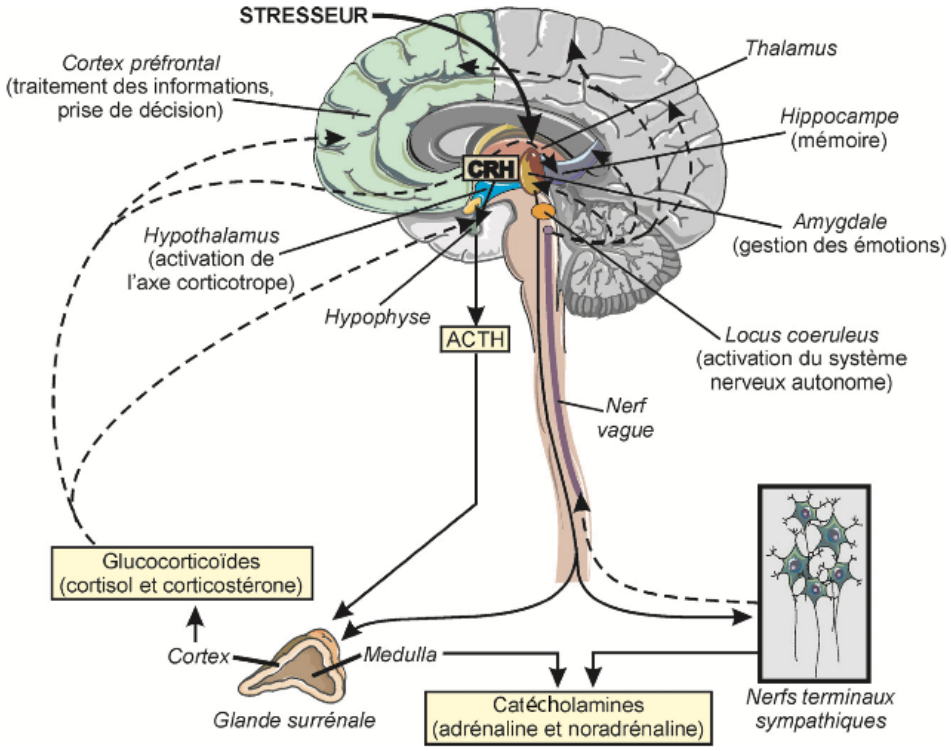


FIGURE 1.5 – Mécanisme physiologique de la réponse aux stress [NATIONAL DE LA SANTÉ ET DE LA RECHERCHE MÉDICALE (INSERM), 2011].

1.4 Génétique de l'axe corticotrope

La génétique de l'axe corticotrope et ses sources de variabilité sont bien documentées (voir [MORMEDE et TERENINA, 2012] pour une littérature abondante sur le sujet). Il a été notamment montré que les tests des fonctions de l'axe corticotrope présentaient une grande stabilité au niveau des caractéristiques individuelles [BERTAGNA et collab., 1994; COSTE et collab., 1994; HENNESSY et collab., 1988; HUIZENGA et collab., 1998; TANAKA et collab., 1993].
Si on se restreint aux animaux d'élevage avec un intérêt particulier pour le porc, on peut citer les nombreux travaux comparatifs entre les Meishan (MS), race d'origine chinoise, et les races de porcs de type européen tels que les Large White. Ainsi, il a été montré que les Meishan donnent naissance à des portées plus nombreuses [BIDANEL, 1993; CANARIO et collab., 2009, 2006], ont des carcasses plus grasses et un meilleur indice de qualité de viande [BIDANEL et collab., 1993, 1990]. Ils possèdent également une activité de l'axe corticotrope et un niveau de cortisol moyen en circulation

sanguine élevés, proches de ceux observés chez les sangliers [DÉSAUTÉS et collab., 1999; WISE et collab., 2001]. Ce niveau d'activité supérieur peut être mesuré non seulement au niveau de la production de cortisol [FOURY et collab., 2007; HAY et MORMEDE, 1998], mais aussi au niveau de la quantité de CBG circulante, supérieur chez les Meishan [GUYONNET-DUPÉRAT et collab., 2006; OUSOVA et collab., 2004] . HAZARD et collab. [2008] ont montré que le niveau plus élevé de production de cortisol pouvait être expliqué (au moins en partie) par la sensibilité plus élevée de l'axe corticotrope à l'ACTH chez ces animaux. PERREAU et collab. [1999] montrent que les récepteurs aux minéralocorticoïdes (MR) dans l'hippocampe sont plus denses chez les Meishan que chez les Large White, tandis que les récepteurs aux glucocorticoïdes (GR) dans l'hypophyse sont plus denses chez les Large White et les travaux de DE KLOET et collab. [1998] et BRADBURY et collab. [1994] suggèrent que la différence de ratio MR/GR entre ces deux races explique aussi en partie les différences phénotypiques marquées dans l'activité de l'axe corticotrope. C'est d'ailleurs en raison de ces différences marquées que Meishan et Large White sont souvent utilisés pour la recherche de locus impliqués dans la variabilité génétique des caractères de production, de comportement et des réponses neuroendocrines de stress chez le porc [DÉSAUTÉS et collab., 2002]. newline Enfin, il a été montré dans plusieurs espèces que l'activité de l'axe corticotrope était un caractère très sensible à la sélection génétique [BROWN et NESTOR, 1973; EDENS et SIEGEL, 1975; GROSS et SIEGEL, 1985; RONECKER, 2010], avec de hauts niveaux d'héritabilité variant entre 0,4 et 0,5.

Comme nous le disions précédemment, la variabilité génétique de l'axe corticotrope existe à plusieurs niveaux. La variabilité génétique de la sensibilité à l'ACTH est la mieux documentée [MORMEDE et TERENINA, 2012]. Chez le porc par exemple, il a été montré que la production de glucocorticoïdes en réponse à une injection d'ACTH était un caractère individuel [HENNESSY et collab., 1988]. Dans une étude récente, LARZUL et collab. [2015] ont ainsi pu calculer un taux d'héritabilité pour la sensibilité à l'ACTH de 0,684. Chez le canard, la réponse des surrénales à l'ACTH semble être un élément important de la différence entre lignées divergentes sélectionnées pour la réponse de l'axe corticotrope à un stress de suspension [RONECKER, 2010]. La recherche de gènes différentiellement exprimés chez le porc et le poulet a mené à la mise en évidence de gènes candidats liés aux différences observables dans la sensibilité à l'ACTH [BUREAU et collab., 2009; HAZARD et collab., 2008; JOUFFE et collab., 2009]. Plusieurs études ont par ailleurs relié la sensibilité de l'axe corticotrope à l'ACTH à certains caractères de production chez les animaux d'élevage. Ainsi, HENNESSY et JACKSON [1987] ont montré chez le porc un lien entre

forte sensibilité à l'ACTH et faible poids, faible taux de croissance et faible efficacité alimentaire. De même, une corrélation négative a été trouvée chez des béliers entre la mesure de cortisol après injection d'ACTH et l'efficacité alimentaire [KNOTT et collab., 2008]. En revanche, GILBERT et collab. [2007] dans une étude portant sur deux lignées de porcs Large White divergentes sur l'efficacité alimentaire n'ont pas montré de différence significative pour l'activité de l'axe corticotrope, ce qui permet de faire l'hypothèse que ces deux caractères sont influencés de manière indépendante par les facteurs génétiques.

La biodisponibilité des hormones est un autre élément particulièrement sujet à variabilité génétique au sein de l'axe corticotrope. DÉSAUTÉS et collab. [2002] ont mis en évidence un lien entre les niveaux de cortisol plasmatique (basal et post-stress) et un locus incluant le gène codant pour la CBG (*SERPINA6*) chez le porc. OUSOVA et collab. [2004] ont confirmé ce résultat en dosant directement la CBG circulante et GUYONNET-DUPÉRAT et collab. [2006] suggèrent qu'il existe au moins un polymorphisme dans ce gène qui puisse influencer la liaison de la CBG au cortisol ou l'expression de l'ARNm de la CBG.

La variabilité génétique de l'axe corticotrope peut également être observée au niveau des récepteurs aux corticoïdes. Bien que celle-ci soit bien documentée chez l'homme, elle l'est beaucoup moins chez les animaux d'élevage. PERREAU et collab. [1999] ont étudié les propriétés des MR et GR chez des porcs de races Meishan et Large White, montrant que les glucocorticoïdes présentent une plus grande affinité pour les MR que pour les GR. MURÁNI et collab. [2010] ont quant à eux montré l'existence d'un SNP (*single nucleotide polymorphism*) pour le gène *NR3C1*, le gène du GR, qui influence la quantité de cortisol dans la circulation sanguine et la taille des surrénales. Cependant, il y a encore peu de connaissances sur l'efficacité des voies de transduction impliquées dans les fonctions des corticostéroïdes dans ce cas.

Finalement, il est aussi envisageable d'étudier la variabilité génétique au niveau des mécanismes de contrôle de plus haut niveau de l'axe corticotrope (hypophyse, hypothalamus,...). C'est une approche difficile à mettre en place chez les animaux d'élevage, mais une solution envisageable serait la recherche de polymorphismes au sein de gènes encodant des protéines de régulation de l'axe corticotrope et des voies nerveuses impliquées dans sa régulation. Par exemple, MURÁNI et collab. [2010] ont mis en évidence de multiples associations entre des SNP dans *NR3C1* et *AVRP1B*, un récepteur à la vasopressine qui interagit avec la CRH pour

stimuler la sécrétion de l'ACTH par l'hypophyse, suggérant ainsi un rôle de la variabilité de la séquence de ces gènes sur le niveau des réponses de stress.

1.5 Justification du projet de recherche

Chez les animaux d'élevage et chez le porc en particulier, les objectifs de sélection de ces dernières années ont clairement été en faveur des caractères liés à la production (vitesse de croissance, efficacité alimentaire, composition en lipides et protéines des carcasses, ...). Cette sélection génétique s'est faite au détriment des caractères dits fonctionnels, liés à la notion de robustesse chez les animaux d'élevage (qualité des aplombs, résistance aux maladies, ...). La robustesse a été définie par KNAP [2005] comme étant la capacité pour les animaux à combiner un haut potentiel de production à une grande résistance aux conditions environnementales afin de maintenir ce haut potentiel de production dans une large variété d'environnements.
Pour répondre aux problématiques spécifiques du contexte économique et sociétal actuel allant vers de nouveaux objectifs de durabilité des élevages, il est alors nécessaire de faire prendre une part plus importante à la robustesse dans les processus de sélection, et ce dans le but d'obtenir des animaux capables de maintenir leur niveau de production dans une large gamme de systèmes d'élevage et de conditions climatiques, tout en optimisant leur bien-être [MORMÈDE et FOURY, 2009]. Il s'agit donc de rééquilibrer la balance entre caractères de production et caractères de robustesse dans les objectifs de sélection.
Cette balance entre production et robustesse est prédite par la théorie de l'allocation des ressources [BEILHARZ, 1998; RAUW, 2009]. Selon cette théorie, les ressources énergétiques d'un individu sont limitées et leur répartition entre les différentes fonctions métaboliques est optimisée en vue d'obtenir la meilleure adaptation possible de l'individu à son environnement. On peut donc s'attendre à ce que la sélection génétique sur des caractères de production ait redirigé les ressources énergétiques vers les fonctions métaboliques liées aux caractères de production, aux dépens des autres caractères (dont les caractères de robustesse). Si les ressources énergétiques ne sont pas suffisantes pour assurer l'expression du plein potentiel de production, il peut exister une interaction entre les facteurs d'environnement et le génotype des animaux qui diminuerait leur résistance aux facteurs de stress. Selon SCHINCKEL [2010b], les porcs dans les conditions commerciales expriment moins de 80% de leur potentiel génétique. Il existe donc une marge de manœuvre permettant de définir

des objectifs de sélection sur des caractères de robustesse qui permettent de grandement améliorer la résistance des animaux aux facteurs de stress tout en minimisant les pertes sur les caractères de production.

Au cœur de cette problématique, se trouve l'axe corticotrope (hypophyse, hypothalamus, surrénales), le système clef (avec le système nerveux autonome) de la réponse au stress. Chez le porc, l'hormone principale lors de la réponse au stress est le cortisol et il a été montré que celle-ci avait une influence plutôt négative sur les caractères de production [HENNESSY et JACKSON, 1987] et plutôt positive sur des caractères de robustesse [CANARIO et collab., 2009]. Il a été montré chez plusieurs espèces que l'activité de l'axe corticotrope était très sensible à la sélection génétique et disposait d'une bonne héritabilité (voir par exemple [RONECKER, 2010]). Il est donc envisageable d'utiliser la sélection sur cet axe pour orienter les objectifs de sélection vers plus de robustesse chez les animaux d'élevage et chez le porc en particulier.

Il est alors important d'obtenir une meilleure compréhension de cet axe et des mécanismes mis en jeu lors des réponses de stress. En effet, proposer des biomarqueurs aux éleveurs permettrait, à travers une sélection génomique, de satisfaire aux objectifs de durabilité des élevages alliant les objectifs de production de viande d'origine animale et de respect de l'environnement et du bien-être animal.

Dans ce but, on se propose d'étudier la variabilité génétique individuelle qui existe au sein de l'axe corticotrope. Différentes sources de variabilité ont déja été décrites au niveau de la sensibilité à l'ACTH et de la biodisponibilité des glucocorticoïdes (à travers la variabilité génétique existant pour la CBG, leur transporteur), au sein des récepteurs aux glucocorticoïdes et des mécanismes de contrôle supérieur. L'étude de chaque source de variabilité séparément peut apporter une compréhension sommaire des mécanismes en jeu à chaque niveau. Cependant, il est important de prendre en compte le fait que chacun de ces différents niveaux est en interaction avec les autres et qu'ils agissent ensemble au sein d'un même système. En effet, comme expliqué auparavant, l'axe corticotrope est sujet à un rétrocontrôle négatif par les corticoïdes eux-mêmes. Il est donc tout à fait envisageable qu'une mutation à un niveau du système puisse avoir des répercussions à tous les autres niveaux, de telle sorte que le changement initial puisse être atténué, ou, au contraire, amplifié [MORMEDE et TERENINA, 2012]. Afin d'utiliser l'information disponible aux différents niveaux et d'obtenir une compréhension globale de l'architecture des mécanismes de réponses au stress, il est donc nécessaire d'utiliser une

approche intégrative, c'est à dire une approche utilisant des données collectées à différents niveaux du vivant (génétique, transcriptomique, métabolomique, clinique,...) et de les étudier, non pas de manière isolée, mais ensemble et de comprendre comment les changements à un niveau se répercutent sur les autres.

C'est dans ce contexte que le projet SUSoSTRESS a été financé par l'Agence Nationale pour la Recherche (ANR-12-ADAP-0008). L'un de ses objectifs est l'élaboration d'un modèle, chez le porc, des variations génétiques de l'axe corticotrope et de son activité physiologique en lien avec les performances des animaux sur les caractères de robustesse et de production. Pour répondre à cet objectif, une population G0 de porcs Large White a été phénotypée. Le Large White a été choisi car il s'agit de l'une des races largement utilisées en croisement pour la production de viande porcine en France, ultra-sélectionnée sur les caractères de production et dans laquelle on a montré une grande variabilité individuelle de l'activité de l'axe corticotrope [LARZUL et collab., 2015].

Le protocole expérimental mis en place pour ce projet fait intervenir des données complexes comportant plusieurs niveaux d'information. D'une part, la population G0 a été soumis à trois types d'expériences simulant 3 types de stress :

— une injection d'ACTH, l'hormone sécrétée par l'hypophyse qui cible les surrénales et provoque la libération de cortisol (l'hormone principale du stress chez le porc) dans la circulation sanguine. Cette expérience stimule de façon directe l'axe corticotrope et permet de mesurer la sensibilité à l'ACTH, source majeure de variabilité génétique, ainsi que les réponses tissulaires au cortisol. C'est également le test qui a été utilisé ultérieurement pour l'expérience de sélection ;

— un test de contrainte de 10 minutes au cours duquel les animaux les mouvements des animaux ont été restreins. Cette expérience permet d'obtenir une réponse comportant à la fois une composante comportementale et biologique au stress ;

— une injection de lipopolysaccharide (LPS). Le LPS est un composant des parois bactériennes à gram négatif. Son injection déclenche une réaction inflammatoire qui simule un stress de type maladie.

Ces trois types de stress et leur implication dans les composantes des réponses au stress est illustré en figure 1.6.

D'autre part, pour chacune des expériences, des prises de sang ont été réalisées à plusieurs pas de temps et plusieurs types de données ont été mesurées : des variables cliniques, transcriptomiques et métabolomiques. Ainsi, l'analyse de la totalité des données disponibles représente un défi

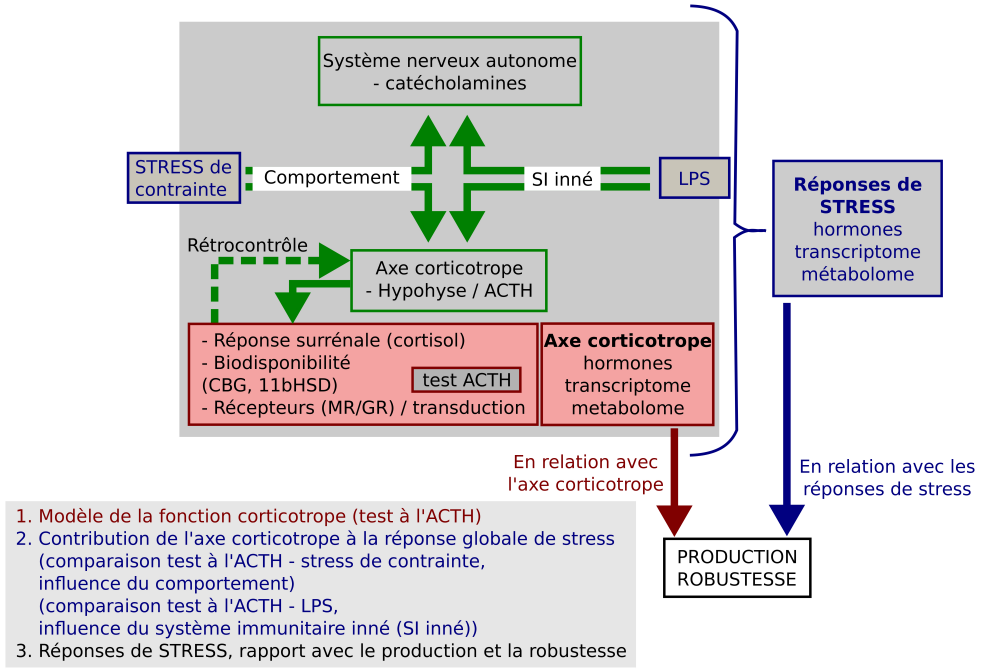


FIGURE 1.6 – Implication de l'axe corticotrope dans les réponses au stress et impact des différents types de stress appliqués aux porcs dans le projet SUSoSTRESS.

pour l'analyse statistique. En effet, idéalement, on souhaiterait analyser ces données dans une démarche de biologie intégrative en combinant à la fois les différentes mesures entre elles et les données issues des 3 expériences entre elles, tout en tenant compte de l'aspect longitudinal des données. Une telle approche serait novatrice pour l'analyse de données biologiques mais demande le développement d'une méthodologie adaptée à la fois à l'intégration multi-tableaux et aux données longitudinales.

Avant de décrire un ensemble de méthodes d'analyses statistiques existantes pour tenter de répondre à cette question méthodologique, nous nous attachons, dans la section suivante, à faire le rappels des grands principes en biologie moléculaire.

1.6 Rappels de biologie moléculaire

La cellule est l'unité biologique, structurelle et fonctionnelle qui compose tous les organismes vivants. C'est la plus petite unité vivante capable de se reproduire. L'activité cellulaire est coordonnée et programmée à partir d'une information stockée sur des macromolécules appelées acides désoxyribonucléiques (ADN). Elle est composée de 4 types de nu-

cléotides : adénine (A), cytosine (C), guanine (G) ou thymine (T). Dans la cellule, l'ADN est structurée en chromosomes, eux mêmes structurés en gènes. Cette information génétique est transcrite en acides ribonucléiques messagers (ARNm) qui servent d'intermédiaires dans la production de protéines. Ces dernières assurent un vaste éventail de fonctions au sein de la cellule, telles que signalisation, transport, etc. Le gène, l'ARNm et la protéine sont les 3 éléments principaux du dogme central de la biologie moléculaire. La définition initiale de ce dernier spécifiait qu'un gène permettait la synthèse d'un ARNm par le processus de la *transcription* et qu' un ARNm synthétisait une protéine par le processus appelé *traduction* [CRICK et collab., 1970]. Les progrès en biologie moléculaire ont cependant montré que les relations entre gènes, ARNm, protéines et fonctions cellulaires sont bien plus complexes avec par exemple, l'existence de variants pour un même gène, de variations des modifications post-transcriptionnelles, d'épissages alternatifs concernant les ARNm ou encore de modifications post-traductionnelles et une information pouvant aller aussi bien du gène vers la protéine que de la protéine vers le gène [SHAPIRO, 2009].

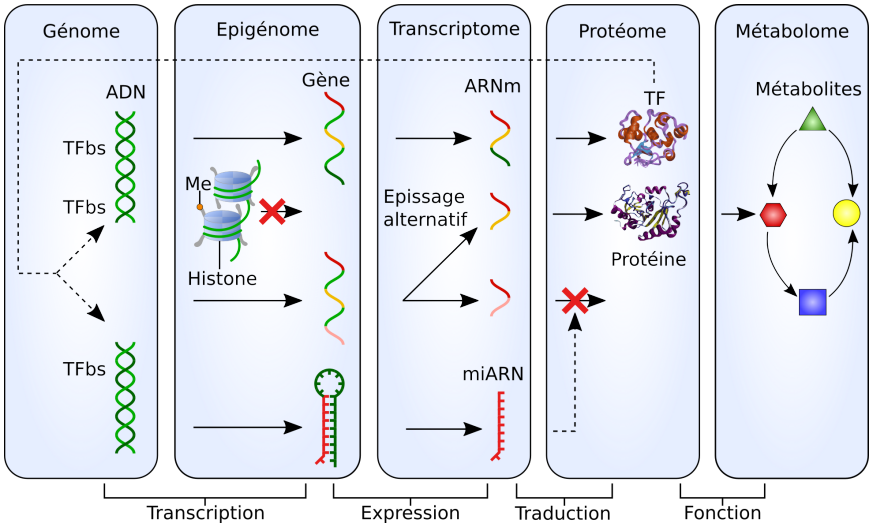


FIGURE 1.7 – Dogme central de la biologie moléculaire. L'expression des gènes dans une cellule passe par la transcription de l'ADN en ARN messager qui est lui même traduit en protéine. De nombreuses sources de variations existent aux différents niveaux d'activité cellulaire et expliquent les différences au niveau macroscopique du phénotype (caractères observables au niveau individuel (d'après RITCHIE et collab. [2015])).

Le fonctionnement cellulaire est illustré en figure 1.7. L'expression des gènes ainsi que la façon dont celle-ci est régulée a pour résultat l'expression de phénotypes macroscopiques observables. Comme pour le dogme

central de la biologie moléculaire, l'évolution des connaissances a fait évoluer la complexité des hypothèses portant sur l'influence des différents niveaux cellulaire sur l'expression d'un caractère (voir figure 1.8). Ainsi, nous sommes passés d'une hypothèse expliquant une relation linéaire entre les variations existant à chaque niveau cellulaire à une hypothèse expliquant les caractères complexes par une combinaison des variations existant aux différents niveaux et dans leurs interactions avec l'environnement.

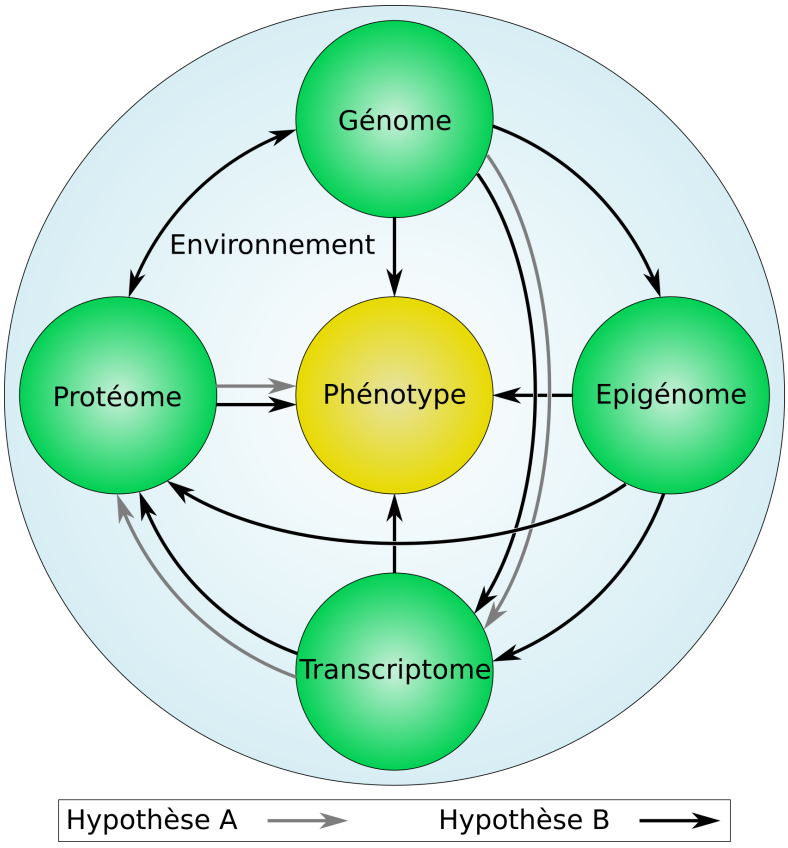


FIGURE 1.8 – Hypothèses alternatives de l'influence des différents niveaux cellulaires sur l'expression de caractères complexes. L'hypothèse A (flèche grise) décrit une relation hiérarchique entre les différents niveaux, de telle sorte que la variation au niveau de l'ADN entraine une variation des ARNm qui entrainent des modifications aux niveaux suivants de façon linéaire. L'hypothèse B (flèche noire) décrit une interaction entre les variations à tous les niveaux cellulaires et avec l'environnement pour expliquer un caractère complexe (d'après [RITCHIE et collab., 2015]).

Dans ce cadre, des approches intégratives ont été développées pour étudier les relations entre gènes et caractères observables en tenant compte de la complexité des interactions entre les différents niveaux cellulaires.

La section suivante, après une description plus précise des types de données disponibles dans les approches intégratives s'attache à décrire l'ensemble des méthodes d'analyses statistiques intégratives utilisées au cours de ce travail.

1.7 Approches intégratives en biologie

Depuis les 10 dernières années, les progrès réalisés en biologie moléculaire, et plus particulièrement dans les techniques de séquençage à haut débit, ont permis d'augmenter considérablement la quantité de données disponibles et de diminuer en parallèle les coûts expérimentaux. En conséquence, les protocoles expérimentaux ont pu être compliqués et il est à présent possible de mesurer des données à plusieurs niveaux du vivant.
Ces progrès ont donc entrainé le développement des approches de type intégrative en biologie (*systems biology*) pour lesquelles, un intérêt existe depuis le début des années 50 [WIENER, 1948]. Cette approche analytique globale s'attache à étudier les organismes en intégrant des données issues de plusieurs sources afin de mieux appréhender les processus complexes se produisant au sein d'un système. L'analyse conjointe des données aux différents niveaux permet ainsi de tirer des informations pertinentes des relations existant entre eux et il devient alors possible de comprendre comment ils interagissent et agissent ensemble au sein d'un même système.
La génétique des systèmes (*systems genetics*) est une branche particulière de la biologie intégrative qui se propose d'analyser globalement les facteurs moléculaires sous-jacents à la variabilité individuelle existant au sein des phénotypes physiologiques ou cliniques dans une population. Elle ne se concentre pas seulement sur les variations génétiques, mais aussi sur les phénotypes intermédiaires entre modifications de séquence d'ADN et phénotypes cliniques : expression de gènes, niveaux de protéines et de métabolites, interactions entre gènes, interactions gènes et environnements, etc [CIVELEK et LUSIS, 2014].

Après un rappel sur les données 'omiques, nous introduisons les deux types de stratégies possibles pour l'intégration de données biologiques : « supervisées » et « non supervisées ». Nous détaillons ensuite le principe des méthodes supervisées et non supervisées abordées dans cette thèse. Enfin, nous abordons le cas particulier des données longitudinales.

1.7.1 Description des données 'omiques

Les différents types de données utilisables dans les approches intégratives et leurs interactions sont illustrées dans la figure 1.9. On se réfère généralement au terme de données *'omiques* :

— le *génome* correspond à l'ensemble des gènes contenus dans une cellule ;

— le *transcriptome* correspond à l'ensemble des ARN produits à partir du génome à un instant t. Chaque cellule possède le même génome. Cependant elles n'expriment pas toutes le même sous-ensemble de leur patrimoine génétique. En effet, la spécialisation des cellules, tant dans leurs formes que dans leurs fonctions a pour résultat l'expression de gènes différents pour des cellules différentes ;

— le *protéome* correspond à l'ensemble des protéines produites à partir des ARNm. Les protéines synthétisent, détruisent et régulent l'activité des métabolites et peuvent varier d'une cellule à l'autre en fonction de la spécialisation de ces dernières. Pour de nombreuses raisons, il n'y a pas de corrélation directe entre le niveau d'expression d'un ARNm et celui de la protéine dérivée. Par exemple, certains processus de régulation peuvent affecter la production de la protéine lors de la traduction ou encore sa vitesse de dégradation. De plus, il est fréquent que des protéines soient exportées en dehors de leur cellule de production ;

— le *métabolome* correspond à l'ensemble des métabolites (petites molécules) produits dans un organisme. Il s'agit du niveau ultime d'expression des gènes et sa mesure permet de donner l'état physiologique d'une cellule.

Les relations entre génome et expression d'un caractère sont complexes. D'une part, l'expression d'un caractère est rarement relié à l'expression d'un seul gène et est bien souvent multigénique et peut être influencé par les interactions entre environnement et génome. En effet, les cellules sont capables de s'adapter à des modifications de l'environnement en modulant l'expression de leurs gènes. Cela se traduit par une modification de l'abondance d'ARN correspondant, ce qui entraine des modifications de la quantité de protéines au sein des cellules, et donc à une baisse ou une hausse de leur activité.

L'intégration de ces données doit se faire par l'application de modèles mathématiques adaptés à la description des relations entre les différents types de données en jeu.

1.7.2 Stratégies d'intégration de données

Lorsque l'on intègre des données biologiques organisées en plusieurs tableaux, on cherche à étudier les relations existant entre eux. On peut alors choisir deux types de stratégies :

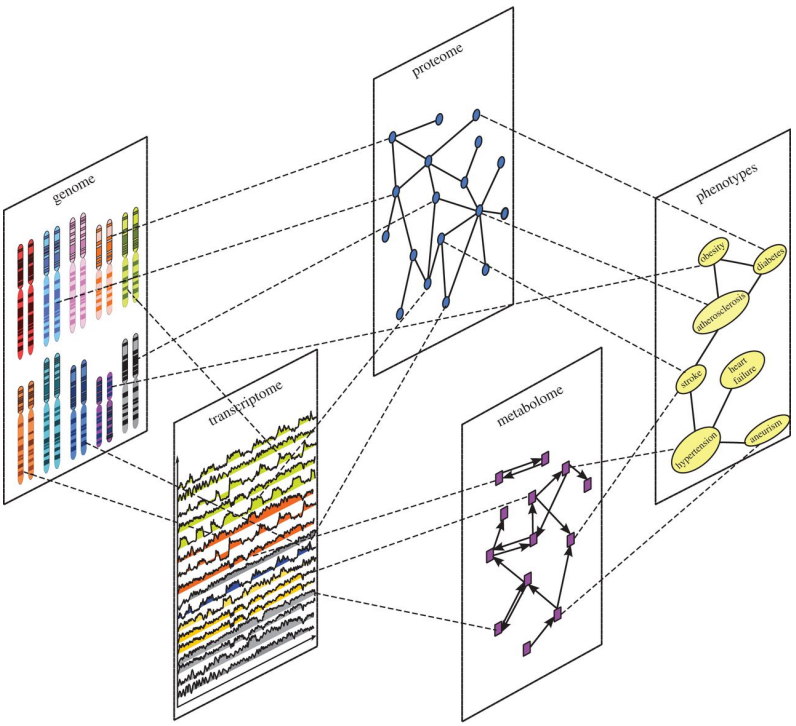


FIGURE 1.9 – Schéma des informations disponibles au sein d'une cellule en biologie moléculaire. Il illustre les types de données qu'il est possible de collecter et d'intégrer (d'après [GLIGORIJEVIĆ et PRŽULJ, 2015]).

— dans le premier cas, on peut chercher à étudier les relations entre les jeux de données de façon symétrique, sans chercher à expliquer l'un par l'autre. On parle alors d'intégration « supervisée » ;

— dans le second cas, on ne considère pas les jeux de données de façon équivalente, mais on cherche à étudier une relation asymétrique entre eux en cherchant à expliquer l'un par l'autre. Il s'agit alors d'intégration « non supervisée » ;

Ces deux approches peuvent elles-même être divisées selon la nature de l'analyse réalisée. Que l'analyse soit supervisée ou non, il est possible de baser l'études des relations entre tableaux en explorant les structures de corrélations (ou covariance) existant directement entre les jeux de données. On peut citer par exemple l'Analyse Canonique des Corrélations (ACC) [HOTELLING, 1936] qui est à la base de toutes les méthodes d'analyses multidimensionnelles et étudie, de façon symétrique, les relations entre deux jeux de données mesurées sur les mêmes individus. De même, la régression PLS (pour *Partial Least Square regression*) [WOLD, 1985] explore les structures de corrélations entre deux jeux de données et peut être utilisée dans une approche non-supervisée (en mode canonique) ou

supervisée, en fonction de l'algorithme appliqué.

Une autre stratégie possible est l'utilisation de méthodes basées sur la décomposition spectrale (en vecteurs et valeurs propres). Ce type d'approches passe par la construction d'une matrice « globale » constituée d'une concaténation de l'ensemble des tableaux pondérés et concaténés en un seul. Parmi elles, on peut citer les exemples de l'Analyse Factorielle Multiple (AFM) [Escofier et Pages, 1994] et des différentes approches dérivées de STATIS (pour Structuration à Trois Indices de la Statistique [Glacon, 1981; Lavit, 1988; Lavit et collab., 1994; L'Hermier des Plantes, 1976]) : STATIS, dual-STATIS, ou encore les Analyses Triadiques Partielles (ATP) [Thioulouse et Chessel, 1987] par exemple. Ces approches sont toutes des approches non supervisées. En AFM, la pondération est faite en fonction de la variabilité existant au sein de chaque tableau individuel. Pour STATIS et ses dérivés, elle est faite en fonction de la similarité entre les tableaux et une matrice consensus résumant au mieux l'ensemble des jeux de données. Cette matrice consensus peut être construite à partir de la matrice de coefficients RV [Escoffier, 1976] (des produits scalaires entre individus dans le cas de STATIS, des matrices de covariance dans le cas de dual-STATIS et des tableaux initiaux dans le cas de l'ATP). On peut également citer en exemple la régression inverse par tranches ou SIR (pour *Sliced Inverse Regression*) [Li, 1991]. Cette dernière est une méthode de régression semi-paramétrique qui applique à un jeu de données multivarié, une décomposition spectrale dans un modèle de régression non-linéaire pour expliquer une variable cible.

Les méthodes citées ici, le sont soit pour leur nature classique en intégration de données biologiques (comme l'ACC et la régression PLS en mode canonique), soit car elles ont été utilisées dans ce travail de thèse. Ainsi, la régression PLS et l'AFM ont été utilisées dans le chapitre 2 pour étudier les relations entre variables cliniques et transcriptome dans deux expériences du projet SUSoSTRESS. Ces méthodes ont permis l'identification de gènes dont les variations d'expression dans les réponses aux stress sont particulièrement liées aux manifestations physiologiques de ces réponses. Les approches SIR et STATIS ont, quant à elles, été impliquées dans la méthode développée lors du chapitre 3, adaptant la SIR à l'analyse de données répétées. Il existe cependant une multitude d'autres approches pour l'analyse conjointe de T tableaux. On peut citer par exemple la décomposition de type Tucker [Kruskal, 1989], qui propose de décomposer les données à 3 indices en un *produit tensoriel*. Une autre méthode consiste à utiliser les rotations Procrustes [Ten Berge, 1977]. Cette approche consiste à transformer les matrices de données initiales de manière à ce qu'elles soient

en moyenne les plus proches possibles des transformations de toutes les autres matrices. Il est également possible de s'intéresser à une approche de style *Generalized Orthogonal Multiple Co-Inertia Analysis PLS* [VIVIEN et SABATIER, 2003] qui généralise le cadre de la régression PLS aux données multiblocs.

Nous faisons le choix dans ce chapitre introductif de ne présenter que les méthodes utilisées dans ce travail de thèse et récapitulées dans le tableau 1.1.

TABLEAU 1.1 – Tabeau récapitulatif des méthodes d'intégration de données abordées dans cette section et des chapitres de la thèse où elles sont abordées.

	corrélations	Décomposition spectrale
supervisé	régression PLS (Chapitre 2)	SIR (Chapitre 3)
non-supervisé	CCA (-) PLS (-)	AFM (Chapitre 2) STATIS (Chapitre 3)

1.7.3 Intégration non-supervisée

Cette section décrit les approches d'intégration non-supervisées utilisées dans le cadre de cette thèse : AFM et dual-STATIS. Ces deux méthodes d'analyses multi-tableaux se basent sur la décomposition spectrale pour définir des composantes principales consensus entre les différents tableaux de données. Aucune approche non-supervisée basée sur l'analyse des corrélations (ou covariance) n'a été utilisée dans les travaux décrits dans ce manuscrit et n'est donc abordée ici.
L'AFM et dual-STATIS sont très similaires, et se déroulent en deux étapes :

1. la recherche de pondérations adéquates pour les différents tableaux ;
2. l'ACP de la matrice globale basée sur la concaténation des données pondérées.

Après avoir défini les notations utilisées dans cette section, nous décrirons rapidement chacune de ces méthodes.

Notations

Soit $\mathbf{X}$ notre jeu de données répétées observées. Plus précisément, p variables ont été mesurées sur les mêmes n individus un nombre T de fois avec $n > p$. Les observations dans $\mathbf{X}$ sont donc référencées par 3 indices : $\mathbf{X} = (x_{ijt})_{i=1,\dots,n, j=1,\dots,p, t=1,\dots,\mathrm{T}}$.
On note alors :

— $\mathbf{X}_{..t}$, la matrice $n \times p$ des observations au temps t, $\mathbf{X}_{.j.}$, la matrice $n \times \mathrm{T}$ des observations pour la variable j et $\mathbf{X}_{i..}$, la matrice $j \times \mathrm{T}$ de la $i^{\text{ème}}$ observation ;

— $\mathbf{x}_{.jt}$, le vecteur de taille n contenant toutes les observations de la variable j au temps t, $\mathbf{x}_{i.t}$, le vecteur de taille p contenant toutes mesures de l'individu i au temps t et $\mathbf{x}_{ij.}$, le vecteur de taille T contenant les observations à tous les pas de temps de l'individu i pour la variable j.

Ces tableaux à 3 indices peuvent aussi être appelés tableaux cubiques et son illustrés en figure 1.10.

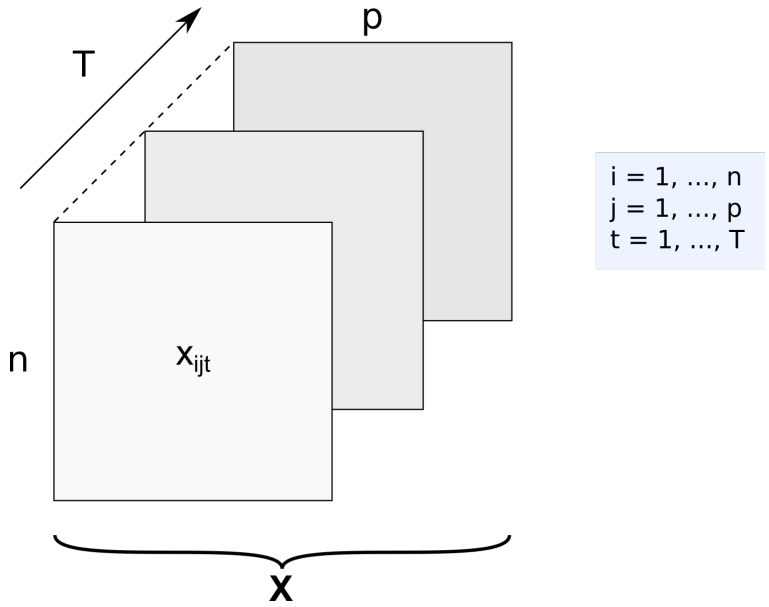


FIGURE 1.10 – Illustration des données dites « cubiques ». Elles sont constituées de p variables mesurées T fois sur les n mêmes individus.

STATIS et dual-STATIS

Les méthodes STATIS et dual-STATIS ont été décrites pour la première fois dans L'HERMIER DES PLANTES [1976] (voir aussi ABDI et collab. [2012] pour une revue de littérature abondante sur STATIS et ses dérivés). Il s'agit d'une approche multivariée utilisée pour explorer les données structurées en blocs, comme c'est le cas pour les données cubiques par exemple.
L'approche STATIS est utilisée lorsque l'on mesure plusieurs blocs de variables sur les n mêmes individus. Les variables de chaque bloc peuvent être différentes ou identiques. L'approche dual-STATIS est très proche de STATIS mais s'applique lorsque l'on a plusieurs blocs d'individus sur

lesquels on mesure un même ensemble de variables. Là où STATIS s'intéresse aux matrices de produits scalaires entre observations, dual-STATIS s'intéresse aux structures de corrélations. Enfin, lorsque l'on étudie un jeu de données de mesures répétées des mêmes variables sur les mêmes individus (jeu de données cubiques), comme c'est le cas pour des données longitudinales par exemple, les deux analyses peuvent s'appliquer selon le but recherché ou bien, on peut utiliser l'analyse Triadique Partielle (ATP) [THIOULOUSE et CHESSEL, 1987]).

Dans le cadre de la recherche de potentiels biomarqueurs, il est plus pertinent de s'intéresser aux structures de corrélations entre variables des différents blocs. C'est pourquoi, dans le cadre de ce travail, on détaillera l'approche dual-STATIS. Dans cette section, nous décrirons les différentes étapes de dual-STATIS. L'idée principale de la méthode est d'étudier les relations entre les structures de corrélations et l'évolution de ces relations entre les différents blocs de variables, par la construction d'une matrice *compromis* qui représente le meilleur résumé de la structure du jeu de données. On se limite au cas des données cubiques, telles que $\mathbf{X}$ décrit en figure 1.10 et on reformule la méthode dual-STATIS classique comme une décomposition en valeur singulière généralisée (GSVD pour *Generalized Singular Value ecomposition*). Les données sont supposées centrées réduites et on définit :

$$\mathbf{Z}_{..t} = (\mathbf{X}_{..t} - \mathbf{1}_n \bar{\mathbf{x}}_t^\top) \Delta_t$$

où $\bar{\mathbf{x}}_t = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_{i.t} \in \mathbb{R}^p$, $\mathbf{1}_n$ est un vecteur de taille n dont tous les éléments sont égaux à 1 et $\Delta_t = \text{Diag}(\hat{\sigma}_{jt}^{-1})_{j=1,\dots,p}$ avec $\hat{\sigma}_{jt}^2$ la variance empirique du vecteur $\mathbf{x}_{.jt} = (x_{ijt})_{i=1,\dots,n}$.

Dual-STATIS se décompose alors en deux étapes principales : *la recherche de la pondération optimale* et *l'ACP de la matrice globale*.

Recherche de la pondération optimale En dual-STATIS, la recherche de pondération peut aussi être appelée « étude de l'interstructure » car elle permet d'étudier les relations entre les différents tableaux.

On choisit pour objet représentatif d'un tableau t sa matrice de covariance Γ_t :

$$\Gamma_t = \frac{1}{n} \mathbf{Z}_{..t}^\top \mathbf{Z}_{..t}$$

ou, si l'on choisit de normer les matrices de covariances afin de leur donner une norme de Frobenius de 1 :

$$\widetilde{\Gamma}_t = \frac{\Gamma_t}{\|\Gamma_t\|_{\text{F}}}$$

avec $\|\Gamma_t\|_F^2 = \text{Trace}(\Gamma_t^\top \Gamma_t)$.

Ce normage permet de comparer deux matrices $\widetilde{\Gamma}_t$ en mesurant le cosinus de l'angle formé par celles-ci, sans être influencé par leurs différences d'échelle. Normer les matrices de covariance de cette matière est équivalent à normer les tableaux initiaux et à travailler avec des matrices $\widetilde{\mathbf{Z}}_{..t}$ où :

$$\widetilde{\mathbf{Z}}_{..t} = \frac{\mathbf{Z}_{..t}}{\sqrt{\|\Gamma_t\|_F}}.$$

La proximité entre les $\widetilde{\Gamma}_t$ est estimé par un coefficient de similarité entre les T matrices $\widetilde{\Gamma}_t$. Ce coefficient de similarité est un produit scalaire de Frobenius et s'écrit :

$$\tilde{c}_{tt'} = \langle \widetilde{\Gamma}_t, \widetilde{\Gamma}_{t'} \rangle_F = \frac{\langle \Gamma_t, \Gamma_{t'} \rangle_F}{\|\Gamma_t\|_F \|\Gamma_{t'}\|_F}.$$

La matrice $\widetilde{\mathbf{C}} = (\tilde{c}_{tt'})_{t,t'=1,\dots,T}$ contient les coefficients de similarité entre les Γ_t et représente l'information globale sur les similitudes et différences entre les structures de corrélations dans les différents blocs de variables. Pour étudier les relations entre les tableaux, on résout le problème d'optimisation suivant :

$$\mathbf{u}_1 = \underset{\mathbf{u}=(u_1,\dots,u_T)^\top \in \mathbb{R}^T, \|\mathbf{u}\|=1}{\arg\max} \Big\| \sum_{t=1}^{T} u_t \widetilde{\Gamma}_t \Big\|_F^2 \tag{1.1}$$

Ce problème d'optimisation est équivalent à trouver une matrice $\Gamma^* = \sum_{t=1}^{T} u_t \widetilde{\Gamma}_t$ qui présente la plus forte similarité moyenne (au sens du produit scalaire de Frobenius) avec les matrices $(\widetilde{\Gamma}_t)_{t=1,\dots,T}$. Une solution est donnée par u_1, le premier vecteur propre de $\widetilde{\mathbf{C}}$. La matrice Γ^* est alors appelée « matrice compromis » et est notée Γ^c. En général, on multiplie les poids trouvés par un même facteur de sorte que leur somme soit égale à 1 :

$$\alpha_t = \frac{\sqrt{\mu_1} u_{1t}}{\sum_{t'=1}^{T} \sqrt{\mu_1} u_{1t'}} = \frac{u_{1t}}{\sum_{t'=1}^{T} u_{1t'}}.$$

La matrice compromis Γ^c est alors une combinaison convexe des matrices $\widetilde{\Gamma}_t$:

$$\Gamma^c = \sum_{t=1}^{T} \alpha_t \widetilde{\Gamma}_t.$$

Cette matrice compromis capture les structures de corrélations les plus stables à travers les blocs de données.

ACP de la matrice globale La seconde étape de dual-STATIS fait l'ACP de la matrice compromis I^c. Il est équivalent de faire la SVD généralisée [1] du triplet $(\widetilde{\mathbf{Z}}, \mathbb{I}_p, \mathbf{D})$ où :

— $\widetilde{\mathbf{Z}} = \begin{bmatrix} \widetilde{\mathbf{Z}}_{..1} \\ \vdots \\ \widetilde{\mathbf{Z}}_{..\mathrm{T}} \end{bmatrix}$ est la matrice globale de taille a $(n\mathrm{T}) \times p$ des données concaténées,

— $\mathbb{I}_p$, la matrice carrée itentité de taille p (qui est la métrique dans l'espace des individus),

— $\mathbf{D} = \frac{1}{n}\mathrm{Diag}(\alpha_1, \ldots, \alpha_\mathrm{T}) \otimes \mathbb{I}_n$, où $\otimes$ est le produit de Kronecker, est la métrique dans l'espace des variables (qui correspond à une pondération des observations). C'est une matrice carrée de taille $(n\mathrm{T}) \times (n\mathrm{T})$.

La SVD généralisée permet de décomposer $\widetilde{\mathbf{Z}}$ en 3 matrices :

$$\widetilde{\mathbf{Z}} = \mathbf{P}\Lambda\mathbf{Q}^\top \qquad \text{avec } \mathbf{P}^\top\mathbf{D}\mathbf{P} = \mathbf{Q}^\top\mathbf{Q} = \mathbb{I}_r \text{ et } \Lambda = \mathrm{Diag}(\sqrt{\lambda_k})_{k=1,\ldots,r},$$

où r est le rang de la matrice $\widetilde{\mathbf{Z}}$. On a alors :

— $\mathbf{P}$ est la matrice de taille $(n\mathrm{T}) \times r$ des vecteurs propres de $\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top\mathbf{D}$,

— $\mathbf{Q}$ est la matrice de taille $p \times r$ des vecteurs propres de $\widetilde{\mathbf{Z}}^\top\mathbf{D}\widetilde{\mathbf{Z}} = I^c$

L'approche fournit un ensemble de composantes principales globales sur lesquelles il est possible :

— de projeter des positions compromis pour les individus et les variables, ce qui permet de visualiser les relations les plus importantes ;

— de projeter les positions spécifiques de chaque bloc de variables pour visualiser comment elles se positionnent par rapport au compromis.

La méthodologie développée en chapitre 3 permet d'obtenir le même type de représentations. Le détail des coordonnées pour chacune de ces positions et comment elles sont obtenues y est décrit.

Analyse Factorielle Multiple

Comme dual-STATIS, l'Analyse Factorielle Multiple (AFM) [Escofier et Pages, 1994] est une méthode d'analyse multi-tableaux qui permet d'analyser des données organisées en blocs décrivant un même ensemble d'individus. L'objectif de l'AFM est également de chercher les similarités entre les blocs de variables analysés et de fournir à la fois une vision globale des individus sur l'ensemble des blocs de variables et une vision des relations existant entre les variables des différents blocs.
Comme dual-STATIS, elle se base sur la décomposition spectrale d'une matrice globale et se décompose en deux étapes.

1. La SVD généralisée est une SVD dans laquelle les métriques dans les espaces des individus et/ou des variables ne sont pas les métriques identités usuelles.

Recherche de la pondération optimale Chaque bloc de variables t correspond à un tableau $\mathbf{X}_{t,t=1,\dots,\mathrm{T}}$, tel que défini en figure 1.10. Dans le cas de l'AFM, la recherche de la pondération optimale se fait avec la volonté d'équilibrer l'inertie axiale maximum de chaque bloc de variables pour qu'elle soit égale à 1 pour tous les blocs . Pour parvenir à équilibrer les blocs de variables de cette manière, on pondère les tableaux de donnée par un coefficient inversement proportionnel à leur inertie afin que les tableaux les plus variables ne prennent pas le pas sur les autres dans la construction de la matrice globale [ABDI et collab., 2013].

On commence donc dans un premier temps par calculer pour chaque bloc de variables t, la matrice de corrélation Γ_t :

$$\Gamma_t = \frac{1}{n}\mathbf{X}_{..t}^{\top}\mathbf{X}_{..t}$$

L'ACP de chaque matrice $\mathbf{X}_{..t}$ fournit T ensembles de valeurs propres $\lambda^{(t)}_{q(q=1,\dots,r)}$ où r est la dimension de projection. Pour que l'ACP de $\mathbf{X}_{..t}$ fournisse une première valeur propre $\lambda_1^{(t)}$, on définit la nouvelle matrice Γ_t^* :

$$\Gamma_t^* = \frac{\Gamma_t}{\lambda_1^{(t)}}$$

La matrice de corrélations globale s'écrit alors :

$$\Gamma^c = \sum_{t=1}^{\mathrm{T}} \frac{\Gamma_t}{\lambda_1^{(t)}}$$

Cette pondération est équivalente à normaliser chaque matrice de corrélation pour les rendre comparables entre elles, comme l'on réduirait des variables en les divisant par leurs écart-types si l'on travaillait sur des vecteurs.

ACP de la matrice globale Comme en dual-STATIS, l'ACP de la matrice de corrélations globales Γ^c est équivalente à faire la GSVD du triplet $(\mathbf{X}, \mathbb{I}_n, \mathbf{A})$ où :

— $\mathbf{X} = \left[\mathbf{X}_{..1} \dots \mathbf{X}_{..\mathrm{T}}\right]$ est la matrice globale de taille $n \times (\mathrm{T}p)$ des données concaténées,

— $\mathbf{A} = \mathrm{Diag}(\frac{1}{\lambda_1^{(1)}}, \dots, \frac{1}{\lambda_1^{(\mathrm{T})}}) \otimes \mathbb{I}_p$, où $\otimes$ est le produit de Kronecker, est la métrique associée aux variables et est une matrice carrée de taille $(p\mathrm{T}) \times (p\mathrm{T})$

— $\mathbb{I}_n$ est la matrice identité de taille n associée aux observations.

Cette GSVD permet de décomposer $\mathbf{X}$ en 3 matrices :

$$\mathbf{X} = \mathbf{P}\Lambda\mathbf{Q}^\top \qquad \text{avec } \mathbf{P}^\top \mathbb{I}_n \mathbf{P} = \mathbf{Q}^\top \mathbf{A}s\mathbf{Q} = \mathbb{I}_r \text{ et } \Lambda = \text{Diag}(\sqrt{\delta_k})_{k=1,\ldots,r},$$

où r est le rang de la matrice $\mathbf{X}$ et δ_k est la $k^{\text{ième}}$ valeur propre de $\mathbf{X}$. On a alors :

— $\mathbf{P}$ est la matrice de taille $n \times r$ des vecteurs propres de $\mathbf{X}\mathbf{X}^\top \mathbb{I}_n$,

— $\mathbf{Q}$ est la matrice de taille $p\text{T} \times r$ des vecteurs propres de $\mathbf{X}^\top \mathbb{I}_n \mathbf{X}$

L'approche fournit, d'une part, un ensemble de composantes principales globales dont la projection permet de positionner les observations les unes par rapport aux autres en tenant compte de l'ensemble des blocs de variables. D'autre part, elle fournit les corrélations des variables de chaque bloc avec les composantes globales, ce qui permet d'identifier les variables liées entre elles d'un bloc de variables à l'autre. Enfin, elle fournit également les corrélations entre les composantes des ACP spécifiques à chaque bloc et de l'ACP globale, ce qui permet d'identifier les blocs de variables qui contribuent le plus à la construction de chaque composante globale (et ainsi, identifier les blocs de variables les plus proches entre eux).

1.7.4 Intégration supervisée

Cette section décrit deux approches d'intégration supervisée utilisées dans le cadre de cette thèse : la régression PLS et la SIR. Ces deux approches permettent de faire l'intégration entre deux jeux de données en cherchant à expliquer un tableau $\mathbf{Y}$ par un tableau $\mathbf{X}$. Les deux approches diffèrent à la fois par la nature de $\mathbf{Y}$ et par les méthodes statistiques sous-jacentes. En effet, la régression PLS se base sur l'étude des corrélations entre les deux jeux de données et peut être appliquée lorsque $\mathbf{Y}$ est un vecteur ou quand il s'agit d'une matrice. La SIR, quant à elle est une méthode de régression semi-paramétrique qui se base sur une décomposition spectrale appliquée dans un modèle de régression et s'applique lorsque $\mathbf{Y}$ est un vecteur mesurant une variable cible réelle.
Après avoir défini les notations relatives au cadre de ces méthodes, nous détaillerons plus précisément chacune de ces approches.

Notations

Soit $\mathbf{X}$ notre jeu de données observé constitué de p variables mesurées sur n individus. Les observations dans $\mathbf{X}$ sont donc référencées par 2 indices : $\mathbf{X} = (x_{ij})_{i=1,\ldots,n, j=1,\ldots,p}$.
On note alors :

— $\mathbf{x}_{.j}$, le vecteur de taille n contenant toutes les observations de la variable j ;

— $\mathbf{x}_{i.}$, le vecteur de taille p contenant toutes mesures de l'individu i.

Enfin, on observe également sur nos n mêmes individus, une variable réelle $\mathbf{Y}$, de dimension $n \times q$, qui peut être univariée ($q = 1$) ou multivariée ($q > 1$).

Ces données sont illustrées en figure 1.11.

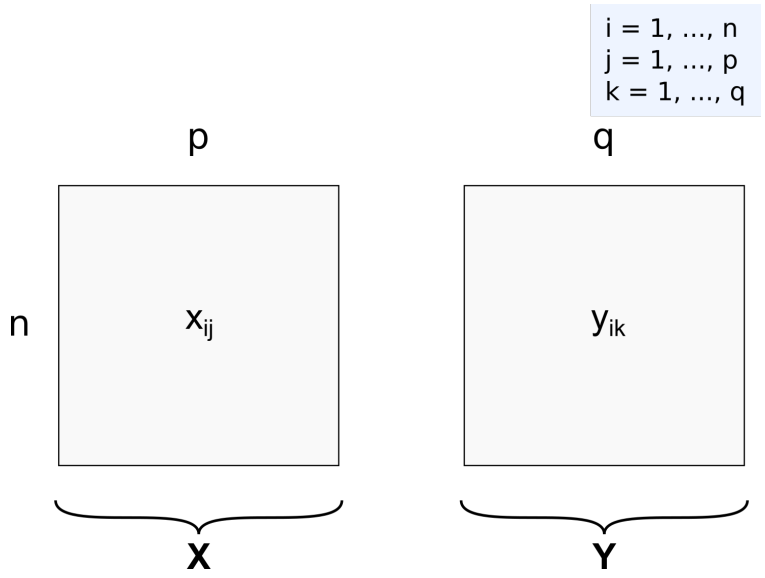


FIGURE 1.11 – IIllustration des données utilisées pour les approches d'intégration supervisées. Elles sont constituées de 2 tableaux : $\mathbf{X}$, constitué de p variables et $\mathbf{Y}$ constitué de q variables ($q \geq 1$). Les deux tableaux ont été mesurés sur les n mêmes individus.

Régression PLS

La régression PLS (*Partial Least Square*) [WOLD, 1985] repose sur la décomposition simultanée de $\mathbf{X}$ en vecteurs *loadings* et *variables latentes* associées. Le principe général de l'approche est de chercher un modèle de régression linéaire de $\mathbf{Y}$ sur un ensemble de composantes orthogonales, construites (comme les variables canoniques en ACC) à partir des combinaisons linéaires des p variables initiales de $\mathbf{X}$. En régression PLS, ces composantes sont construites itérativement par régressions locales successives, de manière à ce qu'elles soient les plus liées à la variable $\mathbf{Y}$ à prédire au sens de la covariance empirique. En d'autres termes, chaque composante de la PLS est construite de manière à optimiser la covariance entre chaque combinaison linéaire des deux tableaux :

$$\underset{\|u\|=\|v\|=1}{\arg\max}\ \mathrm{cov}(\mathbf{X}u_l, \mathbf{Y}v_l) \qquad l = 1, \ldots, r \quad (r \leq \min p, q) \tag{1.2}$$

où $\mathbf{X}u_l = \eta_l$ et $\mathbf{Y}v_l = \xi_l$ sont les *composantes de la PLS*, u_l et v_l, les *vecteurs loadings* et r est la dimension recherchée.
L'algorithme est itératif et se déroule en r étapes. À chaque étape l ($l = 1, \ldots, r$), il s'agit de :

1. chercher les composantes $\eta_1 = \mathbf{X}\mathbf{u}_1$ et $\xi_1 = \mathbf{Y}\mathbf{v}_1$ où $\mathbf{u}$ et $\mathbf{v}$ sont solutions de :
$$\underset{\|u\|=\|v\|=1}{\arg\max}\ \mathrm{cov}(\mathbf{X}\mathbf{u}, \mathbf{Y}\mathbf{v})\ ;$$
2. faire les régressions partielles de $\mathbf{X}$ sur les composantes η et de $\mathbf{Y}$ sur ξ pour obtenir les coefficients de régression partielle c_l et d_l où :
$$c_l = \mathbf{X}^\top \eta / (\eta^\top \eta) \quad \text{et} \quad d_l = \mathbf{Y}^\top \xi / (\xi^\top \xi)\ ;$$
3. faire la déflation des matrices $\mathbf{X}$ et $\mathbf{Y}$. La déflation consiste à extraire les résidus des régressions de $\mathbf{X}$ et $\mathbf{Y}$ sur les remplacer une matrice par une autre, construite à partir de la matrice d'origine. Dans le cadre de la régression PLS, l'étape de déflation consiste à recalculer des matrices $\mathbf{X}$ et $\mathbf{Y}$ telles que :
$$\mathbf{X}_l = \mathbf{X}_{l-1} - \eta_l c_l^\top \quad \text{et} \quad \mathbf{Y}_l = \mathbf{Y}_{l-1} - \eta_l v_l^\top$$
où $\mathbf{X}_l$ est la matrice $\mathbf{X}$ à l'itération l (de même pour $\mathbf{X}_{l-1}$, $\mathbf{Y}_l$ et $\mathbf{Y}_{l-1}$) ;
4. recommencer jusqu'à $l = r$.

La méthode est proche de l'ACC. La différence entre les deux méthodes se trouve dans le critère à optimiser qui est la corrélation entre deux tableaux de données $\mathbf{X}$ et $\mathbf{Y}$ en ACC contre la covariance en régression PLS. Ainsi, la PLS présente l'avantage de ne pas nécessiter d'inverser les matrices pour la résolution du problème d'optimisation. En « mode régression », la déflation de $\mathbf{Y}$ fait bien de la PLS une approche supervisée, car elle fait intervenir la régression de $\mathbf{Y}$ sur la composante η_l construite à partir de $\mathbf{X}$.
Il est à noter que la PLS existe aussi en « mode canonique », qui diffère du mode régression, par la déflation utilisée sur $\mathbf{Y}$. En effet, en mode canonique, celle-ci devient :
$$\mathbf{Y}_l = \mathbf{Y}_{l-1} - \xi_l d_l^\top .$$
Dans ce cas, la déflation ne fait intervenir que les éléments de la régression de $\mathbf{Y}$ sur les composantes ξ calculées à partir de $\mathbf{Y}$ et $\mathbf{X}$ et $\mathbf{Y}$ jouent un rôle symétrique dans la construction des composantes.

En régression PLS, les vecteurs loadings sont directement interprétables, puisqu'ils indiquent comment les variables de chaque tableau de données contribuent à l'explication des relations entre $\mathbf{X}$ et $\mathbf{Y}$, tandis que les composantes de la PLS représentent des coefficients de similarité entre

individus [WOLD et collab., 2004]. Elle présente cependant le désavantage d'offrir des composantes difficilement interprétables, ce qui peut être corrigé par l'utilisation d'approches parcimonieuses pour faire la sélection des variables initiales les plus importantes pour expliquer les relations entre $\mathbf{X}$ et $\mathbf{Y}$ [LÊ CAO et collab., 2008].

Régression inverse par tranches

La dernière méthode que nous aborderons dans cette section est la SIR. La SIR est une méthode de régression semi-paramétrique. De façon classique, un modèle de régression paramétrique est constitué d'une fonction de lien dont les paramètres sont spécifiés et peut être appliqué pour expliquer une variable cible $\mathbf{Y}$ par une variable explicative $\mathbf{X}$. Il est alors possible d'estimer les paramètres du modèle et de décider en aval si le modèle est adapté à la description de notre jeu de données particulier. On peut citer par exemple le modèle linéaire classique dont les paramètres sont les coefficients associés à chaque variable et peuvent être estimer par maximum de vraisemblance. Cependant, lorsque la mise en évidence d'un modèle paramétrique n'est pas simple, il est possible de se tourner vers des modèles non paramétriques. L'objectif est alors de construire une fonction inconnue de régression f à partir de l'information disponible dans $\mathbf{X}$.
Un modèle semi-paramétrique fait intervenir ces deux aspects. Dans le cas de la SIR, si l'on reprend les données telles que décrites dans la figure 1.11 et avec $\mathbf{Y}$, une variable cible réelle univariée ($q = 1$), on peut formuler le modèle comme suit :

$$\mathbf{Y} = f(\mathbf{X}^\top \mathbf{a}_1, \ldots, \mathbf{X}^\top \mathbf{a}_d, \epsilon) \tag{1.3}$$

où $d \ll p$, $f : \mathbb{R}^{d+1} \to \mathbb{R}$ est une fonction arbitraire inconnue, $(\mathbf{a}_k)_{k=1,\ldots,d}$ sont des vecteurs de $\mathbb{R}^p$ à estimer et ϵ est un terme d'erreur. On notera, en outre, $\mathbf{A}$ la matrice de dimensions $p \times d$ dont les colonnes sont les $\mathbf{a}_k$.

L'hypothèse sous-jacente à ce modèle, est qu'il existe un espace de dimension réduite, $S_{Y|X}$, qui est un espace le plus petit possible et est définit tel que la projection de $\mathbf{X}$ sur $S_{Y|X}$ contient toute l'information disponible dans la variable aléatoire X pour expliquer Y. Cet espace, qu'on appelle l'*espace central*, et est aussi appelée espace EDR (pour *effective dimension reduction*) contient toute l'information disponible dans $\mathbf{X}$ sur $\mathbf{Y}$. La SIR est une méthode de régression, mais il est possible de l'utiliser à des fins exploratoires. On utilise alors la matrice $\mathbf{A}$ pour fournir une représentation des données qui correspond à la meilleure projection linéaire de $\mathbf{X}$ expliquant $\mathbf{Y}$.
LI [1991] montre que sous les conditions appropriées, $\mathbf{A}$ peut être estimée par les d premiers vecteurs propres Γ-orthonormés de Γ^e, où Γ est la ma-

trice de variance-covariance empirique de $\mathbf{X}$, $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ et Γ^e est la matrice de covariance empirique de la matrice d'espérance contionnelle $\mathbb{E}(\mathbf{Z}|\mathbf{Y})$ pour $\mathbf{Z} = (\mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}^\top)\Gamma^{-1/2}$. Pour cela, on commence par découper le support de $\mathbf{Y}$ en H tranches $(\tau_h)_{h=1,\dots,\mathrm{H}}$. $\mathbb{E}(\mathbf{Z}|\mathbf{Y})$ peut alors être estimer par la matrice $\mathbf{G}$ telle que :

$$\mathbf{G} = \frac{1}{n_h}\Delta^\top\mathbf{Z} \tag{1.4}$$

où n_h est le nombre d'observations i pour lesquelles y_i se trouve dans la tranche τ_h ($n_h = \sum_i \delta_{\{y_i \in \tau_h\}}$), et Δ est la matrice de taille $n \times \mathrm{H}$ définie telle que $\delta_{ih} = 1$ si $y_i \in \tau_h$ et $\delta_{ih} = 0$ sinon. En pratique, les tranches de $\mathbf{Y}$ sont souvent choisies de manière à avoir un effectif similaire dans chacune des tranches, ce qui peut faire varier l'amplitude des τ_h. Cela permet d'assurer un effectif suffisant dans chacune des tranches [SARACCO et collab., 1999].

Γ^e peut alors être calculée de la façon suivante :

$$\Gamma^e = \mathbf{G}^\top\mathbf{M}\mathbf{G}$$

où $\mathbf{M} = \mathrm{Diag}(\frac{n_1}{n}, \dots, \frac{n_\mathrm{H}}{n})$. Les d premiers vecteurs propres de Γ^e estiment l'espace EDR. La projection des tranches et des variables sur ces composantes permettent de visualiser les variables de $\mathbf{X}$ les plus importantes pour expliquer $\mathbf{Y}$.

1.7.5 Cas particulier des données longitudinales

La complexification des protocoles expérimentaux a induit l'amélioration des techniques de biologie moléculaire et n'a pas seulement permis la collecte et l'analyse conjointe de données à plusieurs niveaux du vivant, mais aussi la collecte de données répétées dans le temps. Mesurer des données répétées dans le temps permet d'étudier la dynamique d'une réaction à une perturbation et pas seulement un instantané de cette réaction à un temps t. Cette approche est indispensable pour la compréhension des mécanismes en jeu dans un système, du fait des rétrocontroles qui peuvent avoir lieu lors de la réponse à une perturbation (un stress par exemple) et qui peuvent entrainer une réponse allant dans un sens en réaction immédiate et dans un sens opposé à plus long terme.
Dans le cadre de cette thèse, on dispose de données 'omiques longitudinales, obtenues en réponse à différents facteurs de stress, que l'on souhaite pouvoir analyser en tenant compte d'une variable cible réelle : la mesure du cortisol (l'hormone principale des réponses au stress) lorsque l'axe corticotrope est au maximum de son activité.
Les données longitudinales peuvent être vues comme un cas particulier de

l'intégration de données. En effet, les approches intégratives sont classiquement utilisées pour étudier les relations entre différents types de données (intégration multi-omique) mais elles peuvent aussi s'appliquer pour l'intégration de données longitudinales. Dans ce cas, les données collectées à T pas de temps peuvent être vues comme T tableaux d'observations indépendantes. Les approches multi-tableaux AFM et dual-STATIS décrites en section 1.7.3 sont donc particulièrement adaptées à l'analyse de ce type de données. Elles présentent cependant l'inconvénient de ne pas conserver la continuité entre les pas de temps. En effet, le temps est alors perçu comme un facteur à plusieurs niveaux et aucune information n'est conservée sur l'ordre des pas de temps, ni sur leur proximité temporelle (on pourrait s'attendre à ce que des pas de temps proches aient des structures de corrélations plus similaires que des pas de temps éloignés). Parmi les alternatives aux approches multi-tableaux, on peut citer :

— la décomposition des sources de variation au sein des données ;

— les modèles de représentation fonctionnelle telles que les fonctions spline.

Décomposition des sources de variation au sein des données Lors de l'analyse de données 'omiques longitudinales, une approche commune est d'appliquer un modèle linéaire mixte univarié sur chacune des variables pour identifier les variables différentiellement exprimées au cours du temps, suivi d'une correction pour tests multiples afin de contrôler le taux de faux positifs [KARLOVICH et collab., 2009].

Si l'on considère à nouveau le tableau $\mathbf{X}$ tel que décrit dans la figure 1.10, on peut considérer qu'au temps t, la variable j pour l'individu i est modélisée par :

$$x_{ijt} = \mu_{jt} + \pi_{ij} + \epsilon_{ijt}$$

où μ_{jt} est l'effet fixe du pas de temps t, π_{ij} est une variable aléatoire qui suit une loi normale $\mathcal{N}(0, \sigma^2_{\pi,j})$ et prend en compte la dépendance entre les observations d'un même individu i et ϵ_{ijt} sont les résidus du modèle, distribués selon une loi normale $\mathcal{N}(0, \sigma^2_{\epsilon,j})$ et tiennent compte des effets non identifiables par le protocole expérimental.
La limite de cette approche est qu'elle applique un modèle séparément sur chaque variable, ce qui ne permet pas de prendre en compte les relations pouvant exister entre variables d'un même tableau de données.

Pour pallier cette limite, LIQUET et collab. [2012] proposent l'analyse multi-niveaux. Cette méthode repose sur une décomposition de la variation dans les données, inspirée des modèles linéaires mixtes. En reprenant

les notations du paragraphe précédent, on peut décomposer $\mathbf{X}$ en :

$$\mathbf{X} = \underbrace{\mathbf{X}_{...}}_{\text{valeur moyenne}} + \underbrace{\mathbf{X}_b}_{\text{variation inter-sujet}} + \underbrace{\mathbf{X}_w}_{\text{variation intra-sujet}}$$

où

— $\mathbf{X}_{...} = \mathbf{1}_{n\mathrm{T}}\, \bar{x}_{...}^{\top}$ représente la valeur moyenne dans $\mathbf{X}$ et où $\bar{x}_{...}^{\top} = (\bar{x}_{.1.}, \ldots, \bar{x}_{.p.})$ $(\bar{x}_{.j.} = \frac{1}{n\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \sum_{i=1}^{n} x_{ijt})$ contient les moyennes empiriques des p variables tous pas de temps confondus et $\mathbf{1}_{n\mathrm{T}}$ est un vecteur unitaire de taille $n\mathrm{T}$;

— $\mathbf{X}_b = \begin{bmatrix} 1_{\mathrm{T}} \bar{x}_{(b)1}^{\top} \\ \vdots \\ 1_{\mathrm{T}} \bar{x}_{(b)n}^{\top} \end{bmatrix}$ est la matrice de variation inter-individus, avec $\bar{x}_{(b)i}^{\top} = (\bar{x}_{i1.} - \bar{x}_{.1.}, \ldots, \bar{x}_{ip.} - \bar{x}_{.p.})$;

— $\mathbf{X}_w = \mathbf{X} - 1_{\mathrm{T}} \bar{x}_{i..}^{\top}$ est la matrice de variation intra-individus, avec $\bar{x}_{i..}^{\top} = (\bar{x}_{i1.}, \ldots, \bar{x}_{ip.})$.

$\mathbf{X}_w$ représente la variabilité intra-individus et est composée à la fois de l'effet du temps et de tous les autres effets non identifiables (erreur). Après décomposition, Liquet et collab. [2012] analysent $\mathbf{X}_w$ en utilisant des analyses multivariées classiques : ACP, régression PLS, etc. En décomposant les sources de variation dans les jeux de données originaux et en tenant compte de la nature répétée des données, l'analyse multi-niveaux permet d'étudier les effets de chaque pas de temps au sein des individus séparément de la réponse inter-individuelle.

L'utilisation de cette méthode est illustrée dans cette thèse. En effet, l'approche multi-niveaux a été utilisée en combinaison avec des ACP et des régressions PLS pour explorer les données de biologie clinique et de transcriptome des expériences d'ACTH et de LPS du projet SUSoSTRESS. Ces applications ont été publiées dans les deux articles qui constituent le chapitre 2.

Utilisant le même principe que l'approche multi-niveaux, l'ASCA (*ANOVA-Simultaneous Component Analysis*) [Jansen et collab., 2005; Smilde et collab., 2005] (initialement développée pour les données métabolomiques et initialement connue sous le nom de ACP sur variables instrumentales [Rao, 1964; Sabatier et collab., 1989]) permet d'étudier les données 'omiques répétées. Cette méthode, qui généralise les ANOVA aux données multivariées, propose de décomposer un jeu de données en plusieurs sous-matrices. Chaque sous-matrice contient les effets dûs à un facteur du protocol expérimental étudié (par exemple, des effets régimes,

traitements, temps,...). Une ACP est ensuite appliquée sur la matrice d'effets d'intérêt.
Décomposer les sources de variation dans les jeux de données et travailler sur les matrices représentant l'effet du temps permet d'explorer le lien entre les variables et le temps, peut être vu comme une façon de pré-traiter les données avant d'appliquer des analyses exploratoires classiques.

Fonctions spline Les stratégies d'analyses multi-tableaux et les approches par décomposition des sources de la variation au sein des données présentent toutes deux le même inconvénient : elles ne permettent pas de conserver la continuité entre deux pas de temps. Une alternative à ces approches est l'utilisation de représentations fonctionnelles telles que les fonctions spline. Lorsque l'on cherche à modéliser l'évolution d'un phénomène complexe, il est fréquent que la fonction décrivant les données soit non linéaire. Le modèle utilisé devient complexe et demande l'ajout de nouveaux paramètres faisant intervenir des polynômes de degré parfois élevé pour en améliorer l'ajustement.
Dans [DÉJEAN et collab., 2007] les auteurs proposent d'utiliser des fonctions spline pour l'analyse de données longitudinales à l'évolution non linéaire dans le temps. Les fonctions spline sont des fonctions polynômiales par morceau. Le principe est de couper les données en plusieurs intervalles et d'estimer un pôlynome sur chaque sous-intervalle. Ce découpage permet d'estimer plusieurs polynômes de degré faible au lieu d'un seul polynôme de degré élevé, rendant ainsi l'estimation du modèle plus simple. Le choix d'un paramètre de lissage permet de déterminer le nombre de segments pour le découpage [DÉJEAN et collab., 2007]. L'objectif est alors de modéliser les changements de niveaux d'expression des variables 'omiques par une fonction lissée d'évolution au cours du temps.
Les fonctions spline ont été largement explorées pour pallier leurs deux plus grosses limites dans le contexte des données à grande dimension : le choix du paramètre de lissage, qui est arbitraire, et leur coût computationnel élevé. Par exemple, PATTERSON et THOMPSON [1971] ont travaillé sur une redéfinition des fonctions spline dans un contexte de modèle linéaire mixte (approche LMMS) pour pallier au choix arbitraire du paramètre de lissage et plus récemment, DURBÁN et collab. [2005] ont proposé une approche à base de fonctions spline tronquées pour réduire les coûts computationnels. Enfin, STRAUBE et collab. [2015] proposent une approche combinant étape de filtrage et modélisation souple de la fonction spline tenant compte de sa variabilité inter et intra-individuelle avec un modèle mixte. Après avoir modélisé les fonctions spline, il devient,

par exemple, possible de faire de la classification sur les coefficients de la décomposition pour identifier des clusters de variables ayant des profils d'évolution similaires.

Comme mentionné précédemment, les méthodes d'intégration multi-tableaux telles que dual-STATIS et AFM présentent l'inconvénient de ne pas conserver la continuité entre les pas de temps. L'approche multi-niveaux présente le même inconvénient. Cependant, il s'agit de méthodes simples à mettre en place et qui sont plus adaptées que les fonctions spline dans les cas où les données sont répétées sur un nombre faible de pas de temps. C'est pourquoi, nous nous sommes tournés vers ces approches pour mener à bien nos objectifs d'intégration de données longitudinales. Ainsi, l'approche multi-niveaux a été utilisée dans le chapitre 2 pour l'analyse conjointe des variables cliniques et transcriptomiques des données du projet SUSoSTRESS. Pour aller plus loin, nous souhaitions aussi pouvoir faire l'intégration de données longitudinales en tenant compte d'une variable réelle cible : la valeur du cortisol (l'hormone principale du stress), lorsque l'axe corticotrope est au maximum de son activité. Les approches à notre disposition n'étant pas satisfaisantes à nos yeux, il nous est alors alors paru nécessaire d'intégrer aux objectifs de cette thèse le développement d'une approche statistique répondant à nos attentes. Nous avons donc décidé d'adapter une analyse multi-tableaux, dual-STATIS, au cadre d'une approche intégrative supervisée faisant intervenir une variable cible : la SIR. Les développements méthodologiques issus de ce travail sont décrit en chapitre 3.

1.8 Objectifs et plan de la thèse

Au cours de ce chapitre introductif, nous avons présenté le contexte général autour de ce travail de thèse. Il y est rappelé que le « stress » est défini comme la réponse non spécifique des organismes à toute stimulation. Chez les espèces animales vertébrées destinées à l'alimentation humaine, l'axe corticotrope est le plus important système neuroendocrinien de réponse au stress. De grandes variations individuelles d'origine génétique ont été décrites dans l'activité de l'axe corticotrope avec des conséquences physiopathologiques importantes. En terme de production animale, des niveaux plus élevés de cortisol ont des effets négatifs sur la croissance et l'efficacité alimentaire et augmentent le ratio gras/maigre des carcasses. Au contraire, le cortisol a des effets positifs sur les caractères liés à la robustesse et à l'adaptation. La sélection intense pour la croissance des tissus maigres durant les dernières décennies a concomitamment réduit la production de cortisol, et nous faisons l'hypothèse que cette réduction

peut être partiellement responsable des effets négatifs de la sélection sur les caractères de robustesse.
Le travail de cette thèse s'inscrit dans le cadre du projet ANR SUSoSTRESS qui a pour objectif la compréhension des mécanismes moléculaires et génétiques sous-jacents à la variabilité individuelle de réponses de stress et a collecté sur 120 porcs des données longitudinales cliniques, transcriptomiques et métabolomiques.

La présente thèse a 2 objectifs principaux :

1. Le premier objectif est le développement d'un modèle fonctionnel permettant de décrire et d'intégrer au mieux l'ensemble des sources de variation génétique du fonctionnement de l'axe corticotrope et plus généralement des réponses de stress dans notre population porcine d'étude (race Large White). Plus précisément, il s'agit d'élaborer un modèle (au sens biologique du terme) décrivant les différentes réponses biologiques de stress et l'influence des variations génétiques (simples et en interaction), dans le but de prédire les leviers les plus efficaces en fonction de l'objectif de sélection. Cet objectif principal peut être découpé en plusieurs objectifs secondaires :
   - intégrer des données de hautes dimensions issus de différents tissus (biologie clinique, transcriptome) ;
   - tenir compte de l'aspect longitudinal des données ;
   - extraire un sous-ensemble de gènes différentiellement exprimés lors des réponses au stress ;
   - mettre les données en relation avec un phénomène cible d'intérêt principal : la mesure du cortisol en réponse à une injection d'ACTH qui représente le niveau d'activité de l'axe corticotrope.
2. Les données 'omiques à intégrer étant de nature longitudinale, il est nécessaire de développer une méthodologie statistique adaptée à notre cadre et aux questions posées.

Cette thèse s'articule autour de ces objectifs : Le chapitre 2 est consacré à l'étude des données cliniques et transcriptomiques du projet SUSoSTRESS. Il est constitué de deux articles. Ces deux articles présentent des résultats obtenus avec des méthodes statistiques existantes et ne comportent pas de développement méthodologique novateur spécifique. Le premier décrit les résultats obtenus en intégrant ensemble les données de biologie clinique et transcriptomique de l'expérience d'injection d'ACTH, dont l'identification de gènes clefs impliqués dans les cascades de signalisation et lien avec les réponses de stress. Il a fait l'objet d'une publication dans une revue à comité de lecture. Le deuxième article décrit ceux

obtenus pour l'intégration des données de biologie clinique et transcriptomique de l'expérience d'injection de LPS et cible plus spécifiquement les relations entre axe corticotrope et régulation du système immunitaire. Cet article est en cours d'écriture et sera soumis prochainement.

Le chapitre 3 est consacré à la contribution méthodologique et décrit la méthode de la « multiway-SIR », développée au cours de cette thèse pour l'intégration de données multidimensionnelles longitudinales. La méthode est illustrée par une application sur les données de biologie clinique de l'expérience d'ACTH. Ce chapitre fait l'objet d'un article en cours d'écriture.

Enfin, le chapitre 4 conclut ce manuscrit par une discussion générale reprenant l'ensemble de ces travaux et les perspectives pour la recherche et la sélection des animaux de production.

Chapitre 2

Etude de la cinétique de réponse à l'ACTH et au LPS, biologie clinique et transcriptome sanguin

Sommaire

2.1 Introduction

Ce travail avait pour objectif d'analyser de façon approfondie (comportement, biologie clinique, métabolome, transcriptome) la cinétique des réponses au stress aigu et d'en analyser la variabilité individuelle. Il travail a été réalisé dans le cadre du projet SUSoSTRESS (Génétique moléculaire des réponses de stress et robustesse chez le Porc), financé par l'Agence Nationale pour la Recherche (ANR-12-ADAP-0008). L'un de ses objectifs est l'élaboration d'un modèle, chez le porc, des variations génétiques de l'axe corticotrope et de son activité physiologique en lien avec les performances des animaux sur les caractères de robustesse et de production. Le Large White a été choisi car il s'agit de l'une des races largement utilisée en croisement pour la production de viande porcine en France, sélectionnée sur les caractères de production et dans laquelle on a montré une grande variabilité individuelle de l'activité de l'axe corticotrope [LARZUL et collab., 2015].

Le projet SUSoSTRESS comporte deux composantes successives : une analyse à haut débit de la variabilité des réponses de stress dans une population fondarice (G0), puis une sélection génétique divergente pendant 3 générations sur la base de la réponse de la glande surrénale à l'ACTH, source majeure de variabilité génétique de l'axe corticotrope. Ma contribution s'est limitée à l'étude de la population fondatrice.

Protocole expérimental du projet SUSoSTRESS, étude de la population fondatrice G0

La population d'étude était composée de 120 porcs Large White (57 mâles et 63 femelles). Ils étaient issus de 28 familles (avec 4 à 5 porcs par famille) et, pour des raisons de logistique, ont été élevés en 3 bandes d'élevage avec 3 semaines d'intervalle entre chaque bande. Les porcs ont été sevrés à l'âge de 4 semaines et les animaux de 2 à 3 familles ont été mélangés après sevrage.

Afin d'étudier la variabilité du fonctionnement de l'axe corticotrope, celui-ci a été étudié au cours de trois expériences afin d'analyser les différentes composantes des réponses de stress :

— une injection d'ACTH, l'hormone sécrétée par l'hypophyse qui cible les surrénales et provoque la libération de cortisol (l'hormone principale du stress chez le porc) dans la circulation sanguine. Cette expérience stimule de façon directe l'axe corticotrope et permet de me-

surer la sensibilité à l'ACTH, source majeure de variabilité génétique, ainsi que les réponses tissulaires au cortisol. C'est également le test qui a été utilisé ultérieurement pour l'expérience de sélection ;

— un test de contrainte de 10 minutes. Au cours de ce test, les 120 animaux ont été placés individuellement dans un filet à la verticale (tête en haut et pattes arrière en bas) et soulevés de terre de manière à ce que les pattes arrières ne touchent plus le sol et que leurs mouvements soient restreints tout au long de l'expérience (voir figure 2.1). Cette expérience permet d'obtenir une réponse comportant à la fois une composante comportementale et biologique au stress ;

— une injection de lipopolysaccharide (LPS). Le LPS est un composant des parois bactériennes à gram négatif. Son injection déclenche une réaction inflammatoire qui simule un stress de type maladie.

FIGURE 2.1 – Photo du dispositif permettant de restreindre les mouvements des porcs dans l'expérience de contrainte du projet SUSoSTRESS.

Ces trois expériences apportent chacune de l'information sur des volets différents des mécanismes de réponses au stress (sensibilité à l'ACTH et au cortisol, réponse comportementale, système immunitaire) et sont donc complémentaires entre elles. Elles ont toutes été réalisées sur les mêmes animaux à 1 semaine d'intervalle : à l'âge de 6 semaines pour l'expérience d'ACTH, de 7 semaines pour l'expérience de contrainte et de 8 semaines pour l'expérience de LPS.
Dans chaque expérience, les données ont été collectées à partir de prises de sang effectuées à plusieurs pas de temps : avant l'expérience (t=0), puis

à t=+1h, t=+4h et le lendemain à t=+24h pour les 3 expériences. Un pas de temps supplémentaire a été mesuré pour l'expérience de contrainte, à t=+10min (c'est à dire à la fin du test de contrainte), afin d'obtenir des mesures les plus proches possibles de la source de stress.

À chaque pas de temps et pour chaque expérience, les données ont été collectées à 3 niveaux :

— au niveau clinique : on a mesuré sur la totalité des 120 animaux la concentration de cortisol dans le sang, ainsi que de métabolites plasmatiques (glucose, AGL,...) et la numération et formule sanguines (globules blancs, proportions des sous-populations de globules blancs, globules rouges, plaquettes, ...) ;

— au niveau métabolomique : à cette échelle, on mesure les changements dans les réactions biochimiques provoquées par chacun des stress. La mesure a été effectuée sur la totalité des 120 animaux ;

— au niveau transcriptomique : à cette échelle, on mesure les changements au niveau de l'expression des gènes. Le transcriptome a été mesuré sur le sang total, à l'aide de puces à ADN Agilent 60K, sur un sous-ensemble de 30 animaux. Le choix du sang comme tissu d'analyse se justifie par la facilité d'obtention, d'où la possibilité de l'étude cinétique chez les mêmes animaux comme dans la présente expérimentation et ultérieurement des études cliniques et populationnelles. C'est également l'approche privilégiée chez l'humain pour les études transcriptomiques du stress ou du système immunitaire par exemple [Chaussabel et collab., 2010; Cole, 2010].

Ce schéma expérimental présente plusieurs avantages. Premièrement, trois stimulus provoquant un large éventail de réponses complémentaires pour la compréhension des mécanismes de réponses au stress ont été appliqués sur une même population. Il est donc possible de comparer les données collectées entre les trois expériences pour analyser la variabilité individuelle. Deuxièmement, les données sont collectées à trois niveaux d'analyse biologique pour les trois expériences, ce qui permet une approche intégrative au sein de chaque expérience. Enfin, pour chaque expérience et chaque type de données, on dispose de données répétées dans le temps, ce qui permet d'obtenir une compréhension de la dynamique de la réponse au stress dans le temps et non un instantané dont les résultats peuvent varier en fonction du moment t où l'on observe la réponse, d'autant plus que les différentes réponses ont une cinétique propre en fonction du stimulus et de la nature de la réponse.

Le tableau 2.1 résume l'ensemble des données disponibles dans le projet SUSoSTRESS et celles qui ont été analysées au cours de cette thèse. Les

données analysées et qui font l'objet de chapitres dans cette thèse et de publications acceptées ou en cours de soumission sont principalement les données de biologie clinique et transcriptomique des expériences d'ACTH et de LPS.

TABLEAU 2.1 – Tableau récapitulatif des données du projet SUSoSTRESS et de leur valorisation dans ce travail. En orange, les données qui ont été analysées au cours de ces 3 ans de thèse. En bleu, les données dont l'analyse est en cours ou reste à faire.

	Test à l'ACTH (axe corticotrope)	Stress de contrainte	Injection de LPS (stress inflammatoire)
Biologie clinique	Chapitre 2 & 3, Article accepté, pré-publication		Chapitre 2, Article en pré-publication
Transcriptome sanguin	Chapitre 2, Article accepté		Chapitre 2, Article en pré-publication
Métabolome			
Comportement			

Dans ce chapitre sont décrits les résultats des analyses de biologie clinique et de transcriptome réalisées après administration d'ACTH et de LPS.

2.2 Article 1 - Time course of the response to ACTH in pig : biological and transcriptomic study

Sautron *et al. BMC Genomics* (2015) 16:961
DOI 10.1186/s12864-015-2118-8

BMC Genomics

RESEARCH ARTICLE Open Access

Time course of the response to ACTH in pig: biological and transcriptomic study

Valérie Sautron[1,2,3*], Elena Terenina[1,2,3], Laure Gress[1,2,3], Yannick Lippi[4], Yvon Billon[5], Catherine Larzul[1,2,3], Laurence Liaubet[1,2,3], Nathalie Villa-Vialaneix[6] and Pierre Mormède[1,2,3]

Abstract

Background: HPA axis plays a major role in physiological homeostasis. It is also involved in stress and adaptive response to the environment. In farm animals in general and specifically in pigs, breeding strategies have highly favored production traits such as lean growth rate, feed efficiency and prolificacy at the cost of robustness. On the hypothesis that the HPA axis could contribute to the trade-off between robustness and production traits, we have designed this experiment to explore individual variation in the biological response to the main stress hormone, cortisol, in pigs. We used ACTH injections to trigger production of cortisol in 120 juvenile Large White (LW) pigs from 28 litters and the kinetics of the response was measured with biological variables and whole blood gene expression at 4 time points. A multilevel statistical analysis was used to take into account the longitudinal aspect of the data.

Results: Cortisol level reached its peak 1 h after ACTH injection. White blood cell composition was modified with a decrease of lymphocytes and monocytes and an increase of granulocytes ($FDR < 0.05$). Basal level of cortisol was correlated with birth and weaning weights. Microarray analysis identified 65 unique genes of which expression responded to the injection of ACTH (adjusted $P < 0.05$). These genes were classified into 4 clusters with distinctive kinetics in response to ACTH injection. The first cluster identified genes strongly correlated to cortisol and previously reported as being regulated by glucocorticoids. In particular, *DDIT4*, *DUSP1*, *FKBP5*, *IL7R*, *NFKBIA*, *PER1*, *RGS2* and *RHOB* were shown to be connected to each other by the glucocorticoid receptor NR3C1. Most of the differentially expressed genes that encode transcription factors have not been described yet as being important in transcription networks involved in stress response. Their co-expression may mean co-regulation and they could thus provide new patterns of biomarkers of the individual sensitivity to cortisol.

Conclusions: We identified 65 genes as biological markers of HPA axis activation at the gene expression level. These genes might be candidates for a better understanding of the molecular mechanisms of the stress response.

Keywords: Stress, Hypothalamic-pituitary-adrenal (HPA) axis, Cortisol, Time-course, Systems biology, Microarray, Pig

Background

Farm animals have been highly selected for favorable production traits such as lean growth rate, feed efficiency, and prolificacy in pigs. In parallel their robustness has tended to decrease, as shown by the sensitivity to diseases, locomotor weakness or behavioral problems [1]. The hypothalamic-pituitary-adrenocortical (HPA) axis plays a major role in physiological homeostasis including metabolism, brain function and behavior, the immune system and inflammatory processes. Together with the autonomic nervous system, it is also involved in stress and adaptive responses to environmental challenges. On the basis of available literature, we have hypothesized that the HPA axis could contribute to the trade-off between production and robustness traits, and that genetic variation in HPA axis activity could be used to select more robust animals [2, 3]. HPA axis activity shows a large variation among individuals and genetic influences are well documented [4]. For example, in pigs, the sensitivity of the adrenal cortex to adrenocorticotropic hormone (ACTH)

*Correspondence: valerie.sautron@toulouse.inra.fr
[1]INRA, UMR 1388 Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France
[2]Université de Toulouse INPT ENSAT, UMR 1388 Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France
Full list of author information is available at the end of the article

and the production of corticosteroid binding globulin (CBG), the carrier of cortisol in blood are the two main mechanisms responsible for genetic differences in circulating cortisol levels [5, 6]. In a previous paper, Hazard et al. [7] have studied at the gene expression level the molecular mechanisms of genetic differences in the adrenal gland response to ACTH. However little is known about the individual variation in the biological activity of cortisol, the main glucocorticoid hormone, and the genetic mechanisms involved.

Corticosteroid hormones exert their biological actions via two intracellular receptors (gluco- and mineralo-receptor) that, upon activation by their ligand, influence the expression of a large number (several hundreds) of genes in a wide range of cell types [8]. In pigs, glucocorticoid receptor polymorphisms have been described with their functional consequences [9–13]. The present experiment was designed to explore in pigs individual variation in the biological response to cortisol in order to identify possible biomarkers of this sensitivity. To that end, juvenile pigs were submitted to an ACTH challenge to release cortisol and the kinetics of the response was measured with biological variables and with gene expression analysis in blood cells. Taken together with biological information, this approach will serve as an important step to understand HPA axis regulation and will identify key genes involved in signaling pathways relevant to stress responses. The final goal of this work is to develop a strategy for further genetic studies in order to overcome the unfavourable consequences of stress in farm animals.

Animals and methods

Animals, treatment and blood sampling

All animal use was performed under European Union and French legislation (directive 201063UE, décret 2013-118). The protocol and procedures were approved by the local (Poitou-Charentes) ethics committee (decision CE2013-1, 21012013). Experimental animals were 120 piglets (63 females and 57 males) randomly selected from 28 litters (4–5 animals per litter) of purebred Large White pigs and produced in 3 successive batches raised 3 weeks apart. They were weaned at 4 weeks and animals from 2–3 litters were mixed at weaning in each post-weaning pen. Experimental animals were not isolated from their littermates. At 6 weeks, each animal was injected in the neck muscles with 222 μg of synthetic 1–24 ACTH (Pepscan Presto BV, Lelystad, The Netherlands) diluted in 1 mL of 0.9 % saline. Injections occurred from 10:00–11:00 AM to avoid nycthemeral variations. Blood samples were collected before the injection ($t = 0$) and 1 ($t = +1$), 4 ($t = +4$) and 24 ($t = +24$) hours later. At each time, animals were slightly restrained on their back in such a way that the effect on their stress level can be regarded as insignificant. Two blood samples were then taken by puncture of a jugular vein in Vacutainer® tubes with 20 G needles. The whole handling procedure lasted less than 30 sec. One 10 mL tube with lithium heparin was used for chemical biology. After centrifugation (2355 g, 10 min), plasma aliquots were frozen at −80 °C until analysis. One 5 mL tube with EDTA (di-potassium salt) was used for blood cell count and an aliquot (400 μL) was mixed with the same volume of DL buffer (Macherey-Nagel), frozen at −20 °C for 4 h and then at −80 °C until analysis for gene expression.

Biological analyses

Cortisol was measured by direct automated immunoassay (AIA-1800, Tosoh Bioscience, San Francisco, CA). Glucose and free fatty acid (FFA), were measured by colorimetry with an ABX Pentra 400 clinical chemistry analyzer from Horiba Medical (Grabels, FR). Blood cell counts were measured with a MS-9-5 hematology analyzer from Melet Schloesing Laboratories (Osny, FR), calibrated for pig blood by the manufacturer. Blood cell count variables included: white cells count, proportion of lymphocytes, monocytes and granulocytes, red cells count, percentage of hematocrit, concentration of hemoglobin, red cells width and volume, concentration of platelets and platelets width and volume. The biological variables contained thus 15 variables measured on the 120 pigs. In addition, birth and weaning weights were also measured for each pig. Outlying observations were visually identified and treated as missing data. Missing data were imputed using a k-nearest neighbour ($k = 5$) imputation (R package **DMwR** [14]). To ensure normality, cortisol, platelets and white cells count were $\log_{10}$ transformed and FFA was transformed using the square-root. Batch effects were removed by aligning the within-batch medians for all measurements.

RNA extraction and whole blood gene expression analysis

A total RNA isolation and purification was done according to the manufacturer's instructions using the Nucleospin RNA Blood kit (Macherey-Nagel, France) followed by DNase treatment. The quality of each RNA sample was checked through the Bioanalyser Agilent 2100 (Agilent Technologies, Massy, France) and low-quality RNA preparations were discarded (RIN < 8).

Microarray description

A porcine microarray GPL16524 (Agilent, 8×60 K) was used to hybridize the RNA samples as previously described [15]. This microarray contained 61,625 spots. Among them, 308 were negative controls and 49 were used for aligning. One probe was duplicated twice on each array. Thus the microarray contained 60,305 unique porcine probes. After quality control, quantile normalization and filtering, 35,429 transcripts were found to be expressed in blood in our experimental conditions.

Hybridization protocol

Blood samples from 30 female pigs from only 2 batches were used. Each of the 15 arrays used contained 8 microarrays which corresponded to the 4 observations of two individuals, each from one batch. This design secured the kinetics of the response for each individual and prevented confounding effects between batch and array. After quality control and filtering, 35,419 probes were kept and $\log_2$ transformed. Technical biases were handled by aligning the within-array medians for all genes and by a quantile normalization within animal (function **normalize.quantiles** in the R package **preprocessCore** [16]). No missing data were reported. Normalized data were submitted to GEO/NCBI with the GSE71207 accession number.

Statistical analyses

All analyses were performed with the R software, version 3.1.0 [17].

Multilevel approach

A multilevel approach was used to investigate the relationships between the repeated measurements while taking advantage of multivariate approaches. The multilevel approach, as described by Liquet et al. [18], uses a split-up variation coming from the mixed-model framework. Let $X = (x_{it}^k)_{i=1,\ldots,n,\, t\in\{0,+1,+4,+24\},\, k=1,\ldots,p}$ be the $(N \times p)$ observation matrix (biological variables or gene expressions) on n animals with 4 times of measurements ($N = n \times 4$). X can be split up as:

$$X = \underbrace{X_{\cdot\cdot}}_{\text{offset term}} + \underbrace{X_b}_{\text{between-animal deviation}} + \underbrace{X_w}_{\text{within-animal deviation}} \quad (1)$$

The matrix $X_{\cdot\cdot}$ represents the offset term defined as $1_N\, x_{\cdot\cdot}^T$ where 1_N is a vector of length N containing only ones and $x_{\cdot\cdot}^T = (x_{\cdot\cdot}^1, \ldots, x_{\cdot\cdot}^p)$ (with $x_{\cdot\cdot}^k = \frac{1}{N} \sum_{t\in\{0,+1,+4,+24\}} \sum_{i=1}^{n} x_{it}^k$). X_b is the between-animal matrix of size $(N \times p)$ defined by concatenating $1_4\, x_{bi}^T$ for each animal into X_b with $x_{bi}^T = \left(x_{i\cdot}^1 - x_{\cdot\cdot}^1, \ldots, x_{i\cdot}^p - x_{\cdot\cdot}^p\right)$ $\left(x_{i\cdot}^k = \frac{1}{4} \sum_{t\in\{0,+1,+4,+24\}} x_{it}^k\right)$. $X_w = X - X_{i\cdot}$ is the within-animal deviation matrix of size $(N \times p)$ with $X_{i\cdot}$ the matrix defined by concatenating the matrices $1_4\, x_{i\cdot}^T$ for every animal into $X_{i\cdot}$, with $x_{i\cdot}^T = \left(x_{i\cdot}^1, \ldots, x_{i\cdot}^p\right)$.

By splitting the different parts of the variation in the data while taking into account the repeated measurements on the subjects, the multilevel step allows us to study the effect of different conditions within a subject separately from the variation between subjects. This method is especially relevant when a high between-subject variability is observed in repeated data: multivariate approaches (such as principal components analysis, PCA [19] and partial least square regressions, PLS [20]) can then be performed on X_w to highlight the most relevant correlations between variables in the dataset, independently from individual variations.

Statistical analysis of plasma metabolites and hormone

First, all variables were subjected to a one-way ANOVA with repeated measures. *P*-values were adjusted using a Benjamini-Hochberg (BH) approach in order to control the false discovery rate (*FDR*) [21]. Variables with an adjusted *P*-value ($FDR < 0.05$) were then subjected to 3 paired *t*-tests to assess the difference between $t = 0$ and $t = +1$, between $t = 0$ and $t = +4$ and between $t = 0$ and $t = +24$. The full list of *P*-values was adjusted using a BH approach.

In addition, the influence of sex on the biological variables was tested using a two-way ANOVA with repeated measures including sex as a variable. *P*-values were adjusted using a BH approach.

Then, variability between individuals before the ACTH injection was studied by performing a PCA on the observations at $t = 0$ for variables responding to ACTH and birth and weaning weights. The overall effect of ACTH over time was investigated with a multilevel PCA as previously described.

Cortisol levels at $t = +1$ is the most relevant measure to assess the sensitivity of the adrenals to ACTH. Hence, correlations between biological variables at $t \in \{0, +1, +4, +24\}$ and the level of cortisol at $t = +1$ were investigated using paired *t*-tests. *P*-values were adjusted using a BH approach.

Statistical analysis of the transcriptome

All transcripts were subjected to 3 paired *t*-tests to assess the difference in expression between $t = 0$ and $t = +1$, between $t = 0$ and $t = +4$ and between $t = 0$ and $t = +24$. The full list of *P*-values was adjusted using a Bonferroni approach. As the Bonferroni approach exerts a more stringent control than the BH approach, it was used to obtain a narrowed list of the most significantly differentially expressed (DE) transcripts. Transcripts with at least one adjusted *P*-value < 0.05 among the three tests were kept. Correlations between the expression levels of different transcripts of the same gene were investigated to highlight genes for which at least 3/4 of the duplicates were significantly DE and had a correlation of at least 0.65 between them. The most significant transcripts per annotated gene were kept and the multilevel approach was used to extract the within-subject deviation matrix for further analysis.

A hierarchical ascending classification (HAC) was performed using the Ward method with an Euclidean distance matrix based on the correlations between genes. This allowed for the characterization of clusters of genes

having the same evolution over time. Significance of the difference between time measurements was assessed using 3 paired t-tests at the average gene level between $t = 0$ and $t = +1$, between $t = 0$ and $t = +4$ and between $t = 0$ and $t = +24$. P-values were adjusted within the clusters using a BH approach.

Integration

Relations between the main stress variable, the cortisol, the biological variables and the gene expressions were studied using different approaches.

As for biological data, correlations between gene expression at $t \in \{0, +1, +4, +24\}$ and the level of cortisol at $t = +1$ were investigated. More precisely, the within-subject components of the transcriptomic data and of the cortisol level was extracted using the method described in [18] and Section "Multilevel approach". Then, a paired t-test was performed between the within-subject expression at $t \in \{0, +1, +4, +24\}$ and the within-subject level of cortisol at $t = +1$. The full list of P-values was globally adjusted using a BH approach. In order to assess the significativity of change in expression after removing the contribution of cell population changes, a linear mixed model was fitted for every DEG

$$x_{it} = \beta_0 + \beta_{1,t} + \beta_2(\mathrm{L/G})_{it} + U_i + \epsilon_{it}$$

where x_{it} is the expression of the DEG for the animal number i ($i = 1, \ldots, 120$) and the time step t ($t \in \{0, +1, +4, +24\}$), $(\mathrm{L/G})_{it}$ is the lymphocytes/granulocytes ratio for the same experiment and U_i is the individual random effect. Both time step (as a factor) and $(\mathrm{L/G})_{it}$ were supposed to have fixed effects on gene expression. Significance of the time effect in this model was checked by testing $\beta_{1,t} = \beta_{1,0}$ for $t \in \{+1, +4, +24\}$ and correcting p-values by time point for multiple tests with a BH approach ($FDR < 0.05$). The effect of changes in white cell populations was also assessed by testing $\beta_2 = 0$ and a correction for multiple tests was applied using a BH approach ($FDR < 0.05$).

A multilevel PLS, *i.e.*, a PLS performed on the within-subject components of the biological and transcriptomic data, as described in [18] and Section "Multilevel approach", was used to investigate the overall relationships between biological and transcriptomic data. A sparse version of the PLS (L^1 penalty) as described in [20] was used to select a small subset of variables to explain every axis.

Sequence annotation

Each cDNA sequence was compared to Refseq_rna mammalian database using the NCBI blastn program (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Resulting hits were sorted out according to their closeness to the pig genome, their coverage and sequence identity. The selected cDNA sequences were submitted to the HUGO (Human Genome Organization) gene nomenclature committee, using their RefSeq IDs (http://genenames.org). Then, HUGO gene symbols or official gene symbol were used as gene names.

Functional enrichment and pathway analysis

Functional enrichment was performed on each list of clustered genes identified by HAC and on the list of genes of which expression is significatively explained by cell population changes in blood (according to the mixed model described in Section "Integration"). Functional annotation of genes was provided by the BioMart database [22]. To set the statistical enrichment of a particular biological function, a Fisher's exact test was performed, using the list of genes expressed on the microarray as the reference list of genes. Resulting P-values were adjusted for multiple tests using a BH approach. A minimum of 3 genes per gene ontology and a $FDR < 0.05$ were necessary to consider a biological function to be enriched.

A pathway is an interconnected arrangement of processes, representing the functional roles of genes in the genome. Functional integration of gene expression, *i.e.*, identification of gene networks, was performed using the 'Gene Ontology' database AmiGO (http://amigo.geneontology.org). The significantly up- or down-regulated genes could be assembled into networks using 'Ingenuity Pathway Analysis' (http://www.ingenuity.com) under licence. This application provides computational algorithms to identify the enriched biological pathways, functions and biological mechanisms of selected genes and proposes also enriched regulators as transcription factors.

A regulatory network could be constructed with the information provided by the option 'Upstream Regulator'. This option proposes a list of regulators known to have a significant effect on some of the targeted genes in the input list. Ingenuity also provides computational algorithms to identify and to dynamically generate significant biological networks. Networks are ranked by a score that takes into account the number of focus genes and the size of the networks. This score ($-\log_{10}(P\text{-value})$) indicates the probability for genes to be associated in the same network by chance. A higher score means a smaller probability for genes to be observed in the same network by chance. We chose networks displaying direct relationships between genes. Path Designer (an Ingenuity tool) was used to improve the readability of the networks. Nodes added by Ingenuity were discarded when they were not necessary to connect our genes of interest and the resulting network was merged with the regulatory network.

Results and discussion

Plasma cortisol, metabolites and hematology

Table 1 shows baseline values of the biological variables, birth weight and weaning weight. Mean evolution over time of the biological variables is shown in Fig. 1. A global effect of time was observed for 14 out of the 15 biological variables ($FDR < 0.05$). However, when considered at each time point individually, only 9 variables presented a level significantly different from their basal level for at least one time point ($t = +1$, $+4$ or $+24$).

As expected [23], ACTH induced a strong cortisol response peaking 1 h after injection ($FDR = 1.39e^{-10}$). Plasma cortisol levels at $t = +1$ were 2.7-fold higher than and highly correlated with basal levels ($r^2 = 0.63$, $FDR = 1.95e^{-14}$). They were significantly lower than baseline values at $t = +4$ ($FDR < 2.2e^{-16}$) and at $t = +24$ ($FDR = 1.39e^{-10}$), as can be expected from the feedback of cortisol on its own secretion. Plasma glucose levels showed a slight increase at $t = +4$ after ACTH injection, but FFA levels showed a large increase at $t = +1$ ($\times 3.21, FDR < 2.2e^{-16}$), with a strong variation across animals, and were almost back to basal levels at $t = +4$. The values measured at $t = +1$ were correlated with basal values ($r^2 = 0.42, FDR = 2.30e^{-06}$). Cortisol induces a weak increase in circulating glucose and also potentiates the effect of other counter-regulatory hormones [24, 25] and increases FFA levels via acute lipolysis [26].

The data obtained here for clinical hematology measures are in the range of published values in pigs. These values show large variations with age and breed among other sources [27–29]. Although the total number of leucocytes was only marginally influenced by ACTH, large changes in leucocyte subpopulations were observed with an increase of the proportion of granulocytes and a decrease of the proportion of lymphocytes and monocytes at $t = +1$ and $t = +4$. These effects are consistent with the literature [30, 31] and result from the redistribution of leucocytes between blood and tissues [32]. Red blood cell related variables (red cells count, hematocrit and hemoglobin levels) were decreased after injection of ACTH and remained low at $t = +24$. Platelets were not influenced by injection of ACTH.

Sex did not influence any of the variables ($FDR > 0.05$).

Between-subject variability at $t = 0$

Results of the PCA at $t = 0$ on variables responding to ACTH and birth and weaning weights are shown in Fig. 2. As red cells count (RC), hematocrit (Hct) and hemoglobin (Hgb) are redundant variables, decision was made to keep only RC for the PCA. No outliers are identified on the projection of the individuals on the two first dimensions of the PCA. On this PCA, sex was not found to be related to the main variability (*i.e.*, to the first axes of the PCA) in the dataset.

The first dimension shows an opposition between the proportion of granulocytes (positively correlated with this axis) and the proportion of lymphocytes (negatively correlated with this axis). The second axis shows an opposition between cortisol (positively correlated with this axis) and

Table 1 Reference values (at $t = 0$) for the biological variables, birth weight and weaning weight ($n = 120$)

	Units	Min	Max	Mean	Sem	*F* value	FDR
Cortisol	log10(ng/mL)	1.04	1.94	1.55	0.04	445.27	< 0.01
Free fatty acids	$(mmol/L)^2$	0.03	0.39	0.16	0.01	327.69	< 0.01
Glucose	mmol/L	5.62	9.84	7.67	0.17	24.36	< 0.01
White cells	log10(G/L)	0.96	1.50	1.23	0.02	2.90	0.04
Lymphocytes	%	38.95	69.53	56.63	1.12	50.59	< 0.01
Monocytes	%	5.70	11.15	7.82	0.19	15.44	< 0.01
Granulocytes	%	18.40	52.70	34.20	1.10	51.96	< 0.01
Red cells	T/L	3.88	6.60	5.48	0.08	139.26	< 0.01
Mean corpuscular volume	fL	43.10	65.22	53.67	0.77	29.66	< 0.01
Hematocrit	%	20.12	36.96	28.74	0.48	165.87	< 0.01
Hemoglobin	g/dL	7.70	12.10	9.75	0.14	155.15	< 0.01
Red blood cell distribution width	fL	30.25	35.05	32.56	0.15	17.89	< 0.01
Platelets	log10(G/L)	1.85	3.13	2.61	0.03	14.07	< 0.01
Mean platelet volume	fL	8.00	14.20	10.05	0.23	0.68	0.57
Platelet distribution width	%	9.90	11.90	10.76	0.09	4.08	< 0.01
Birth weight	Kg	0.40	2.68	1.50	0.07	NR	NR
Weaning weight	Kg	5.46	16.56	9.4	0.35	NR	NR

F value and FDR are for the test of the global time effect on each variable. *NR* non relevant since the measure is the same at all time steps

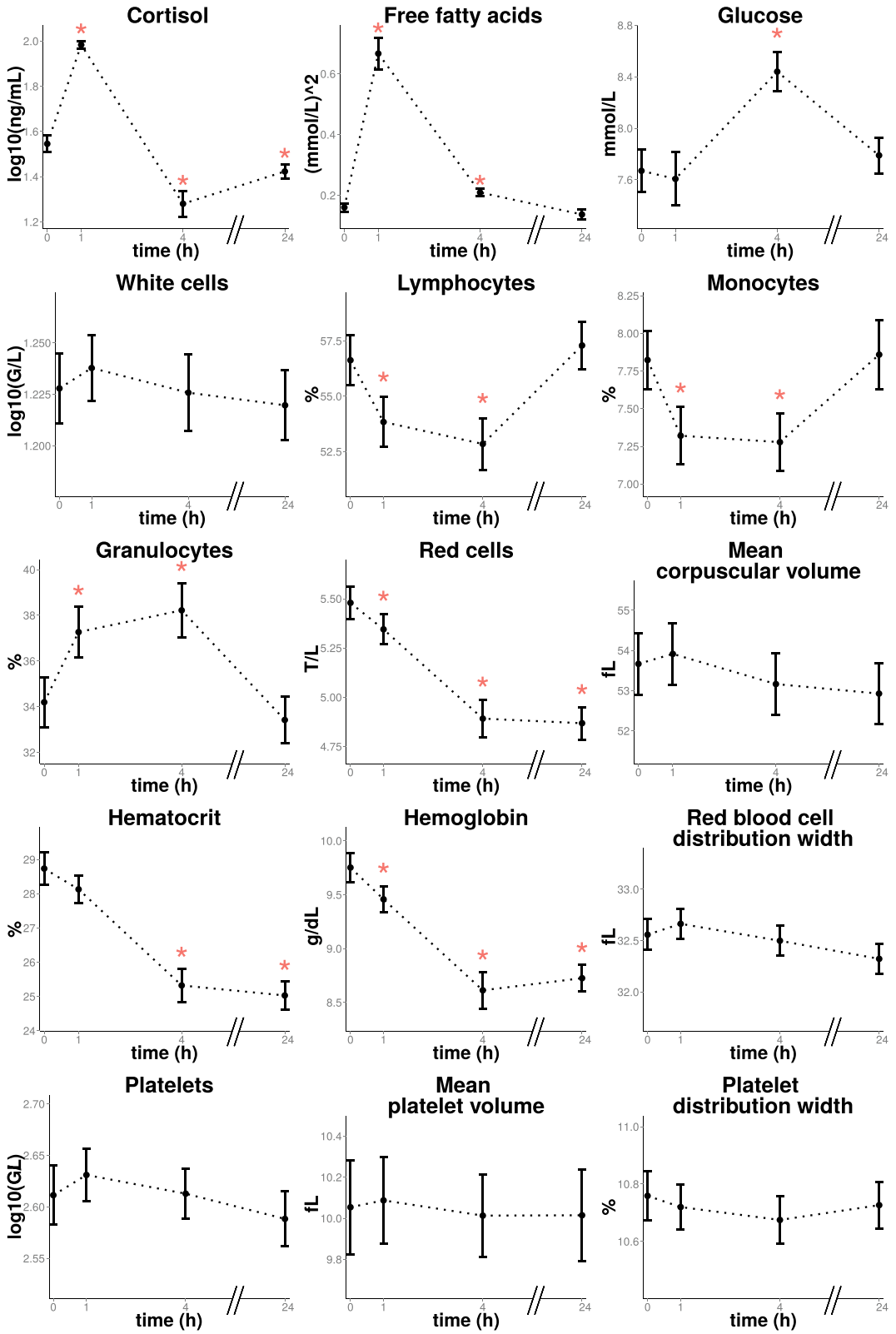


Fig. 1 Mean evolution of the biological variables overtime. (*): time measurement at which the expression of the variable is significantly different ($FDR < 0.05$) from the expression at $t = 0$. Vertical bars correspond to + and − SEM at each time point

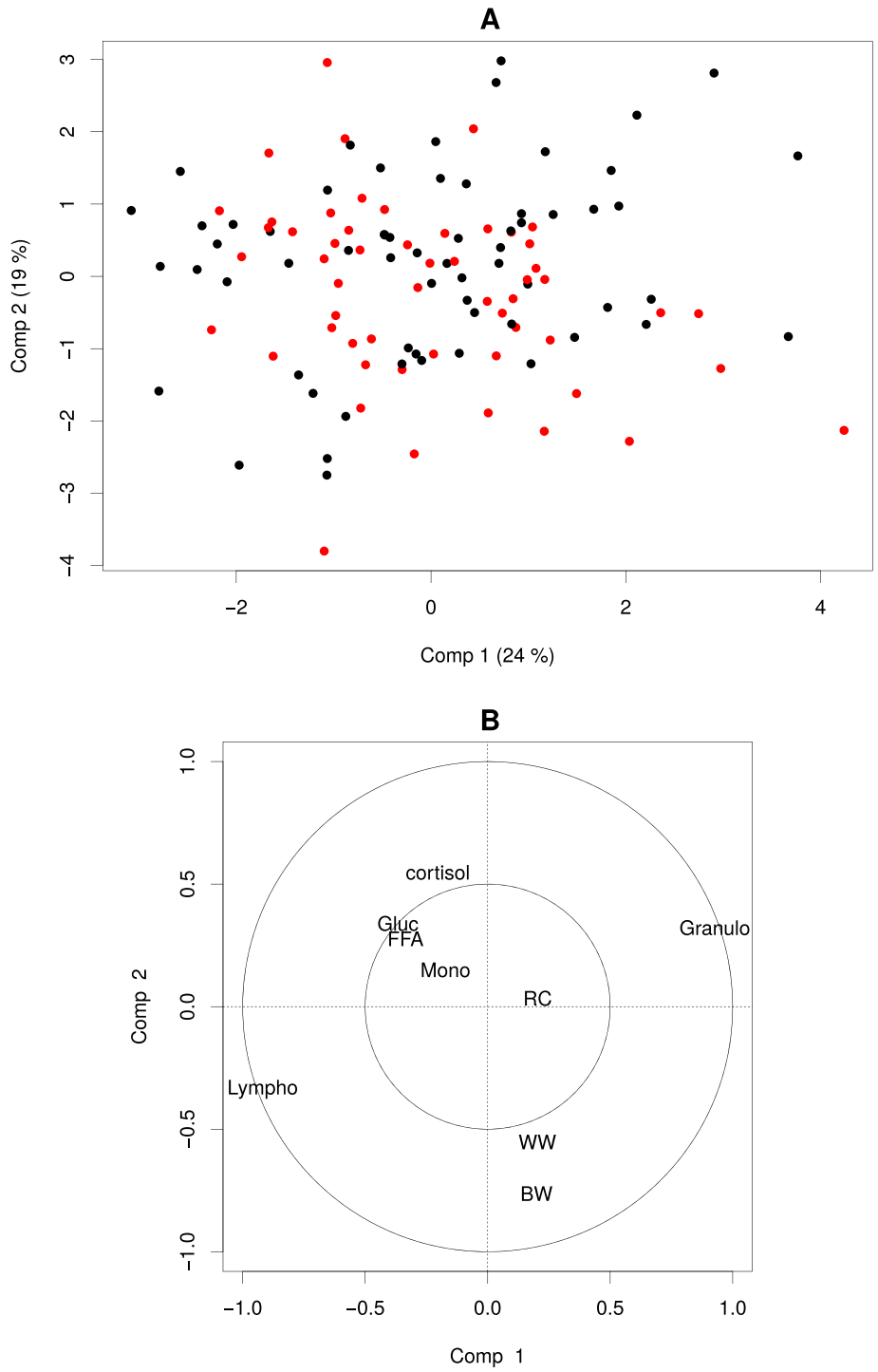


Fig. 2 PCA on the biological variables identified as responding to ACTH, the birth and weaning weights at $t = 0$. Colors symbolize the sex: Black = Female; Red = Male **a** Projection of the individuals on dimensions 1 and 2; **b** Projection of the variables on dimensions 1 and 2; BW = birth weight; WW = weaning weight; Lympho = lymphocyte ratio; Mono = monocytes ratio; Granulo = granulocyte ratio; RC = red cell count; Gluc = glucose; FFA = free fatty acids

birth and (to a lesser extent) weaning weights (negatively correlated with this axis) with a strong opposition between these variables on the whole plan formed by the first and second axis. The other variables were not correlated with either of the first two dimensions of the PCA.

Overall effect of the injection of exogenous ACTH on clinical biology variables

Extraction of the within-subject data matrix prior to the application of a PCA analysis allows for the separation of the observations according to their time of measurement (see Fig. 3). The first component of the multi-level PCA opposes the observations at $t = 0$ and $t = +24$ (positive coordinates on this axis) to the observations at $t = +1$ (negative coordinates on this axis), this time step corresponding to the peak of cortisol. The second component opposes the observations at $t = +4$ (positive coordinates on this axis) to the observations at $t = +1$ (negative coordinates on this axis). The representation of the variables shows that the first axis is mainly driven by an opposition between the proportion of granulocytes, FFA, cortisol and red cell count (high measures at $t = +1$), on one side, and lymphocytes and monocytes (high measures at $t = 0/+24$), on the other side. The second axis shows an opposition between glucose (positively correlated with this axis, high measures at $t = +4$) and cortisol, FFA and red count (negatively correlated with this axis, high measures at $t = +1$).

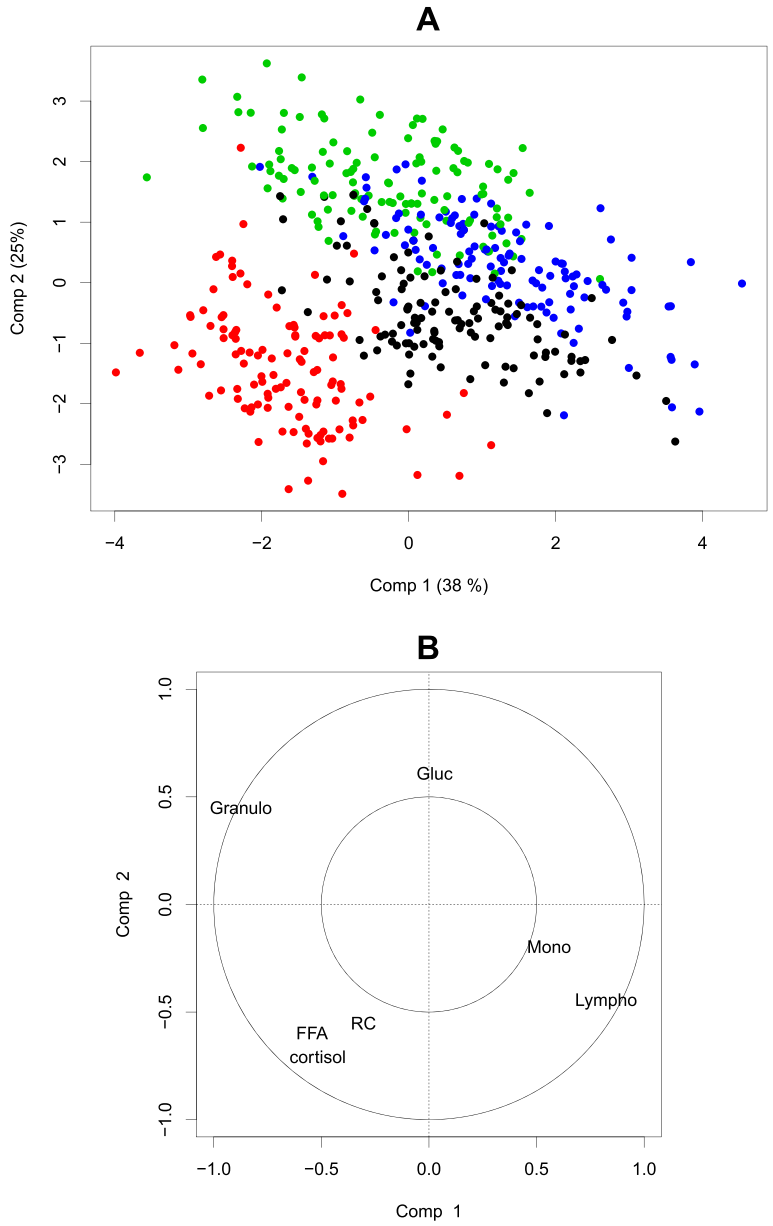


Fig. 3 Multilevel PCA on the biological variables responding to ACTH. Colors symbolize the time of measurement; Black: $t = 0$; Red: $t = +1$; Green: $t = +4$; Blue: $t = +24$; **a** Projection of the individuals on dimensions 1–2; **b** Projection of the variables on dimensions 1–2; Lympho = lymphocyte ratio; Mono = monocyte ratio; Granulo = granulocyte ratio; RC = red cell counts; Gluc = glucose; FFA = free fatty acids

Specific links to the level of cortisol at $t = +1$

Correlations between cortisol at $t = +1$ and other variables at $t = 0$ allows for the identification of variables which baseline levels may be directly or indirectly linked to the intensity of the cortisol level in response to ACTH, a measure of individual variation in HPA axis activity. Correlations are shown in Table 2. Only glucose and FFA levels at $t = 0$ were significantly positively correlated with the level of cortisol at $t = +1$ ($FDR < 0.05$). Correlations between cortisol at $t = +1$ and other variables at $t = +1, t = +4$ or $t = +24$ allows for the identification of variables which are directly linked to the level of cortisol when it reaches its peak during the stress response. FFA at $t = +1$ were positively correlated with cortisol at $t = +1$ and glucose at $t = +1, t = +4$ and $t = +24$ was negatively correlated with cortisol at $t = +1$ ($FDR < 0.05$). No other variable was found to be significantly linked to the intensity of the cortisol level in response to ACTH.

Differentially expressed genes

We used a comprehensive gene expression profiling by means of microarray analysis to identify clusters of genes differentially expressed in peripheral blood cells, taking into consideration the kinetics of the response with 4 time points ($t \in \{0, +1, +4, +24\}$). Differential analysis revealed 158 DE transcripts (adjusted $P < 0.05$) matching 65 unique genes (The complete list with features is provided in Additional file 1). Among them, 23 genes were differentially expressed at $t = +1$ (5 down regulated/18 up-regulated), 25 were differentially expressed at $t = +4$ (8 down-regulated/17 up-regulated) and 17 were differentially expressed at $t = +24$ (all down-regulated). The only gene DE at both $t = +1$ and $t = +4$ was *SUCNR1* (Table 3). The adjusted *P*-values were smaller for tests between $t = 0$ and $t = +1$ and between $t = 0$ and $t = +4$ than between $t = 0$ and $t = +24$ (see Additional file 2). This shows that the transcripts were more differentially expressed between $t = 0$ and $t = +1$ and between $t = 0$ and $t = +4$ than between $t = 0$ and $t = +24$. Main effects of cortisol released by ACTH injection on gene expressions are thus observed at $t = +1$ and $t = +4$ with a return to baseline levels at $t = +24$.

Table 2 Correlation coefficients between the biological variables at $t = 0, t = +1, t = +4$ and $t = +24$ and cortisol at $t = +1$ ($n = 120$)

Variables	$t = 0$ (SE)	$t = +1$ (SE)	$t = +4$ (SE)	$t = +24$ (SE)
Free fatty acids	**0.35 (< 0.07)**	**0.45 (< 0.07)**	−0.13 (0.09)	0.04 (0.09)
Glucose	**0.30 (0.08)**	**−0.25 (0.08)**	**−0.27 (0.08)**	**−0.27 (0.08)**
Lymphocytes	−0.09 (0.09)	−0.10 (0.09)	−0.13 (0.09)	−0.07 (0.09)
Monocytes	−0.05 (0.09)	−0.02 (0.09)	−0.05 (0.09)	−0.04 (0.09)
Granulocytes	0.12 (0.09)	0.14 (0.09)	0.15 (0.09)	0.09 (0.09)
Red cells	0.06 (0.09)	0.07 (0.09)	−0.11 (0.09)	0.06 (0.09)
Hematocrit	0.06 (0.09)	0.13 (0.09)	−0.08 (0.09)	0.12 (0.09)
Hemoglobin	0.13 (0.09)	0.18 (0.08)	−0.15 (0.09)	0.03 (0.09)

SE: standard error of the correlation coefficient; **in bold**: significantly $\neq 0$ ($FDR < 0.05$)

HAC performed on the within-subject deviation matrix with the list of DE genes identified 4 groups of genes. Figure 4 shows that the 65 unique DE genes allow for an almost perfect classification of the observations with respect to their time of measurement. For every cluster, Fig. 5 shows the average evolution of each gene (gray) and the average evolution over the genes in the cluster (red).

Each cluster was then subjected to a functional analysis (results shown in Additional file 3). In each cluster, genes were DE ($FDR < 0.01$) at each time point except for $t = +24$ in cluster 3 ($FDR = 0.57$).

The first cluster (17 genes) was characterized by genes increasing with a peak of expression at $t = 1$ and stable between $t = +4$ and $t = +24$. The DE genes of this cluster could be assembled into a functional network principally involved in neuroimmune functions. The present analysis reveals novel effects of ACTH on at least five genes related to immunoregulation (*FKBP5*, *IL7R*, *CEBPD*, *CEBPB* and *NFKB1A* in cluster 4). *FKBP5* (FK506 binding protein 51) is a decisive factor for the physiological stress response [33] and has an important role in stress-related phenotypes [34]. It modifies steroid hormone receptor sensitivity [35]. *CEBPB*, *DUSP1*, *FKBP5* and *NFKB1A* genes from this cluster are also involved in glucocorticoid receptor signaling. Glucocorticoids exert their classic anti-inflammatory role by acting on nearly all cell types of the immune system. The CCAAT/enhancer binding proteins (C/EBPs) are key regulators of cell differentiation and are also involved in the expression and production of inflammatory cytokines [36]. The increase of Period 1 gene (*PER1*) expression in peripheral blood cells by glucocorticoids was previously reported in humans [37]. Physical and inflammatory stressors induce the release of the adrenal glucocorticoid hormone that rapidly alter the expression of *PER1* in peripheral tissues through a GRE enhancer present in the gene promotor [38–40]. This gene is involved in the circadian rythm, in which the glucocorticoid mechanism plays a predominant role [41]. Another DE gene *DDIT4* (regulated in development and DNA damage response 1) was described as a surrogate biomarker of the efficiency of glucocorticoid receptor blockade in skeletal muscle [42]. Britto and collaborators showed that *DDIT4* expression was low under basal conditions but was highly increased in response to several catabolic stressors, like hypoxia and glucocorticoids [43]. Glucocorticoids were shown to up-regulate *DUSP1* in peripheral tissues [44] but constrain the increase of *DUSP1* gene expression in the central components of the HPA axis [45]. In vitro studies have shown that glucocorticoid suppression of some MAP-kinase dependent

Table 3 List of 65 unique genes differentially expressed in response to ACTH in pigs ($n = 30$)

	Gene name	Adjusted P	Time point	Expression	Cluster
1	ADCY2	1.87E-03	1	UP-regulated	1
2	CEBPB	3.72E-09	1	UP-regulated	1
3	CEBPD	6.65E-07	1	UP-regulated	1
4	CPT1A	2.28E-06	1	UP-regulated	1
5	CXCR4	9.47E-06	1	UP-regulated	1
6	DDIT4	3.96E-07	1	UP-regulated	1
7	DUSP1	7.53E-03	1	UP-regulated	1
8	FKBP5	4.39E-06	1	UP-regulated	1
9	G30866	8.88E-05	1	UP-regulated	1
10	G39878	9.56E-04	1	UP-regulated	1
11	IL7R	6.63E-05	1	UP-regulated	1
12	MXD1	1.71E-03	1	UP-regulated	1
13	NFKBIA	2.72E-03	1	UP-regulated	1
14	PER1	9.37E-05	1	UP-regulated	1
15	PIK3IP1	2.89E-04	1	UP-regulated	1
16	RGS2	8.52E-08	1	UP-regulated	1
17	RHOB	4.08E-02	1	UP-regulated	1
18	TXNIP	1.78E-03	1	UP-regulated	1
19	ALOX5AP	1.20E-03	4	UP-regulated	2
20	ANG1	1.92E-02	4	UP-regulated	2
21	BASP1	4.11E-02	4	UP-regulated	2
22	C2H19orf59	1.06E-02	4	UP-regulated	2
23	CD14	3.99E-04	4	UP-regulated	2
24	CD24	1.82E-04	4	UP-regulated	2
25	CHI3L1	1.73E-02	4	UP-regulated	2
26	CHIT1	2.16E-02	4	UP-regulated	2
27	CLC4D	2.00E-03	4	UP-regulated	2
28	CRLD2	4.40E-02	4	UP-regulated	2
29	G42218	6.47E-03	4	UP-regulated	2
30	MEGF9	1.92E-04	4	UP-regulated	2
31	PDPN	2.08E-02	4	UP-regulated	2
32	RAB31	2.74E-02	4	UP-regulated	2
33	S100A12	5.47E-03	4	UP-regulated	2
34	S100A8	5.26E-03	4	UP-regulated	2
35	S100A9	2.92E-03	4	UP-regulated	2
36	CCL8	1.36E-04	1	DOWN-regulated	3
37	ALOX15	2.03E-07	4	DOWN-regulated	3
38	CAMK1	2.84E-09	4	DOWN-regulated	3
39	CSTA	9.20E-09	4	DOWN-regulated	3
40	FBP1	1.03E-04	4	DOWN-regulated	3
41	G36094	6.98E-10	4	DOWN-regulated	3
42	SLCO2B1	5.40E-13	4	DOWN-regulated	3
43	SUCNR1	2.58E-08	1 & 4	DOWN-regulated	3
44	CD79B	7.92E-04	1	DOWN-regulated	4
45	HHEX	3.32E-02	1	DOWN-regulated	4
46	MZB1	5.04E-03	1	DOWN-regulated	4
47	ST14	2.51E-02	1	DOWN-regulated	4
48	LOC396700	3.78E-06	4	DOWN-regulated	4
49	AKAP13	1.89E-02	24	DOWN-regulated	4
50	ARHGAP31	2.37E-03	24	DOWN-regulated	4
51	CLK1	2.52E-02	24	DOWN-regulated	4
52	DCAF15	1.56E-02	24	DOWN-regulated	4
53	FGR	6.95E-03	24	DOWN-regulated	4
54	G48605	8.08E-04	24	DOWN-regulated	4
55	HOPX	1.54E-02	24	DOWN-regulated	4
56	IGLV_7	3.07E-02	24	DOWN-regulated	4
57	LAS1L	4.12E-02	24	DOWN-regulated	4
58	LOC100626276	1.25E-03	24	DOWN-regulated	4
59	LOC396781	3.36E-02	24	DOWN-regulated	4
60	MAPK6	1.49E-03	24	DOWN-regulated	4
61	ORAI1	1.83E-03	24	DOWN-regulated	4
62	S100A1	5.43E-04	24	DOWN-regulated	4
63	TPST2	9.40E-04	24	DOWN-regulated	4
64	TRMT2A	3.61E-02	24	DOWN-regulated	4
65	XCL1	3.01E-02	24	DOWN-regulated	4

Genes are divided into clusters corresponding to the kinetics of the response to ACTH. Full description of the genes including their probe name and localisation is displayed in Additional file 1

cellular processes depends on glucocorticoid mediated up-regulation of *DUSP1* gene expression [46].

The second cluster (17 genes) was characterized by genes with an increase between $t = 0$ and $t = +4$ and a decrease between $t = +4$ and $t = +24$. This cluster with genes up-regulated at $t = +4$ is largely related to biological processes such as inflammatory and immune response and genes of which products are located in the plasma membrane. Among these genes, two are particularly interesting. *CD14* gene is a component of the innate immune system and has been shown to be sensitive to stress in pigs [47]. *MEGF9* gene was shown to be induced by cortisol in human fetal cells in vitro [48].

The third cluster (8 genes) includes the genes decreasing between $t = 0$ and $t = +4$ and returning to a basal level between $t = +4$ and $t = +24$. No ontology was significantly enriched by genes of this cluster. It is interesting to underline here the *ALOX15* gene (arachidonate 15-lipoxygenase) which is a member of the ALOX family and related to cancer and immune responses. This gene was also reported as a dexamethasone-responsive gene with nearby glucocorticoid receptor-binding sites [49].

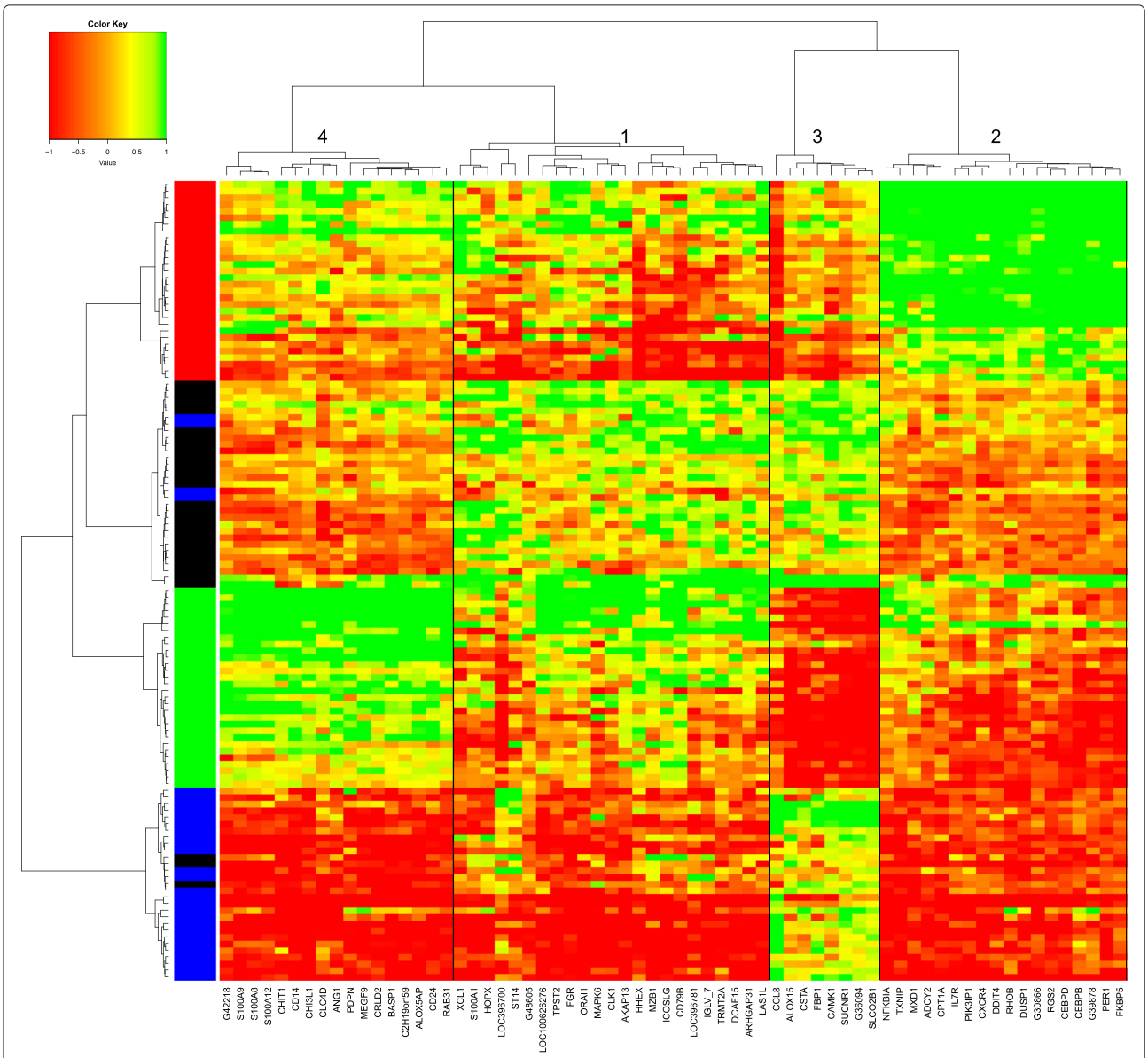


Fig. 4 Hierarchical ascending classification of the 65 unique DE genes. A Ward method was used with an Euclidean distance matrix based on the correlations between genes. Genes are shown in column. Observations are shown in line with one line being a combination pig × time. Colors on the row dendrogram identify the time of measurement. Black: $t = 0$; Red: $t = +1$; Green: $t = +4$; Blue: $t = +24$. Numbers on the column dendrogram identify each cluster

The genes related to the fourth cluster (22 genes) decrease between $t = 0$ and $t = +1$, increase between $t = +1$ and $t = +4$ and decrease between $t = +4$ and $t = +24$. The fourth cluster corresponds to genes with an overall expression decreasing between $t = 0$ and $t = +24$. They are significantly linked to biological processes such as protein phosphorylation and kinase activity. Among the genes involved in this cluster *ARHGAP31* and *ARHGAP* family genes were found to be differentially expressed in macrophages treated with dexamethasone [50, 51].

Integration of biological and gene expression data

All DE genes were found significantly differentially expressed over time in the mixed model described in Section "Integration" ($FDR < 0.05$). These genes are thus differentially expressed over time even when adjusting for changes in white cells populations. Among them, 34 genes had their expression significantly negatively influenced by L/G ratio (see Additional file 1), meaning that these genes are over-expressed when L/G ratio decreases. Genes with a significant effect of L/G ratio were mainly identified as

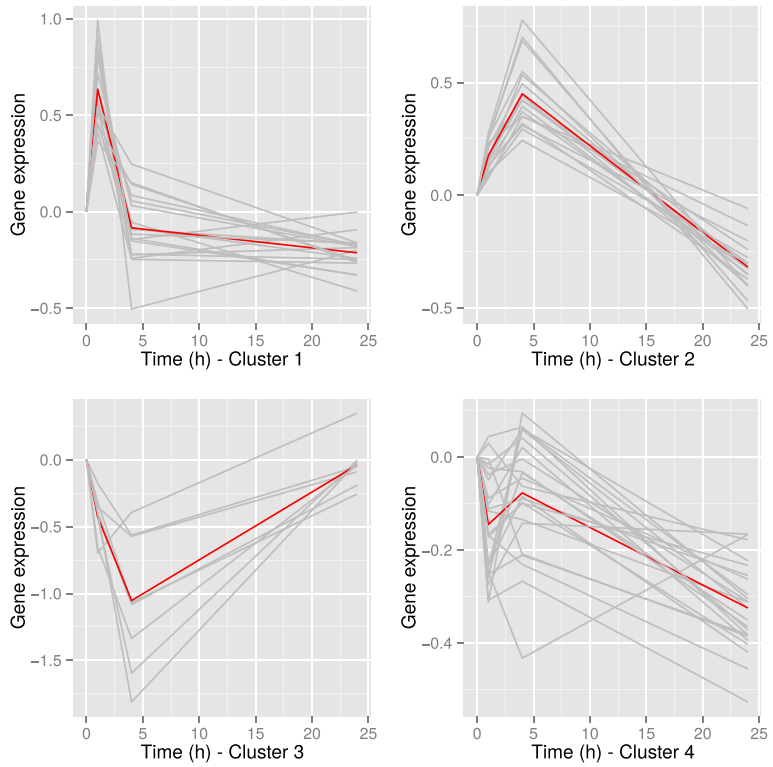


Fig. 5 Average evolution of the genes in each of the cluster identified by the HAC on the 65 unique DE genes. Evolution of each gene is translated so that it is equal to 0 at $t = 0$; Gray: Average evolution of each of the genes in the cluster. Red: Average evolution over all genes in the cluster (cluster 1: 18; cluster 2: 17; cluster 3: 8; cluster 4: 22)

genes of cluster 2, over-expressed at $t = +4$ (17/17) and cluster 1, over-expressed at $t = +1$ (12/18) and to a lesser extent as genes of cluster 4, under-expressed at $t = +24$ (5/22). No gene of cluster 3 was significantly explained by L/G ratio. Results of the functional analysis of this list of 34 genes are shown in Additional file 4. Biological functions significantly enriched include regulation of apoptotic process, response to lipopolysaccharide, inflammatory and innate immune response, defense response to bacterium and positive regulation of NF-kappaB transcription factor activity.

The 65 DE genes and the biological variables were then subjected to a multilevel PLS. Figure 6 shows that the first axis of the multilevel PLS opposes observations at $t = +1$ after injection to all others, while the second axis opposes observations at $t = +4$ *vs.* all others, similarly as what was already established in multi-level PCA of the biological variables in Section "Overall effect of the injection of exogenous ACTH on clinical biology variables".

On the first axis, cortisol and FFA levels are strongly positively correlated with the expressions of *CEBPB*, *RGS2*, *RHOB*, *PER1*, *FKBP5*, *CEBPD*, *DDIT4*, *CPT1A* and *DUSP1*. All these genes belong to the first cluster identified earlier and are linked to molecular functions such as protein binding and transcription regulation.

The second axis of the multilevel PLS is characterized by the opposition between the proportion of lymphocytes and monocytes *vs.* the proportion of granulocytes. This axis is positively correlated with *SUCNR1*, *SLCO2B1*, *FBP1* and *LOC396700*. These genes belong to the third cluster and are related to glycolysis and glycogenesis. *SUCNR1* (succinate receptor 1) is decreased at $t = +1$ and increased at $t = +4$. Succinate has a wide range of metabolic actions and regulates the functions of macrophages [52]. The axis is negatively correlated with *CD14*, *CLC4D*, *CHIT1*, *MEGF9* and *C2H19orf59*. These genes belong to the second cluster which is linked to molecular functions such as inflammatory response, but their relationships with cortisol or stress are not yet clearly established.

Eight genes (*DDIT4*, *DUSP1*, *FKBP5*, *IL7R*, *NFKBIA*, *PER1*, *RGS2*, *RHOB*, Fig. 7) are functionally connected to each other by NR3C1. The NR3C1 (nuclear receptor subfamily 3, group C, member 1) is the glucocorticoid receptor, which can function both as a transcription factor that binds to glucocorticoid response elements in the promoters of glucocorticoid responsive genes, and as a regulator of other transcription factors. Functional consequences of glucocorticoid receptor polymorphisms were reported in pigs [9–13]. Mutations in *NR3C1* have

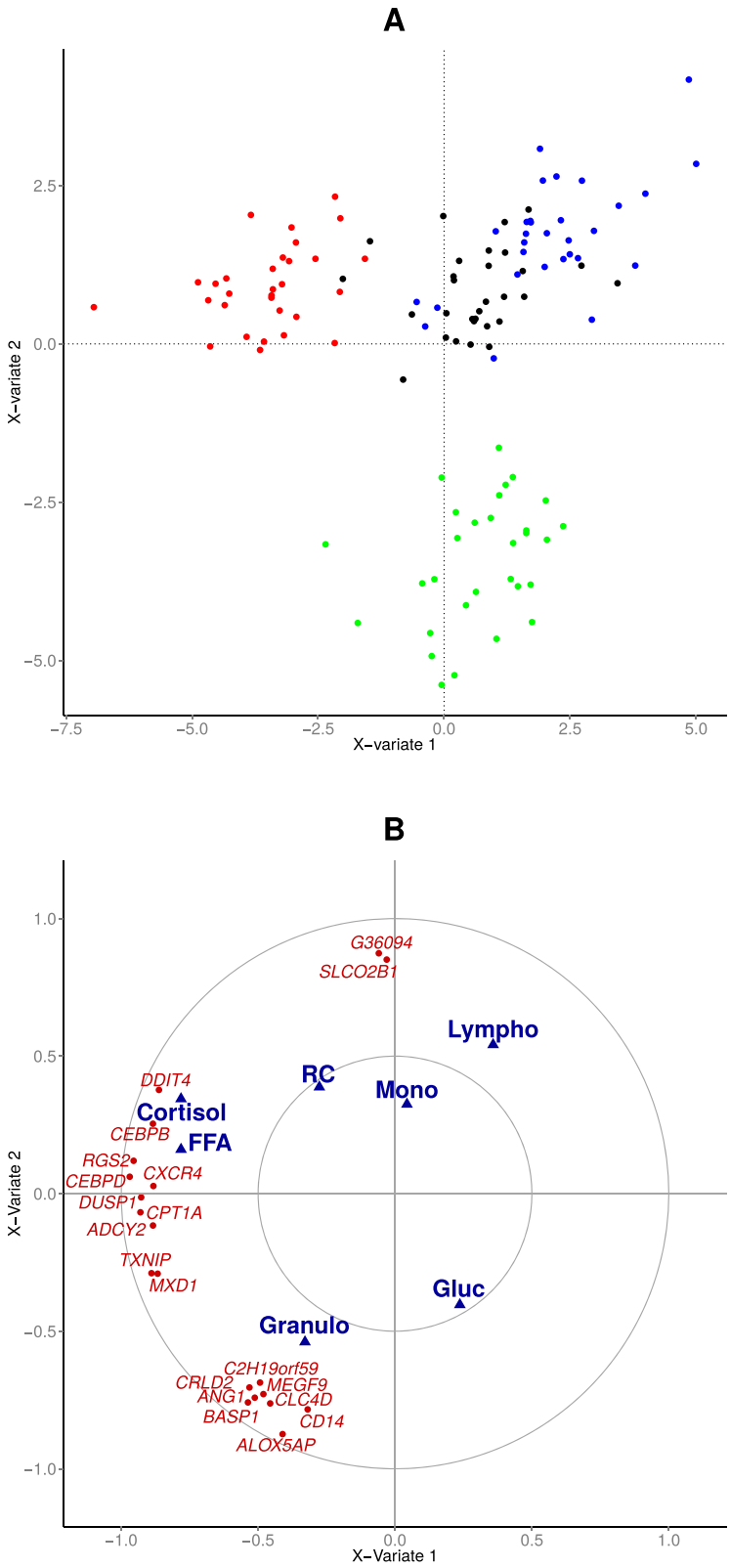


Fig. 6 PLS regression predicting the biological variables responding to ACTH from the DEG expression **a** projection of the observations with one point being a combination: time × pig; Black: $t = 0$; Red: $t = +1$; Green: $t = +4$; Blue: $t = +24$. **b** projection of the variables; Blue: biological variables; Red: gene expressions; 10 genes/components were kept using a sparse approach

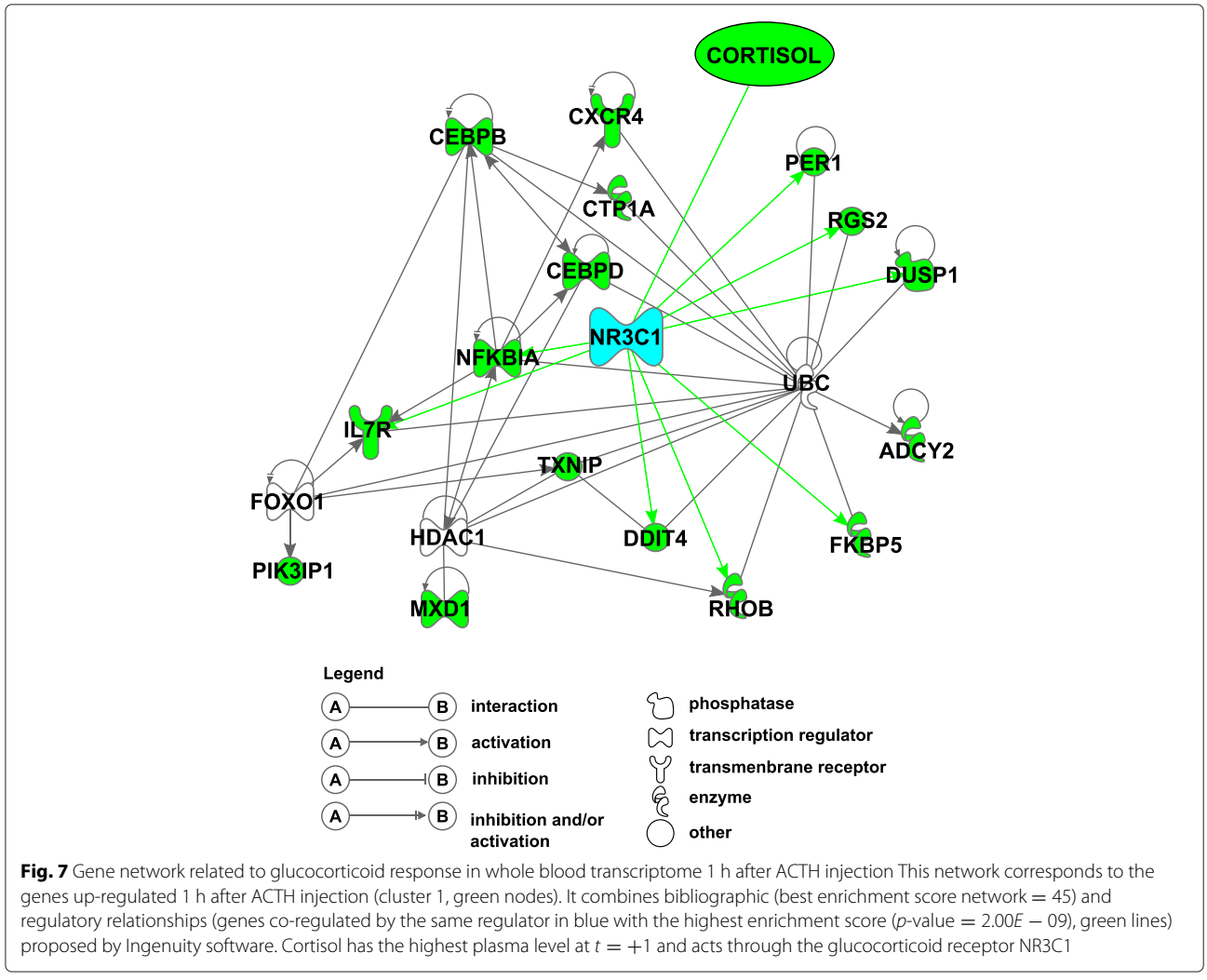


Fig. 7 Gene network related to glucocorticoid response in whole blood transcriptome 1 h after ACTH injection This network corresponds to the genes up-regulated 1 h after ACTH injection (cluster 1, green nodes). It combines bibliographic (best enrichment score network = 45) and regulatory relationships (genes co-regulated by the same regulator in blue with the highest enrichment score (p-value = $2.00E-09$), green lines) proposed by Ingenuity software. Cortisol has the highest plasma level at $t = +1$ and acts through the glucocorticoid receptor NR3C1

been previously demonstrated to be associated with generalized glucocorticoid resistance [53]. It is interesting to highlight the DE genes that encode transcription factors. They play a crucial role in regulating gene expression and are fit to regulate diverse cellular processes by interacting with other proteins. Most of them have not yet been described as important in transcription networks involved in stress responses. If the genes are co-expressed it is highly probable that they are co-regulated. This knowledge can provide new patterns of biomarkers of the individual sensitivity to cortisol that is our field of interest in this study.

Our results are in accordance with several studies on the effects of glucocorticoid hormones on peripheral blood cells. Numerous genes related to cluster 1 and shown as ACTH responsive were found differentially expressed in stress-related investigations. Five genes found in our study (*CXCR4*, *DUSP1*, *FKBP5*, *IL7R*, *TXNIP*) were proposed as markers of differential glucocorticoid sensitivity [54, 55]. *NFKBIA*, *DUSP1*, *CEBPD*, *FKBP5* genes were also found to be associated with up- and down-regulated clusters in response to continuous 24 h cortisol infusion [56]. Ponsuksili and collaborators [57] describe *NFKBIA*, *CEBPB* and *CEBPD* as genes of which hepatic expression levels are correlated with plasma cortisol concentrations. Up-regulation of *PER1* gene upon GR activation was confirmed by genome-wide study of glucocorticoid receptor binding sites in neuronal PC12 cells [58]. However, *DDIT4* was shown to be down-regulated by GR activation rather than up-regulated in this analysis.

While looking for genes of which expressions at $t = 0$, $t = +1$, $t = +4$ or $t = +24$ were significantly correlated with the level of cortisol at $t = +1$, only two genes were identified: *TRMT2A* ($FDR = 0.04$), a gene involved in the methylation of tRNA, and *LOC100626276* ($FDR = 0.04$), a gene of which function has not been identified yet. There is a negative relationship between the expression of cortisol at $t = +1$ and the expression of these two genes at $t = 0$ (-0.45 and -0.63 for *TRMT2A* and *LOC100626276*, respectively). This implies that when

their baseline expression is higher, the intensity of the cortisol response to ACTH decreases.

Conclusions

The present work shows the interest of transcriptomic data analysis at multiple levels. In other studies, genetic markers found through an analysis of transcription factor binding sites of differentially expressed genes in peripheral blood cells have been proposed in humans to identify the chronic stress related to psychopathological conditions [59, 60]. In farm animals, this approach was used in horses [61]. These studies show chronic stress-related changes in the balance between the expression of stress-related genes regulated by glucocorticoids and those regulated by inflammation-related factors. Furthermore, recent data in humans show that the immune system function can also be assessed through blood transcriptomics in health and disease [62].

In the present study, we identified 65 genes differentially expressed in peripheral blood cells of pigs in response to ACTH at different times after injection. It therefore supplies biological markers of HPA axis activation at the gene expression level, and the knowledge on functional gene clusters will help to elucidate the biological processes involved. Moreover, these genes might be candidates for a better understanding of the molecular mechanisms related to stress responses. Thus, blood transcriptome analysis appears as a promising avenue to develop multidimensional biological markers related to robustness. These markers should be used in the study of the genetic mechanisms of adaptation in farm animals that will help to deliver genetic strategies to animal breeders in order to balance production objectives and robustness of animals as well as their welfare [2].

Availability of supporting data

The data sets supporting the results of this article are available in Gene Expression Omnibus (GEO repository, http://www.ncbi.nlm.nih.gov/geo/,) through the accession number GSE71207).

Additional files

Additional file 1: List of 65 unique genes differentially expressed in response to ACTH in pigs ($n = 30$) '.xls' file. Genes are divided into clusters corresponding to the kinetics of the response to ACTH. Probe names are those used on the microarray Agilent GPL16524.

- Gene name: Name of the gene
- Probe Name, Agilent GPL16524: Probe names used on the microarray Agilent GPL16524
- L/G coefficient: L/G ratio coefficient estimate
- Adjusted pval (LG): adjusted *P*-value associated with the L/G ratio coefficient
- adj.pval: adjusted *P*-value of the test at the time measurement where the most significant duplicate of the gene is DE
- Time point: time measurement where the gene is DE
- Expression: whether the DEG is up or down-regulated
- Cluster: cluster in which the gene is classified by HAC
- Gene description: informations on the gene's molecular function
- Location: chromosomal location of the gene. (XLS 23 kb)

Additional file 2: Distribution of the rank of the significant adjusted *P*-values in the tests for DE transcripts between $t = 0$ and $t = +1$, $t = 0$ and $t = +4$ and $t = 0$ and $t = +24$ '.pdf' file. *P*-values are smaller at $t = +1$ and $t = +4$ than at $t = +24$ implying that the transcripts were overall more differentially expressed between $t = 0$ and $t = +1$ and between $t = 0$ and $t = +4$ than between $t = 0$ and $t = +24$. (PDF 4 kb)

Additional file 3: Complete list of enriched GO (Biological process (BP), Molecular function (MF) and Cellular Component (CC) for each of the cluster identified with the hierarchical ascending clustering '.xls' file. Features the GO items, the corresponding functions, the class of ontology, the number of genes in the input list (enriching a GO and total number) and in the reference list (enriching a GO and total number), the raw and the adjusted Fisher's exact test *P*-value and the list of genes. (XLS 10 kb)

Additional file 4: Complete list of enriched GO (Biological process (BP), Molecular function (MF) and Cellular Component (CC) for 34 genes for which L/G ratio had a significant effect '.xls' file. Features the GO items, the corresponding functions, the class of ontology, the number of genes in the input list (enriching a GO and total number) and in the reference list (enriching a GO and total number), the raw and the 679 adjusted Fisher's exact test *P*-value and the list of genes. (XLS 7 kb)

Abbreviations

ACTH: adrenocorticotropic hormone; BH: Benjamini-Hochberg; BW: birth weight; CBG: corticosteroid binding globulin; DE: differentially expressed; FDR: false discovery rate; GR: glucocorticoid receptor; GRE: Glucocorticoid response element; HAC: Hierarchical ascendant classification; Hct: Hematocrit; Hgb: Hemoglobin; HPA: Hypothalamic-pituitary-adrenocortical; L/ G: Lymphocytes/ granulocytes; PCA: prinicpal component analysis; PLS: partial least square regression; RC: red cells; SEM: standar error of the mean; WW: weaning weight.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VS and NVV developed and performed all the statistics. PM and ET developed and undertook the experimental design and ensured the biological interpretation. LG, ET and LL ensured the experimental transcriptomic analysis. YL provided the transcriptomic data set VS, ET and LL performed the transcriptomic related biology interpration. LG performed the sampling and the data management. CL performed the genetics and the choice of animals. YB was responsible of the animal management. VS, NVV, ET, PM and LL wrote the paper. PM is the project manager. All authors read and approved the final manuscript.

Acknowledgements

This project received financial support from the French National Agency of Research (SUSoSTRESS, ANR-12-ADAP-0008) which also funds the PhD thesis of VS together with the Région Midi-Pyrénées. We thank the team of the Magneraud center which took care of animal breeding, GeT-TRiX platform (Toulouse) for expression data production, GenPhySE INRA Toulouse (more specificaly Nathalie Iannuccelli, Katia Fève and Juliette Riquet) for experimental support, ANEXPLO platform (Toulouse) for biological dosages, PEGASE INRA Saint-Gilles (more specifically Rafael Comte) for cortisol assay, Aurélie Ducan for hematology analysis, Magali San Cristobal and Pascal Martin for helpful discussions on the analyses and the French National Research Agency for funding the research project.

Author details

[1]INRA, UMR 1388 Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France. [2]Université de Toulouse INPT ENSAT, UMR 1388 Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France. [3]Université de Toulouse INPT ENVT, UMR 1388 Génétique, Physiologie et Systèmes d'Elevage, F-31076 Toulouse, France. [4]INRA, UMR 1331 ToxAlim, F-31027 Toulouse, France. [5]INRA, UE 1372 GenESI, F-17700 Surgeres, France. [6]INRA, UR 0875 MIAT Mathématiques et Informatique Appliquées de Toulouse, F-31326 Castanet-Tolosan, France.

Received: 6 August 2015 Accepted: 20 October 2015
Published online: 17 November 2015

References

1. Rauw W, Kanis E, Noordhuizen-Stassen E, Grommers F. Undesirable side effects of selection for high production efficiency in farm animals: a review. Livest Prod Sci. 1998;56(1):15–33.
2. Mormede P, Terenina E. Molecular genetics of the adrenocortical axis and breeding for robustness. Domest Anim Endocrinol. 2012;43(2):116–31.
3. Mormède P, Foury A, Terenina E, Knap P. Breeding for robustness: the role of cortisol. Animal. 2011;5(05):651–7.
4. Mormede P, Foury A, Barat P, Corcuff JB, Terenina E, Marissal-Arvy N, et al. Molecular genetics of hypothalamic–pituitary–adrenal axis activity and function. Ann N Y Acad Sci. 2011;1220(1):127–36.
5. Désautés C, Bidanel J, Milan D, Iannuccelli N, Amigues Y, Bourgeois F, et al. Genetic linkage mapping of quantitative trait loci for behavioral and neuroendocrine stress response traits in pigs. J Anim Sci. 2002;80(9): 2276–285.
6. Larzul C, Terenina E, Foury A, Billon Y, Louveau I, Merlot E, et al. The cortisol response to ACTH in pigs, heritability and influence of corticosteroid-binding globulin. 2015. In press.
7. Hazard D, Liaubet L, SanCristobal M, Mormède P. Gene array and real time pcr analysis of the adrenal sensitivity to adrenocorticotropic hormone in pig. BMC genomics. 2008;9(1):101.
8. Necela BM, Cidlowski JA. Mechanisms of glucocorticoid receptor action in noninflammatory and inflammatory cells. Ann Am Thorac Soc. 2004;1(3):239–46.
9. Murani E, Reyer H, Ponsuksili S, Fritschka S, Wimmers K. A substitution in the ligand binding domain of the porcine glucocorticoid receptor affects activity of the adrenal gland. PLoS ONE. 2012;7(9):e45518.
10. Yang X, Liu R, Albrecht E, Dong X, Maak S, Zhao R. Breed-specific patterns of hepatic gluconeogenesis and glucocorticoid action in pigs. Archiv Tierzucht. 2012;1:152–62.
11. Reyer H, Ponsuksili S, Wimmers K, Murani E. Transcript variants of the porcine glucocorticoid receptor gene (nr3c1). Gen Comp Endocrinol. 2013;189:127–33.
12. Reyer H, Ponsuksili S, Wimmers K, Murani E. Association of n-terminal domain polymorphisms of the porcine glucocorticoid receptor with carcass composition and meat quality traits. Anim Genet. 2014;45(1): 125–9.
13. Terenina E, Babigumira BM, Le Mignon G, Bazovkina D, Rousseau S, Salin F, et al. Association study of molecular polymorphisms in candidate genes related to stress responses with production and meat quality traits in pigs. Domest Anim Endocrinol. 2013;44(2):81–97.
14. Torgo L. Data Mining with R: Learning with Case Studies. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Boca raton: Taylor & Francis; 2010.
15. Voillet V, SanCristobal M, Lippi Y, Martin PG, Iannuccelli N, Lascor C, et al. Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity. BMC Genom. 2014;15(1): 797.
16. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93.
17. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. https://www.R-project.org.
18. Liquet B, Lê Cao KA, Hocini H, Thiébaut R. A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. BMC Bioinform. 2012;13(1):325.
19. Saporta G. Probabilités, Analyse des Données et Statistique. Paris: Editions Technip; 2011.
20. Lê Cao KA, Rossouw D, Robert-Granié C, Besse P. A sparse pls for variable selection when integrating omics data. Stat Appl Genet Mol Biol. 2008;7(1):35.
21. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B Met. 1995;57: 289–300.
22. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The biomart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015;43(W1):W589-98.
23. Hennessy D, Stelmasiak T, Johnston N, Jackson P, Outch K. Consistent capacity for adrenocortical response to acth administration in pigs. Am J Vet Res. 1988;49(8):1276–83.
24. Eigler N, Saccà L, Sherwin RS. Synergistic interactions of physiologic increments of glucagon, epinephrine, and cortisol in the dog: a model for stress-induced hyperglycemia. J Clin Invest. 1979;63(1):114.
25. Shamoon H, Hendler R, Sherwin RS. Synergistic interactions among antiinsulin hormones in the pathogenesis of stress hyperglycemia in humans. J Clin Endocr Metab. 1981;52(6):1235–41.
26. Peckett AJ, Wright DC, Riddell MC. The effects of glucocorticoids on adipose tissue lipid metabolism. Metabolism. 2011;60(11):1500–10.
27. Flori L, Gao Y, Laloë D, Lemonnier G, Leplat JJ, Teillaud A, et al. Immunity traits in pigs: substantial genetic variation and limited covariation. PLoS One. 2011;6(7):22717.
28. Sutherland M, Rodriguez-Zas S, Ellis M, Salak-Johnson J. Breed and age affect baseline immune traits, cortisol, and performance in growing pigs. J Anim Sci. 2005;83(9):2087–95.
29. Friendship R, Lumsden J, McMillan I, Wilson M. Hematology and biochemistry reference values for ontario swine. Can J Comparat Med. 1984;48(4):390.
30. Wallgren P, Wilén IL, Fossum C. Influence of experimentally induced endogenous production of cortisol on the immune capacity in swine. Vet Immunol Immunop. 1994;42(3):301–16.
31. Salak-Johnson JL, McGlone JJ, Norman RL. In vivo glucocorticoid effects on porcine natural killer cell activity and circulating leukocytes. J Anim Sci. 1996;74:584–92.
32. Dhabhar FS. Stress-induced augmentation of immune function—the role of stress hormones, leukocyte trafficking, and cytokines. Brain Behav Immun. 2002;16(6):785–98.
33. Touma C, Gassen NC, Herrmann L, Cheung-Flynn J, Büll DR, Ionescu IA, et al. Fk506 binding protein 5 shapes stress responsiveness: modulation of neuroendocrine reactivity and coping behavior. Biol Psychiatry. 2011;70(10):928–36.
34. Binder EB. The role of fkbp5, a co-chaperone of the glucocorticoid receptor in the pathogenesis and therapy of affective and anxiety disorders. Psychoneuroendocrinology. 2009;34:186–95.
35. Storer CL, Dickey CA, Galigniana MD, Rein T, Cox MB. Fkbp51 and fkbp52 in signaling and disease. Trends Endocrinol Metab. 2011;22(12):481–90.
36. Cloutier A, Guindi C, Larivée P, Dubois CM, Amrani A, McDonald PP. Inflammatory cytokine production by human neutrophils involves c/ebp transcription factors. J Immunol. 2009;182(1):563–71.
37. Cuesta M, Cermakian N, Boivin DB. Glucocorticoids entrain molecular clock components in human peripheral cells. FASEB J. 2015;29(4): 1360–70.
38. Takahashi S, Yokota S-i, Hara R, Kobayashi T, Akiyama M, Moriya T, et al. Physical and inflammatory stressors elevate circadian clock gene mper1 mrna levels in the paraventricular nucleus of the mouse. Endocrinology. 2001;142(11):4910–7.
39. Hida A, Koike N, Hirose M, Hattori M, Sakaki Y, Tei H. The human and mouse period1 genes: five well-conserved e-boxes additively contribute to the enhancement of mper1 transcription. Genomics. 2000;65(3): 224–33.
40. Yamamoto T, Nakahata Y, Tanaka M, Yoshida M, Soma H, Shinohara K, et al. Acute physical stress elevates mouse period1 mrna expression in mouse peripheral tissues via a glucocorticoid-responsive element. J Biol Chem. 2005;280(51):42036–43.
41. Burioka N, Takata M, Endo M, Miyata M, Takeda K, Chikumi H, et al. Treatment with β2-adrenoceptor agonist in vivo induces human clock gene, per1, mrna expression in peripheral blood. Chronobiol Int. 2007;24(1):183–9.
42. Kumari R, Willing LB, Jefferson LS, Simpson IA, Kimball SR. Redd1 (regulated in development and dna damage response 1) expression in skeletal muscle as a surrogate biomarker of the efficiency of glucocorticoid receptor blockade. Biochem Biophys Res Commun. 2011;412(4):644–7.
43. Britto FA, Begue G, Rossano B, Docquier A, Vernus B, Sar C, et al. Redd1 deletion prevents dexamethasone-induced skeletal muscle atrophy. Am J Physiol Endocrinol Metab. 2014;307(11):983–93.
44. Clark AR, Martins JRS, Tchen CR. Role of dual specificity phosphatases in biological responses to glucocorticoids. J Biol Chem. 2008;283(38): 25765–9.

45. Osterlund CD, Thompson V, Hinds L, Spencer RL. Absence of glucocorticoids augments stress-induced mkp1 mrna expression within the hypothalamic–pituitary–adrenal axis. J Endocrinol. 2014;220(1):1–11.
46. Burke SJ, Goff MR, Updegraff BL, Lu D, Brown PL, Minkin Jr SC, et al. Regulation of the ccl2 gene in pancreatic β-cells by il-1β and glucocorticoids: role of journal=mkp-1, PLoS ONE. 2012;7(10):e46986.
47. Oster M, Muráni E, Ponsuksili S, Richard B, Turner SP, Evans G, et al. Transcriptional responses of pbmc in psychosocially stressed animals indicate an alerting of the immune system in female but not in castrated male pigs. BMC Genom. 2014;15(1):967.
48. Salaria S, Chana G, Caldara F, Feltrin E, Altieri M, Faggioni F, et al. Microarray analysis of cultured human brain aggregates following cortisol exposure: Implications for cellular functions relevant to mood disorders. Neurobiol Dis. 2006;23(3):630–6.
49. Reddy TE, Gertz J, Crawford GE, Garabedian MJ, Myers RM. The hypersensitive glucocorticoid response specifically regulates period 1 and expression of circadian genes. Mol Cell Biol. 2012;32(18):3756–67.
50. Uhlenhaut NH, Barish GD, Ruth TY, Downes M, Karunasiri M, Liddle C, et al. Insights into negative regulation by the glucocorticoid receptor from genome-wide profiling of inflammatory cistromes. Mol Cell. 2013;49(1):158–71.
51. So A, Chaivorapol C, Bolton EC, Li H, Yamamoto KR. Determinants of cell-and gene-specific transcriptional regulation by the glucocorticoid receptor. PLoS Genet. 2007;3(6):94–4.
52. Mills E, O'Neill LA. Succinate: a metabolic signal in inflammation. Trends Cell Biol. 2014;24(5):313–20.
53. Donner KM, Hiltunen TP, Jänne OA, Sane T, Kontula K. Generalized glucocorticoid resistance caused by a novel two-nucleotide deletion in the hormone-binding domain of the glucocorticoid receptor gene nr3c1. Eur J Endocrinol. 2013;168(1):9–18.
54. Donn R, Berry A, Stevens A, Farrow S, Betts J, Stevens R, et al. Use of gene expression profiling to identify a novel glucocorticoid sensitivity determining gene, bmprii. FASEB J. 2007;21(2):402–14.
55. Menke A, Arloth J, Pütz B, Weber P, Klengel T, Mehta D, et al. Dexamethasone stimulated gene expression in peripheral blood is a sensitive marker for glucocorticoid receptor resistance in depressed patients. Neuropsychopharmacology. 2012;37(6):1455–64.
56. Kamisoglu K, Sleight K, Nguyen TT, Calvano SE, Coyle SM, Corbett SA, et al. Effects of coupled dose and rhythm manipulation of plasma cortisol levels on leukocyte transcriptional response to endotoxin challenge in humans. Innate Immun. 2014;20(7):774–84.
57. Ponsuksili S, Du Y, Murani E, Schwerin M, Wimmers K. Elucidating molecular networks that either affect or respond to plasma cortisol concentration in target tissues of liver and muscle. Genetics. 2012;192(3):1109–22.
58. Polman JAE, Welten JE, Bosch DS, de Jonge RT, Balog J, van der Maarel SM, et al. A genome-wide signature of glucocorticoid receptor binding in neuronal pc12 cells. BMC Neurosci. 2012;13(1):118.
59. Cole SW. Elevating the perspective on human stress genomics. Psychoneuroendocrinology. 2010;35(7):955–62.
60. O'Donovan A, Sun B, Cole S, Rempel H, Lenoci M, Pulliam L, et al. Transcriptional control of monocyte gene expression in post-traumatic stress disorder. Dis Markers. 2011;30(2-3):123–32.
61. Lansade L, Valenchon M, Foury A, Neveux C, Cole SW, Layé S, et al. Behavioral and transcriptomic fingerprints of an enriched environment in horses (equus caballus). PloS one. 2014;9(12):114384.
62. Chaussabel D, Pascual V, Banchereau J. Assessing the human immune system through blood transcriptomics. BMC Biol. 2010;8(1):84.

2.3 Article 2 - Time course study of the response to LPS targeting the pig immune response gene networks

RESEARCH

Time course study of the response to LPS targeting the pig immune response gene networks

Elena Terenina[1*], Valérie Sautron[1], Caroline Ydier[1], Darya Bazovkina[2], Amelie Sevin[1], Laure Gress[1], Yannick Lippi[3], Yvon Billon[4], Laurence Liaubet[1], Pierre Mormède[1] and Nathalie Villa-Vialaneix[5]

*Correspondence:
elena.terenina@toulouse.inra.fr
[1]INRA, UMR 1388 GenPhySE, Université de Toulouse, INRA, INPT, ENVT, F-31326 Castanet-Tolosan, France
Full list of author information is available at the end of the article

Abstract

À FAIRE

Keywords: Stress; Hypothalamic-pituitary-adrenal (HPA) axis; Cortisol; Time-course; Systems biology; Microarray; Pig

Background

Over time, farms have evolved towards factory production units. This has led to a decline of the welfare of animals that becomes an important concern for consumers [1]. Moreover, this type of farming has led to the selection of animals with high production traits such as rapid growth, lean meat, or large litters. However, the strong selection focus on these characteristics is suspected to reduce functional traits, such as viability of the new borns or disease resistance. Consequently, the genetic potential of animals is usually not fully expressed in commercial conditions, due to the limiting influence of the environment. Robustness is a specific quality of an individual to express a high production potential in a wide variety of environmental conditions and is now a major specific breeding goal in the context of sustainable farm animal breeding. Various strategies are available to increase robustness, and we have suggested that the reinforcement of the neuroendocrine stress responses may favour the processes of adaptation and dampen the negative consequences of the environment [2]. The hypothalamic-pituitary-adrenocortical (HPA) axis is the main neuroendocrine system involved in adaptation to stress and is strongly influenced by genetic factors [3]. It is therefore a primary candidate for the selection of more robust animals [2].

In modern intensive livestock production, pigs are easily threatened by different types of inflammation. Immunological stress is a comprehensive process involving immunological, neurological, and endocrinological responses [4]. The reciprocal "subjugation" of the brain and the immune system via cytokines and stress hormones is now well demonstrated [5, 6]. The resulting balance has more recently been demonstrated at the level of blood cell transcriptome [7], with chronic stress increasing the expression of genes regulated by inflammatory mediators and decreasing those regulated by glucocorticoid hormones [8]. This approach has been used to evaluate the negative consequences of adverse environmental conditions, mostly in humans but also in farm animals (horses [9]). More recently, individual differences have also been described as related to personality dimensions in humans [10].

However the relationships with individual variations of HPA axis activity, including genetic factors, is still unexplored.

We have shown previously large variations in biological and transcriptomic responses to an ACTH stimulation test [11]. The present study aims at describing blood transcriptomic, hormonal, and metabolic responses of pigs to a systemic challenge using lipopolysaccharide (LPS), a major component of the outer membrane in gram-negative bacteria [12]. LPS provokes an acute inflammatory syndrome resulting eventually in all kinds of pathophysiological damages [13]. The objective is to analyse the individual variation of the biological responses in relation to the activity of the HPA axis measured by the level of cortisol released by LPS and by an ACTH stimulation test [14].

Animals and methods

Animals, treatment and blood sampling

All animal use was performed under European Union and French legislation (directive 201063UE, décret 2013-118). The protocol and procedures were approved by the local (Poitou-Charentes) ethics committee (decision CE2013-1, 21012013). Experimental animals were 120 piglets (63 females and 57 males) randomly selected from 28 litters (4-5 animals per litter) of purebred Large White pigs and produced in 3 successive batches raised 3 weeks apart. They were weaned at 4 weeks and animals from 2-3 litters were mixed at weaning in each post-weaning pen. Experimental animals were not isolated from their littermates. At 8 weeks, each animal was injected in the neck muscles with LPS (*E. coli* serotype 055:B5, Sigma-Aldrich, Saint Quentin Fallavier, FR) at a dose of 15 μg/kg body weight. Injections occurred from 10:00-11:00 AM to avoid nycthemeral variations. Blood samples were collected before the injection ($t = 0$) and 1 hour ($t = +1$), 4 hours ($t = +4$) and 24 hours ($t = +24$) after injection. At each time, animals were slightly restrained on their back in such a way that the effect on their stress level can be regarded as insignificant. Two blood samples were then taken by puncture of a jugular vein in Vacutainer ® tubes with 20 G needles (Becton-Dickinson, le Pont de Claix, FR). Tympanic temperature was measured at the same time with a digital thermometer (Thermoscan PRO 4000, Braun Welch Allyn, Hechingen, GE). The whole handling procedure lasted less than 30 sec. One 10 ml tube with lithium heparin was used for chemical biology. After centrifugation (2355 g, 10 min), plasma aliquots were frozen at -80 °C until analysis. One 5 ml tube with EDTA (di-potassium salt) was used for blood cell count and an aliquot (400 μl) was mixed with the same volume of DL buffer (Macherey-Nagel, Hoerdt, FR), frozen at -20 °C for 4 h and then at -80 °C until analysis for gene expression.

Biological analyses

Cortisol was measured by direct automated immunoassay (AIA-1800, Tosoh Bioscience, San Francisco, CA). Glucose and free fatty acid (FFA), were measured by colorimetry with an ABX Pentra 400 clinical chemistry analyzer from Horiba Medical (Grabels, FR). Blood cell counts were measured with a MS-9-5 hematology analyzer from Melet Schloesing Laboratories (Osny, FR), calibrated for pig blood by the manufacturer. Blood cell count variables included: white cells count, proportion of lymphocytes, monocytes and granulocytes, red cells count, hematocrit,

concentration of hemoglobin, red cells width and volume, concentration of platelets and platelets width and volume. Thus, the biological variables contained 15 variables measured on the 120 pigs. In addition, birth and weaning weights were also measured for each pig. Outlying observations were visually identified and treated as missing data. Missing data were imputed using a k-nearest neighbor imputation ($k = 5$; R package **DMwR** [15]). To ensure normality, cortisol, platelet and white cell counts were $\log_{10}$ transformed and FFA was transformed using the square root. Batch effects were removed by aligning the within-batch medians for all measurements.

RNA extraction and whole blood analysis

A total RNA isolation and purification was done according to the manufacturer's instructions using the Nucleospin RNA Blood kit (Macherey-Nagel, FR) followed by DNase treatment. The quality of each RNA sample was checked through the Bioanalyser Agilent 2100 (Agilent Technologies, Massy, FR) and low-quality RNA preparations were discarded (RIN < 8).

Microarray description

A porcine microarray GPL16524 (Agilent, 8×60 K) was used to hybridize the RNA samples as described previously in [11]. This microarray contained 61,625 spots. Among them, 308 were negative controls and 49 were used for alignment. One probe was duplicated twice on each array. Thus the microarray contained 60,305 unique porcine probes. After quality control, quantile normalization and filtering, 27,837 transcripts were found to be expressed in blood in our experimental conditions.

Hybridization protocol

Blood samples from 30 female pigs from only 2 batches were used. Blood samples of 2 pigs at one time step each were of poor quality and thus not used. Each of the 15 used arrays contained 8 microarrays which corresponded to the 4 observations of two individuals, each from one batch. This design secured the kinetics of the response for each individual and prevented confounding effects between batch and array. After quality control and filtering, 27,837 probes were kept and $\log_2$ transformed. Technical biases were handled by aligning the within-array medians for all genes and by a quantile normalization within animal (function **normalize.quantiles** in the R package **preprocessCore** [16]). Missing data were imputed using a k-nearest neighbor imputation using the same method as described in Section Biological analyses.

Statistical analyses

All analyses were performed with the R software, version 3.2.2 [17].

Statistical analysis of plasma metabolites and cortisol

First, all variables were subjected to a one-way ANOVA with repeated measures. In order to control the false discovery rate (FDR) [18], p-values were adjusted using a Benjamini-Hochberg (BH) approach (Table 1). Variables with an adjusted p-value (FDR < 0.05) were then subjected to 3 paired t-tests to assess the difference between $t = 0$ and $t = +1$, between $t = 0$ and $t = +4$ and between $t = 0$ and $t = +24$. The full list of p-values was adjusted using a BH approach (Figure 1).

In addition, the influence of sex on the biological variables was tested using a two-way ANOVA with repeated measures including sex as a variable. p-values were adjusted using a BH approach.

Cortisol levels at $t = +1$ is the most relevant measure to assess the sensitivity of the adrenals to ACTH (data from [11]). Hence, correlations between biological variables at $t \in \{0, +1, +4, +24\}$ and the level of cortisol at $t = +1$ were investigated using paired t-tests. p-values were adjusted using a BH approach.

Statistical analysis of the transcriptome

Differentially expressed probes (DEP)

The whole blood is composed of different types of white cells with distinct roles which express different kinds of transcripts [19]. It is thus likely that a modification in blood cell composition may influence the gene expression level without having cells actually express transcripts differently. As blood cell composition was found to vary over time after LPS injection, we used the $\frac{\text{Lymphocyte}}{\text{Granulocyte}}$ (L/G) ratio as a covariate in our analyses.

Three different approaches were used to identify relevant probes:

Firstly, we identified probes differentially expressed at each time step while taking blood cell composition into account. Blood cell composition was measured by the L/G ratio. Three models (one for each time step t' where $t' \in \{+1, +4, +24\}$) were fitted to each probe using observations at $t = 0$ and $t = t'$.

$$\text{expr}_{it} = \mu_0 + \tau_{t'}\mathbb{I}_{\{t=t'\}} + \beta^{t'} L/G_{it} + \epsilon_{it} \quad (1)$$

with $i = 1, \ldots n$ is animal i. expr_{it} is the expression of the probe being studied for animal i at time step t ($t \in \{0, t'\}$), μ_0 is the specific contribution of time step $t = 0$, $\tau_{t'}$ is the effect of time step t', $\beta^{t'}$ is the effect of L/G ratio in this model and $\epsilon_{it} \sim N(0, \sigma_e^2)$ is an error term.

We then tested the contribution of time step t' against the null hypothesis H_0: $\tau_{t'} = 0$. The full list of p-values was globally adjusted using a Bonferroni approach. As the Bonferroni approach exerts a more stringent control than the BH approach, it was used to obtain a narrowed list of the most significant probes. Probes with at least one adjusted p-value < 0.01 were probes for which the expression adjusted by the L/G ratio was significantly different from the basal level. In the sequel, this list of genes will be referred to as **(M3)**.

Secondly, we identified probes for which the L/G ratio effect is different according to the time step. To that aim, we compared a complete model, including all time step contributions and the L/G ratio effect according to the time step (Equation (2)):

$$\text{expr}_{it} = \tau_t + \beta_t L/G_{it} + \epsilon_{it} \quad (2)$$

(with $t \in \{0, 1, 4, 24\}$ and β_t is the interaction effect between time step t and the L/G ratio of individual i at time step t), to a reduced model, including only the average L/G ratio and all time step contributions (Equation (3)):

$$\text{expr}_{it} = \tau_t + \beta L/G_{it} + \epsilon_{it}. \quad (3)$$

An F-test was then performed to test the null hypothesis, H_0: $\beta_0 = \beta_1 = \beta_4 = \beta_{24}$, against the alternate hypothesis, H_1: $\exists t_1, t_2$ such as $\beta_{t_1} \neq \beta_{t_2}$. Multiple testing was handled by applying a BH approach (FDR < 0.05). Probes for which the test was significant were probes for which the effect of L/G varied over time. In the sequel, this list of genes will be referred to as **(M1)**.

Finally, we studied correlations between all probes and cortisol level when it reaches its peak in blood circulation after LPS injection. Thus, Pearson correlations, ρ, were computed between DEP expression at each time step and cortisol level at $t = +4$. A correlation test was then performed to test the null hypothesis, H_0: $\rho = 0$ against H_1: $\rho \neq 0$. Multiple testing was handled by using a BH approach (FDR $< 5\%$). This list of genes will be referred as **(M2)** in the sequel.

In addition, to link probes responding to a LPS injection with a measure of the HPA axis activity, we studied correlations between all probes expressions and the cortisol level at $t = +1$ after ACTH injection, as measured on the same pigs in [11].

All lists of DE probes were then annotated and duplicated probes were removed by keeping only DEP with the smallest FDR per annotated gene and all non-annotated genes. Remaining genes will be referred as differentially expressed genes (DEG) in the sequel.

Sequence annotation

Each cDNA sequence was compared to Refseq_rna mammalian database using the NCBI blastn program (http://blast.ncbi.nlm.nih.gov/Blat.cgi). Resulting hits were sorted out according to their closeness to the pig genome, their coverage and sequence identity. The selected cDNA sequences were submitted to the HUGO (Human Genome Organization) gene nomenclature committee, using their RefSeq IDs (http://genenames.org). Then, HUGO gene symbols or official gene symbols were used as gene names.

Pathway analysis

A pathway is an interconnected arrangement of processes, representing the functional roles of genes in the genome. Functional integration of gene expression, *i.e.*, identification of gene networks, was performed using the "Gene Ontology" database AmiGO (http://amigo.geneontology.org). The significantly up- or down-regulated genes for each list as well as for genes common to all lists could be assembled into networks using "Ingenuity Pathway Analysis" (http://ingenuity.com) under licence. This application provides computational algorithms to identify the enriched biological pathways, functions and biological mechanisms of selected genes and also proposes enriched regulators as transcription factors.

A regulatory network could be constructed with the information provided by the option "Upstream Regulator". This option proposes a list of regulators known to have a significant effect on some of the targeted genes in the input list. Ingenuity also provides computational algorithms to identify and to dynamically generate significant biological networks. Networks are ranked by a score that takes into account the number of targeted genes and the size of the networks. This score ($-\log_{10}(p\text{-value})$) is the probability for a group of genes to be observed in the same network by chance. We chose networks displaying direct relationships between genes. Path Designer (an

Ingenuity tool) was used to improve the readability of the networks. Nodes added by Ingenuity were discarded when they were not necessary to connect our genes of interest and the resulting network was merged with the regulatory network.

Functional analysis of **(M3)** list

As **(M3)** drew a large list of DEG, genes of this list was first subjected to a functional enrichment using **biomaRt** [20], a bioconductor package that allows accessing and retrieving Ensembl data from the R software and topGo [21], a bioconductor[1] package that allows enrichment testing. The statistical enrichment of a particular biological process was tested with a Fisher's exact test, using the list of genes expressed on the microarray as the reference list of genes. Resulting p-values were adjusted for multiple tests using a BH approach. Functions with a minimum of 10 genes per gene ontology and a FDR < 0.01 were considered to be enriched biological functions. Genes enriching generic functions (such as morphogenesis, transcription, locomotion and others) were removed. The remaining genes were then subjected to a hierarchical ascending classification (HAC) using the Ward method with a distance based on the correlations between genes. This allowed for the identification of clusters of genes having a similar pattern of evolution over time.

Time course analyses

In the case of time course analyses, the approach previously described (applying a univariate linear model on each variable followed by multiple test correction) is common. However, this approach disregards the dependancies between genes and does not allow for a global view of the relationships between the repeated measurements in high dimensional data. A multilevel approach was thus used to investigate the relationships between the repeated measurements while taking advantage of multivariate approaches [22].

The multilevel approach, as described by Liquet et al. [22], uses a split-up variation inspired by the mixed-model framework.

Let $X = (x_{it}^k)_{i=1,\ldots,n,\, t\in\{0,+1,+4,+24\},\, k=1,\ldots,p}$ be the $(N \times p)$ observation matrix (clinical biology variables or gene expressions) on n animals with 4 times of measurements ($N = n \times 4$). X can be split up as:

$$X = \underbrace{X_{\cdot\cdot}}_{\text{offset term}} + \underbrace{X_b}_{\text{between-animal deviation}} + \underbrace{X_w}_{\text{within-animal deviation}} \qquad (4)$$

The matrix $X_{\cdot\cdot}$ represents the offset term defined as $1_N\, x_{\cdot\cdot}^T$ where $(x_{\cdot\cdot}^k = \frac{1}{N} \sum_{t\in\{0,+1,+4,+24\}} \sum_{i=1}^{n} x_{it}^k)$. 1_N is a $(N \times 1)$ matrix containing ones and $x_{\cdot\cdot}^T = (x_{\cdot\cdot}^1, \ldots, x_{\cdot\cdot}^p)$. X_b is the between-animal matrix of size $(N \times p)$ defined by concatenating $1_4\, x_{bi}^T$ for each animal into X_b with $x_{bi}^T = (x_{i\cdot}^1 - x_{\cdot\cdot}^1, \ldots, x_{i\cdot}^p - x_{\cdot\cdot}^p)$ $(x_{i\cdot}^k = \frac{1}{4} \sum_{t\in\{0,+1,+4,+24\}} x_{it}^k)$. $X_w = X - X_{i\cdot}$ is the within-animal deviation matrix of size $(N \times p)$ with $X_{i\cdot}$ the matrix defined by concatenating the matrices $1_4\, x_{i\cdot}^T$ for every animal into $X_{i\cdot}$, with $x_{i\cdot}^T = (x_{i\cdot}^1, \ldots, x_{i\cdot}^p)$.

[1] http://www.bioconductor.org

By splitting the different parts of the variation in the data while taking into account the repeated measurements on the subjects, the multilevel step allows to study the effect of different conditions within a subject separately from the variation between subjects. This method is especially relevant when a high between-subject variability is observed in repeated data: multivariate approaches were then performed on X_w to bring out the most relevant correlations between variables in the dataset, independently from individual variations. First a multilevel PCA was performed on the biological variables to study the overall effect of LPS on plasma metabolites and cortisol over time. Then, a multilevel multiple factor analysis (MFA) [23] was used to investigate the overall relationships between clinical biology and transcriptomic data.

Results and discussion

Plasma cortisol, metabolites, and blood cell counts

Baseline values of biological variables and the global time effect, and birth and weaning weights are shown in Table 1. Figure 1 shows the evolution of the main variables over time.

Table 1 Reference values (at $t = 0$) for the biological variables, birth weight and weaning weight ($n = 120$). Results of the ANOVA for time effect (F and significativity (sig)). *: FDR < 0.01; **: FDR < 0.001; ***: FDR $< 10^{-12}$.

	units	min	max	mean	SEM	F	sig
Tympanic temperature	°C	36.100	40.257	39.168	0.050	258.110	***
White cells	$\log_{10}$(G/l)	0.491	1.472	1.188	0.011	572.970	***
Lymphocytes	%	46.600	91.900	67.477	0.555	112.180	***
Monocytes	%	3.900	16.200	8.557	0.191	69.500	***
Granulocytes	%	2.500	35.600	22.608	0.527	78.210	***
L/G ratio		1.355	36.760	3.466	0.295	64.650	***
Red cells	T/l	1.490	7.330	5.163	0.054	69.420	***
Mean corpuscular volume	fl	39.700	63.700	52.008	0.333	62.120	***
Hematocrit	%	6.800	37.400	26.828	0.299	78.990	***
Hemoglobin	g/dl	6.900	12.800	8.947	0.092	46.680	***
Red blood cells distribution width	fl	29.100	33.800	32.029	0.081	74.440	***
Platelets	$\log_{10}$(G/l)	2.330	2.998	2.667	0.011	227.400	***
Mean platelet volume	fl	7.600	13.000	9.682	0.102	71.210	***
Platelet distribution width	%	9.600	12.000	10.771	0.045	122.790	***
Cortisol	$\log_{10}$(ng/ml)	1.041	2.033	1.475	0.017	370.240	***
Free fatty acids	$\sqrt{(\text{mmol/l})}$	0.079	0.560	0.162	0.005	111.040	***
Glucose	mmol/l	5.850	9.525	8.035	0.061	123.990	***
Bilirubin	μmol/l	4.660	13.000	8.523	0.190	178.610	***
Birth weight	kg	0.400	2.680	1.492	0.033		
Weaning weight	kg	5.460	16.564	9.486	0.174		

In pigs like in other species, LPS is responsible for the fever and inflammatory reaction induced by gram-negative bacterial infection, as shown by the increase in the circulating levels of pro-inflammatory cytokines and acute phase proteins, as well as the changes in white blood cell counts [24–27]. Tympanic temperature peaked at $t = +4$ (40.8°C vs 39.1°C) and returned to basal levels at $t = +24$. The characteristic changes of white blood cell count to LPS were observed, with a decrease of total count, maximal at $t = +4$ (5.70 vs 15.35 G/l) and the mirror changes in the respective proportions of lymphocytes and granulocytes. This indicated that the lymphocytes/granulocytes ratio (L/G) was a good measure to use in order to take into account these changes that result mainly from the redistribution of lymphocytes into the tissues [25]. The L/G ratio was maximal at $t = +1$ (9.32 vs 3.67) and back to basal levels at $t = +4$. The red blood cell count and associated measures

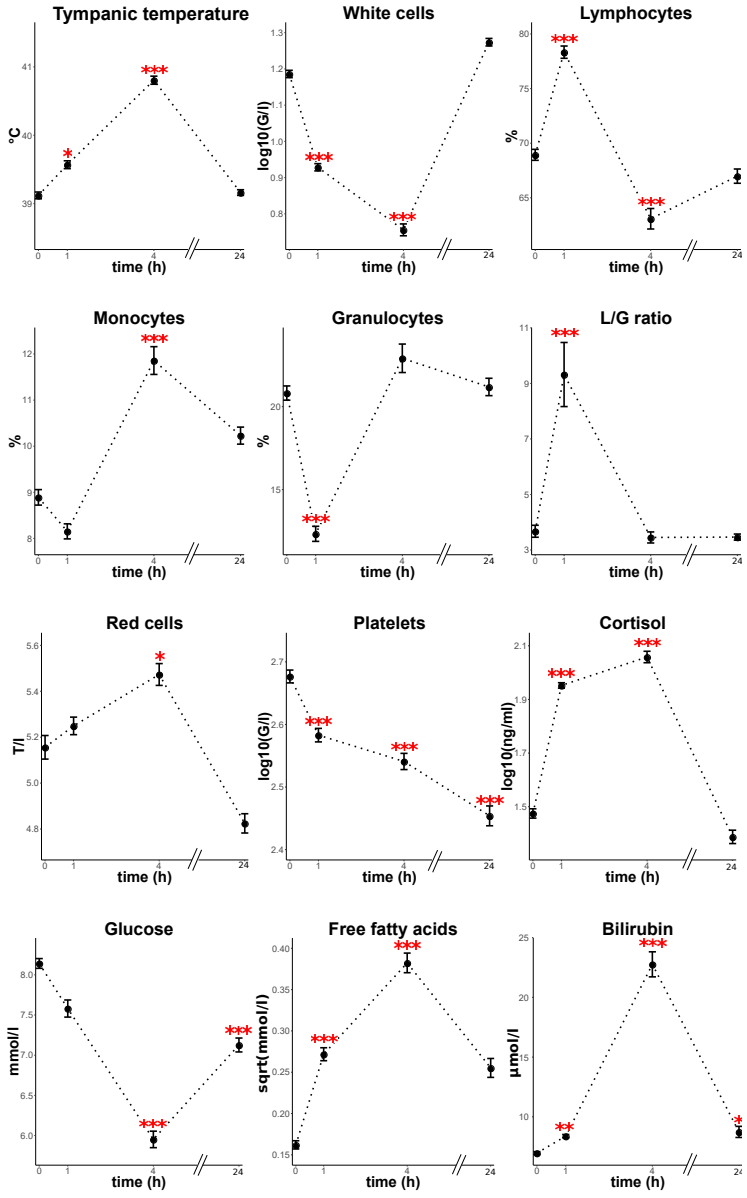


Figure 1 Mean evolution of the biological variables over time. Vertical bars correspond to + and - SEM at each point.

(hematocrit and hemoglobin concentration) showed a biphasic change, with an initial increase, maximal at $t = +4$ (5.47 vs 5.16 T/l) and a subsequent long-lasting decrease (4.82 T/l at $t = +24$). The platelet count showed a steady decrease until at least $t = +24$ (284 vs 475 G/l). These measures were not influenced by sex, except the mean red cell volume and hematocrit that were slightly lower in males (FDR $<$ 0.05).

LPS also induces profound endocrine and metabolic changes and our results are consistent with previously published data in pigs [24, 25, 27]. A large increase in circulating levels of cortisol (and catecholamines, not measured here) has been described and these hormonal changes can be involved in the release of the mediators of inflammation [25, 27]. Cortisol levels peaked at $t = +4$ with a 3.83-fold increase (114.3 vs 29.8 ng/ml). Circulating glucose levels were reduced by 26.9% to 5.95 mmol/l at $t = +4$. It was shown previously in mice that this hypoglycaemia cannot be explained by changes in insulin concentrations that are also reduced by LPS [28], but it could result from the increased glycolysis in muscles and immune cells, as well as from a reduced hepatic glucose production [29]. The circulating concentration of free fatty acids increased from 0.026 to 0.146 mmol/l at $t = +4$. This can result from the lipolytic action of catecholamines and cortisol that are massively released by LPS [11, 30] and from LPS-induced changes in hepatic and fat tissue lipid metabolism [31, 32]. A sharp increase in bilirubin concentrations was also measured at $t = +4$ (17.72 vs 2.14 μmol/l), reflecting the hepatic toxicity of LPS [33, 34]. None of these biochemical measures was influenced by sex.

Overall effect of LPS on clinical biological variables

The overall effect of LPS over time was investigated with a multilevel PCA (Figure 2). The first component of the multilevel PCA opposes the observations at $t = 0$ (negative coordinates on this axis) to the observations at $t = +4$ (positive coordinates on this axis), this time step corresponding to the peak of LPS effect. The second component opposes the observations at $t = +24$ (positive coordinates on this axis) to the other observation times (negative coordinates on this axis). The representation of the variables shows that the first axis is mainly driven by an opposition between free fatty acids (FFA), bilirubin, temperature and cortisol (high measures at $t = +4$), and white cell count and glucose (low measures at $t = +4$). The second axis is mainly driven by L/G ratio and platelet count that are low at $t = +1$.

No biological variable was found to be correlated to cortisol level at $t = +1$ after ACTH injection that measures individual differences in HPA axis activity [14].

Differentially expressed genes related to key immune functions

In our study, we used a comprehensive gene expression profiling by means of microarray analysis to identify clusters of genes differentially expressed in peripheral blood cells, taking into consideration the kinetic of the response with 4 time points ($t \in \{0, +1, +4, +24\}$). LPS induces dramatic changes in blood cell number and lymphocyte/granulocyte (L/G) ratio that introduces a confusion between time and cell type effects, and a major challenge for the interpretation of transcriptomic data. Therefore we based the interpretation of the results on three different lists of genes, **(M1)**, **(M2)**, and **(M3)**.

Analysis of each list of genes

The **first list of genes (M1)**, consists of 154 unique genes (209 transcripts, Additional file 1 — List of 154 unique genes differentially expressed in list **(M1)**) for which the contribution of the L/G ratio to the expression varied over time steps.

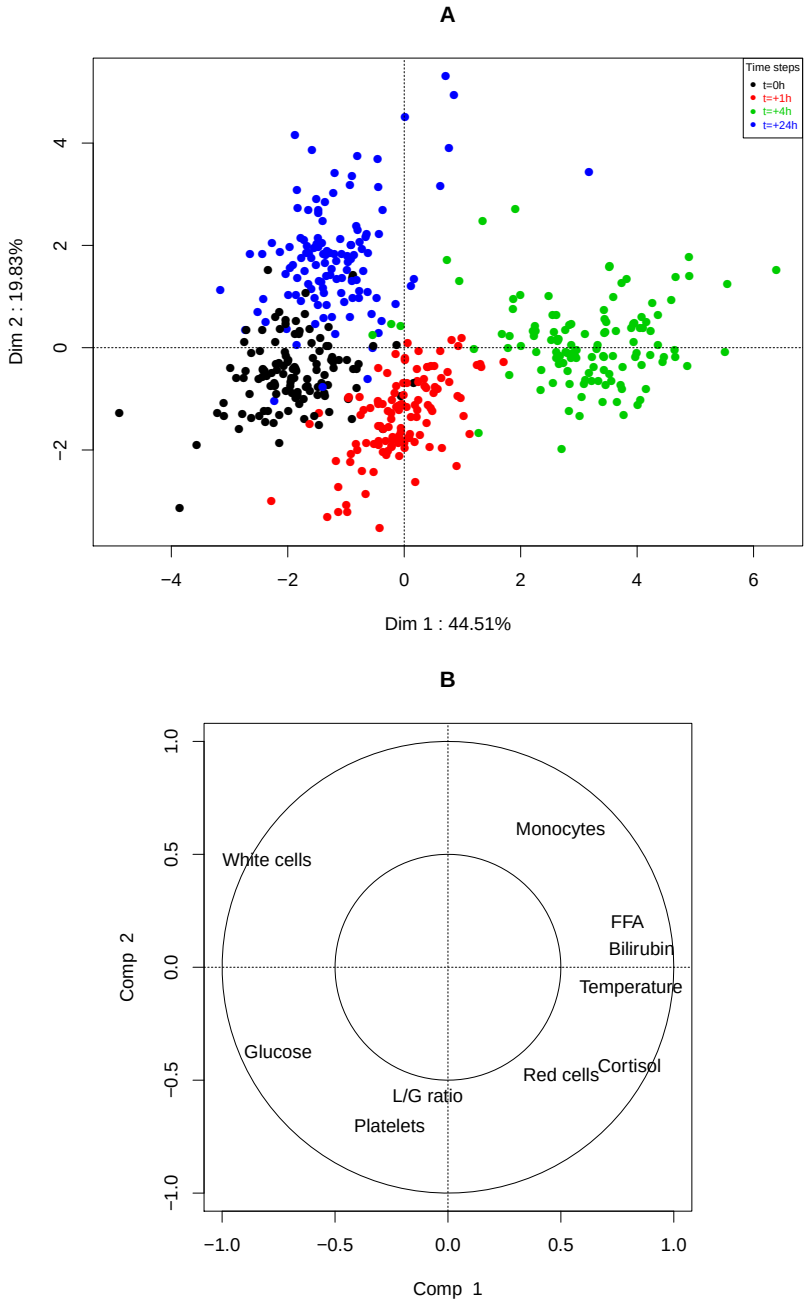


Figure 2 Multilevel PCA on the biological variables responding to LPS. Colors represent the time step; Black: $t = 0$; Red: $t = +1$; Green: $t = +4$; Blue: $t = +24$; **A**: Projection of the individuals on dimensions 1 - 2; **B**: Projection of the variables on dimensions 1 - 2.

Among these genes, 132 genes were further assembled into six functional networks that notably revealed hematological system development and function, tissue morphology, cancer, organismal injury and abnormalities, reproductive system disease, cellular growth and proliferation. It is important to note that none of these genes

is related to immunity and inflammation. The average evolution of all these genes shows the same expression profile. This group of genes is characterized by genes decreasing with a peak of expression at $t = +1$ and stable between $t = +4$ and $t = +24$. According to this analysis, it is unlikely that genes for which the contribution of the L/G ratio to the expression varied over time steps are directly involved in the immune response to LPS injection.

The **second list of genes (M2)** consists of 116 unique genes (185 transcripts, Additional file 2 — List of 116 unique genes differentially expressed in list **(M2)**) for which the expression was found to be correlated to the level of cortisol at $t = +4$. This time point was chosen as the peak of plasma cortisol concentration after LPS. The most significant functions are: cellular function and maintenance – function of blood cells (32 genes); cellular movement, immune cell trafficking – leucocyte migration (36 genes); lymphoid tissue structure and development, tissue morphology – quantity of lymphatic system cells (34 genes); cellular function and maintenance – function of leucocytes (29 genes); hematological system development and function, tissue morphology – quantity of leucocytes (36 genes); cellular movement, hematological system development and function, immune cell trafficking – cell movement of leucocytes (33 genes); immunological disease – systemic autoimmune syndrome (37 genes) (Figure 3, Additional file 3 — Biological functions enriched by differentially expressed genes in list **(M2)** ($n = 30$). '.xls' file).

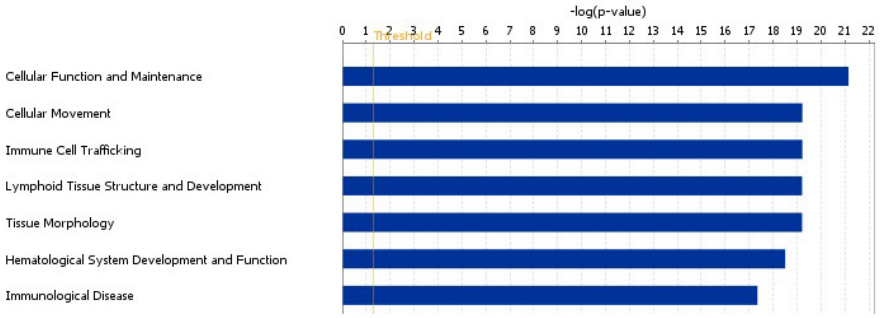


Figure 3 Biological processes significantly enriched by the 116 DEG in **(M2)**.

The **third list of genes (M3)** consists of 9,530 unique genes (22,794 transcripts, Additional file 4 — List of 9,530 unique genes differentially expressed in list **(M3)**) for which the expression adjusted by the L/G ratio was significantly different from the basal level. **(M3)** was submitted to gene ontology and enrichment analysis. These analyses showed 106 classes significant at FDR < 0.05. Due to the important number of DEG, generic classes were removed (such as morphogenesis, transcription, locomotion and others). In each group, we chose a number of representative genes giving two hundred eighty-four genes that were grouped into 6 functional classes:

The "immunity and inflammation" class (175 genes) is related to the inflammatory cascade after activation of leukocytes by LPS via TLR4 receptor (a receptor for bacterial lipopolysaccharide). TLR4 is a critical driver of immune responses to bacterial infections. Signals from TLR4 promote NF-κB and AP-1 activation, leading to inflammatory gene expression [35] (DEG for TLR4, TNF, JUNB, and NF-B pathway).

The "chemotaxis" class is composed of 59 genes. Among them *ABHD2*, *ACADS*, *AIF1*, *ANXA7*, *ARPC1A*, *ARPC2*, *CD97*, *CHL1*, *CLIC1*, *CNTFR*, *COQ3*, *DGKD*, *DNASE2*, *GP1BA*, *GPI*, *HCLS1*, *HPS6*, *IL1RN*, *IL8RA*, *KAT5*, *LOC100523056*, *LSP1*, *MAN2B1*, *PARK7*, *PTPN6*, *SMAD7*, *SPG21*, *TMEM173*, *TMSB10*, *TMSB4X*, *TRDMT1*, and *TSPO* genes are related to immune cell trafficking. This observation is in agreement with the observed blood cell redistribution.

The "apoptosis" class (33 genes) includes *C5AR1*, *CCL24*, *CCR1*, *CCR3*, *CXCL13*, *IRG1*, *ALDOC*, *C3AR1*, *CADM1*, *CAPN3*, *HEXA*, *ID3*, *MAEA*, *PLAU*, *PRDX5*, *PROC*, and *CXCR2* genes related to apoptosis and inflammatory response, and *TNFSF13B* and *NFKBIA* involved in cell-activating factor signalling pathway.

Twelve genes (*CD9*, *ANXA5*, *COMT*, *DDIT3*, *ADAM10*, *BAD*, *SOD2*, *ADRB2*, *CLN8*, *LTA*, *TGFBR1* and *PTEN*) form a "calcium ion transport" class.

The "metabolism" class includes four genes (*EDN1*, *COFILIN*, *PLA2G4A*, and *CORO1A*), and the "hormonal responses" class includes one gene (*HMOX1*).

Clustering of differentially expressed genes **(M3)**

The 279 remaining genes from this third list of 284 genes were finally grouped into 4 clusters (Figure 4, Additional file 5 — List of 284 unique genes differentially expressed in list (M3) included in non-generic biological functions ($n = 30$). 'xlsx' file) according to the kinetic of their response.

The first cluster includes 10 genes up-regulated at $t = +1$ and related to immune cell trafficking (*CXCL5, CCL4, LTA, CCL20, CXCL2, EDN1, NFKBIA, JUNB, TNFAIP3, ALOX12*). In this network, JUNB (JunB proto-oncogene, AP-1 transcription factor subunit) is in the central position together with *NFKBIA* and chemokines (*CXCL5, CCL4, CCL20, CXCL2*). Inflammation is a powerful protective mechanism which is coordinated and controlled by cytokines and chemokines and, as expected, we detected an increase in the expression level of members of the CXCL family. AP-1 also participates in the immune response; it is activated by the TLR signalling pathway [36] and can induce expression of interleukins [37–39]. The AP-1 family is a family of Jun (C-Jun, JunB, and JunD) homodimers and Jun heterodimers with Fos (c-Fos, FosB, Fra-1, and Fra-2) [39, 40]. Hormone activation of the glucocorticoid receptor in leukocytes results in a profound suppression of pro-inflammatory gene networks such as the NF-κB mediated transcription of pro-inflammatory cytokine genes and *CCL4, CXCL2, LTA* were described by [41] as glucocorticoid-regulated genes. These findings show that wide variation in glucocorticoid sensitivity exists between individuals which may influence susceptibility to inflammatory diseases [11].

The second cluster includes 18 genes up-regulated at $t = +4$ and related to connective tissue disorders and inflammatory diseases. Key genes are *C3, C3AR1, CXCL10, CXCL13, FAS, ICAM1, IL1RN, MMP13, RETN, SOD2, TLR4*, and *TNFAIP6*. Toll-like receptor 4 (*TLR4*) is essential for initiating the innate response to lipopolysaccharide from Gram-negative bacteria by acting as a signal-transducing receptor. As the pig industry faces a unique array of related pathogens, it is anticipated that the genotype of swine *TLR4* could be of crucial importance in future strategies aimed at improving genetic resistance to infectious diseases [42].

The third cluster includes the genes down-regulated at $t = +4$. This network groups 164 DEG related to the inflammatory response. This cluster is associated

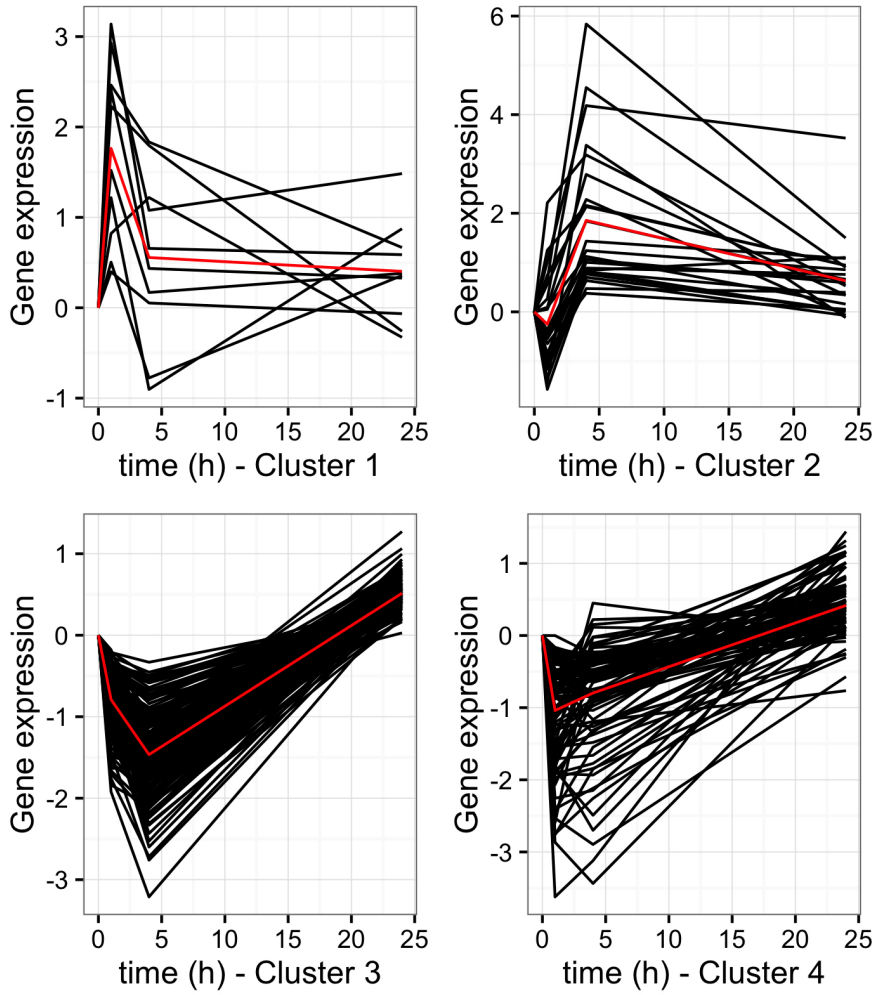


Figure 4 Black: Average evolution the genes in each of the clusters identified by the HAC on the 284 DEG identified in the third list. Evolution of each gene is translated so that it is equal to 0 at $t = 0$; Red: Average evolution over all genes in the cluster (cluster 1: 10 genes, cluster 2: 18 genes, cluster 3: 164 genes, cluster 4: 92 genes).

with functions linked to immunological disease, cancer, cell death and survival, immune cell trafficking, and belongs to a series of twelve canonical pathways, including leukocyte extravasation signalling, NF-kB activation, and glucocorticoid receptor signalling. *P38MAPK*, ubiquitin (*UBC*), and transforming growth factor beta receptor1 (TGFBR1) are in central positions in this network, which groups down-regulated genes involved in intracellular biochemistry modifications and in remodelling.

The fourth cluster includes the genes down-regulated at $t = +1$. These genes are related to apoptosis, NF-kB, and death receptor signalling canonical pathways.

Comparison of all lists of genes

Figure 5 shows the overlap of the three lists of genes (**(M1)**, **(M2)**, and **(M3)**). Twenty two genes are common between the three analyses: *ABHD2,*

Figure 5 Venn diagram showing DEG common to all lists of genes (**M1**, **M2** and **M3**; 22 DEG).

C3, C3AR1, C5AR1, CAPN3, CCDC47, CD163, CXCL13, DBN1, DGAT2, FAS, GYG1, HMOX1, NFAM1, PDXK, SELL, SERPING1, SOD2, TLR4, TNFRSF1A, TNFSF13B, TXNIP. Genes common to all three analyses are good candidates as genes majorly involved in the immune response. IPA analysis showed that these genes form two functional networks, NW1 and NW 2 (Table 2). The first functional network (NW1, Figure 6) is related to infectious diseases, cellular movement, hematological system development and function, cell-to-cell signalling and interaction. These genes form a node connected to ESR1. This gene encodes the estrogen receptor 1, a ligand-activated transcription factor. Estrogen receptors are also involved in pathological processes including breast cancer, endometrial cancer, and osteoporosis [43]. Among the genes that form these networks, two are particularly interesting, *TLR4* and *CD163*. The TLR4 gene was described as one of the important immunological factors influencing for example the development of mycoplasma pneumonia of swine [44]. TLR4 dysregulation is promoted aberrant cytokine production in bacterial sepsis [45]. The expression of porcine *CD163* (a scavenger receptor belonging to a cysteine-rich superfamily) on monocytes/macrophages correlates with permissiveness to African swine fever infection [46]. Cell entry of simian hemorrhagic fever virus is also dependent on *CD163* [47].

Table 2 Gene networks (NW) with commons DEG between the three lists of genes (in bold). ESR1* gene is common in both networks

NW	Genes in Network	Genes in present study	Top Diseases and Functions
1	**ABHD2**, *ALB*, *Calmodulin*, **CAPN3, CD163, DBN1**, *ESR1**, **FAS**, *GPR158, IgG, ITPKA*, **NFAM1, PDXK, SELL, SERPING1**, *SMARCA4, SYK*, **TLR4, TNFRSF1A**, *TREM1, YWHAZ*	12	Infectious Diseases, Cellular Movement, Hematological System Development and Function, Cell-To-Cell Signaling and Interaction
2	*2-methoxyestradiol*, **ABHD2**, *Cbp/p300*, **CCDC47**, *CDKN1A*, **CXCL13, DGAT2**, *EGFR, EP300, ESR1**, **GYG1, HMOX1**, *NR3C1*, *PPARG*, **SOD2**, *STAT1*, **TNFSF13B, TXNIP**	10	Cell Death and Survival, Cellular Development, Cellular Growth and Proliferation

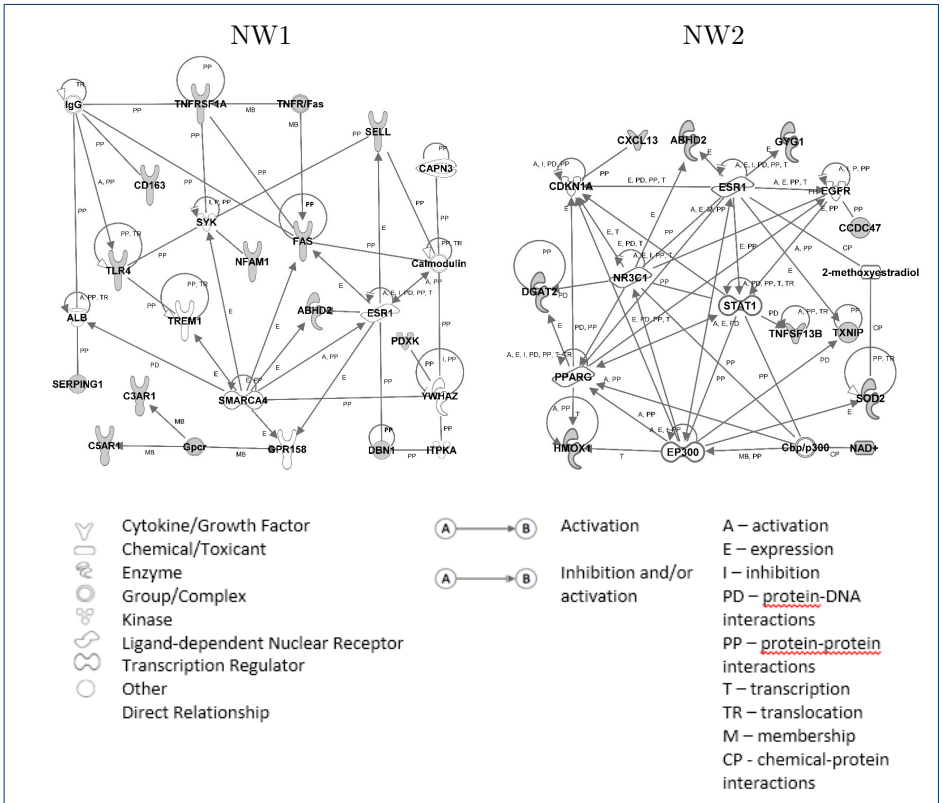


Figure 6 Functional networks formed by the 22 DEG common to **M1**, **M2** and **M3**. NW1: related to infectious diseases, cellular movement, hematological system development and function, cell-to-cell signalling and interaction. NW2: related to cell death and survival, cellular development, cellular growth and proliferation.

The second network (NW2) is related to cell death and survival, cellular development, cellular growth and proliferation. This network forms four functional nodes connected by *NR3C1, STAT1*, and *ESR1* as illustrated in Figure 6. The *NR3C1* (nuclear receptor subfamily 3, group C, member 1) gene encodes the glucocorticoid receptor, which can function both as a transcription factor that binds to glucocorticoid response elements in the promoters of glucocorticoid responsive genes, and as a regulator of other transcription factors. Signal transducer and activated transcription 1 (*STAT1*) has been identified as a point of convergence for the cross

talk between the pro-inflammatory cytokine interferon γ (IFNγ) and the Toll-like receptor-4 (*TLR4*) ligand LPS in immune cells [48]. LPS activates *STAT1* via the NF-κB pathway [49].

Conclusion

Our study raises a methodological challenge for statistical analysis. Indeed, we have presented here an integrative biological approach combining different statistical models and biological measures taking into consideration the longitudinal aspect of the data. This innovative analysis of biological data requires the development of a methodology adapted to both the multi-dimensional and longitudinal data.

LPS stimulation was chosen because it is standard to study general inflammation processes in many species. Immunological stress is the status of animals challenged by bacteria or viruses. It is associated with immunological, neurological, and endocrinological responses [4]. A four time point kinetic was studied. It has been reported that time points earlier than 24 hours are more relevant to decipher the onset of the response to stimulus as shown in kinetics studies in cow [50], pig [51], mouse [52] or human [53]. Moreover, kinetic studies have revealed that many genes return to their basal expression level by 24-48 hours of stimulation, suggesting that homeostasis is restored at that time [50, 51]. Our results provide many candidate genes to test for kinetic studies and ongoing complementary studies focused on this topic.

In conclusion, we have demonstrated that there are specific biomarkers indicative of an LPS-stimulated inflammatory response. Furthermore, these responses persist for prolonged periods of time and at significant expression levels, making them good candidate markers for evaluating the efficacy of anti-inflammatory drugs. The majority of the genes identified have known roles in the inflammatory process. Subsequently, these biomarkers may serve collectively as an indication of inflammation in swine. The knowledge gained from this series of experiments may aid in the development of a model for further studies.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VS, CY and NVV developed and performed all the statistics. PM and ET developed and undertook the experimental design, performed experimentations and ensured the biological interpretation. LG, ET and LL ensured the experimental transcriptomic analysis. YL provided the transcriptomic data set. VS, ET, DB and LL performed the transcriptomic related biology interpration. LG, DB, and AS performed the sampling, the gene expression study and the data management. YB was responsible of the animal management. VS, NVV, ET and PM wrote the paper and supervised all experimentations. All authors read and approved the final manuscript.

Acknowledgements

This project received financial support from the French National Research Agency (SUSoSTRESS, ANR-12-ADAP-0008) which also funded the PhD thesis of VS together with the Région Midi-Pyrénées. We thank the team of "le Magneraud" experimental station who took care of animal breeding, GeT-TRiX platform (Toulouse) for expression data production, GenPhySE INRA Toulouse (more specificaly Nathalie Iannuccelli, Katia Fève and Juliette Riquet) for experimental support, ANEXPLO platform (Toulouse) for biological assays, PEGASE INRA Saint-Gilles (more specifically Rafael Comte) for cortisol assay, Aurélie Ducan for hematology analysis, Aurélie Sécula-Tircazes for Fluidigm analysis, Magali San Cristobal and Pascal Martin for helpful discussions on the analyses and Mechnikov program (Embassy of France in Russia) for DB financial support.

Author details

[1]INRA, UMR 1388 GenPhySE, Université de Toulouse, INRA, INPT, ENVT, F-31326 Castanet-Tolosan, France. [2]Department of Behavioral Neurogenomics, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia. [3]INRA, UMR 1331 ToxAlim, F-31027 Toulouse, France. [4]INRA, UE 1372 GenESI, F-17700 Surgeres, France. [5]MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France.

References

1. Tawse J. Consumer attitudes towards farm animals and their welfare: a pig production case study. Bioscience Horizons. 2010;3(2):156–165.
2. Mormede P, Terenina E. Molecular genetics of the adrenocortical axis and breeding for robustness. Domestic Animal Endocrinology. 2012;43(2):116–131.
3. Mormede P, Foury A, Barat P, Corcuff JB, Terenina E, Marissal-Arvy N, et al. Molecular genetics of hypothalamic-pituitary-adrenal axis activity and function. Annals of the New York Academy of Sciences. 2011;1220(1):127–136. Available from: 10.1111/j.1749-6632.2010.05902.x.
4. Song C, Jiang J, Han X, Yu G, Pang Y. Effect of immunological stress to neuroendocrine and gene expression in different swine breeds. Molecular Biology Reports. 2014;41(6):3569–3576.
5. Dantzer R, O'Connor JC, Freund GG, Johnson RW, Kelley KW. From inflammation to sickness and depression: when the immune system subjugates the brain. Nature Reviews Neuroscience. 2008;9(1):46–56.
6. Pavlov VA, Tracey KJ. Neural circuitry and immunity. Immunologic Research. 2015;63(1-3):38–57.
7. Irwin MR, Cole SW. Reciprocal regulation of the neural and innate immune systems. Nature Reviews Immunology. 2011;11(9):625–632.
8. Cole SW. Elevating the perspective on human stress genomics. Psychoneuroendocrinology. 2010;35(7):955–962.
9. Lansade L, Valenchon M, Foury A, Neveux C, Cole SW, Layé S, et al. Behavioral and transcriptomic fingerprints of an enriched environment in horses (equus caballus). PloS ONE. 2014;9(12):e114384.
10. Vedhara K, Gill S, Eldesouky L, Campbell BK, Arevalo JMG, Ma J, et al. Personality and gene expression: do individual differences exist in the leukocyte transcriptome? Psychoneuroendocrinology. 2015;52:72–82.
11. Sautron V, Terenina E, Gress L, Lippi Y, Billon Y, Larzul C, et al. Time course of the response to ACTH in pig: biological and transcriptomic study. BMC Genomics. 2015;16(961):PMC4650497.
12. Hou X, Zhang J, Ahmad H, Zhang H, Xu Z, Wang T. Evaluation of antioxidant activities of ampelopsin and its protective effect in lipopolysaccharide-induced oxidative stress piglets. PloS ONE. 2014;9:e108314.
13. Westphal M, Stubbe H, Sielenkämper A, Borgulya R, Van Aken H, Ball C, et al. Terlipressin dose response in healthy and endotoxemic sheep: impact on cardiopulmonary performance and global oxygen transport. Intensive Care Medicine. 2003;29(2):301–308.
14. Larzul C, Terenina E, Foury A, Billon Y, Louveau I, Merlot E, et al. The cortisol response to ACTH in pigs, heritability and influence of corticosteroid-binding globulin. Animal. 2015;9(12):1929–1934.
15. Torgo L. Data Mining with R: Learning with Case Studies. CRC Data Mining and Knowledge Discovery Series. Boca Raton, Florida, USA: Chapman and Hall; 2010.
16. Bolstad BM, Irizarry RA, Åstrand 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.
17. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015. Available from: http://www.R-project.org.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B. 1995;57:289–300.
19. Gao Y, Flori L, Lecardonnel J, Esquerré D, Hu ZL, Teillaud A, et al. Transcriptome analysis of porcine PBMCs after in vitro stimulation by LPS or PMA/ionomycin using an expression array targeting the pig immune response. BMC Genomics. 2010;11:292.
20. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package **biomaRt**. Nature Protocols. 2009;4(8):1184–1191.
21. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006;22(13):1600–1607.
22. Liquet B, Lê Cao KA, Hocini H, Thiébaut R. A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. BMC Bioinformatics. 2012;13(1):325.
23. Thurstone LL. Multiple factor analysis. Psychological Review. 1931;38(5):406.
24. Myers MJ, Farrell DE, Baker JD, Cope CV, Evock-Clover CM, Steele NC. Challenge differentially affects cytokine production and metabolic status of growing and finishing swine. Domestic Animal Endocrinology. 1999;17(4):345–360.
25. Williams PN, Collier CT, Carroll JA, Welsh TH, Laurenz JC. Temporal pattern and effect of sex on lipopolysaccharide-induced stress hormone and cytokine response in pigs. Domestic Animal Endocrinology. 2009;37(3):139–147.
26. Dänicke S, Brosig B, Kersten S, Kluess J, Kahlert S, Panther P, et al. The *Fusarium* toxin deoxynivalenol (DON) modulates the LPS induced acute phase reaction in pigs. Toxicology Letters. 2013;220(2):172–180.
27. Llamas Moya S, Boyle L, Lynch P, Arkins S. Age-related changes in pro-inflammatory cytokines, acute phase proteins and cortisol concentrations in neonatal piglets. Neonatology. 2006;91(1):44–48.
28. Harizi H, Homo-Delarche F, Amrani A, Coulaud J, Mormede P. Marked genetic differences in the regulation of blood glucose under immune and restraint stress in mice reveals a wide range of corticosensitivity. Journal of Neuroimmunology. 2007;189(1):59–68.
29. Sun H, Huang Y, Yin C, Guo J, Zhao R, Yang X. Lipopolysaccharide markedly changes glucose metabolism and mitochondrial function in the *longissimus* muscle of pigs. Animal. 2016;p. 1–9.
30. Eigler N, Saccà L, Sherwin RS. Synergistic interactions of physiologic increments of glucagon, epinephrine, and cortisol in the dog: a model for stress-induced hyperglycemia. Journal of Clinical Investigation. 1979;63(1):114.
31. Guo J, Liu Z, Sun H, Huang Y, Albrecht E, Zhao R, et al. Lipopolysaccharide challenge significantly influences lipid metabolism and proteome of white adipose tissue in growing pigs. Lipids in Health and Disease. 2015;14(68).
32. Liu Z, Liu W, Huang Y, Guo J, Zhao R, Yang X. Lipopolysaccharide significantly influences the hepatic triglyceride metabolism in growing pigs. Lipids in Health and Disease. 2015;14:64.

33. Stanek C, Reinhardt N, Diesing AK, Nossol C, Kahlert S, Panther P, et al. A chronic oral exposure of pigs with deoxynivalenol partially prevents the acute effects of lipopolysaccharides on hepatic histopathology and blood clinical chemistry. Toxicology Letters. 2012;215(3):193–200.
34. Trauner M, Fickert P, Stauber RE. Inflammation-induced cholestasis. Journal of Gastroenterology and Hepatology. 1999;14(10):946–959.
35. Rosadini CV, Kagan JC. Early innate immune responses to bacterial LPS. Current Opinion in Immunology. 2016;44:14–19.
36. Liu W, OuLiu X, Yang J, Liu J, Li Q, Gu Y, et al. AP-1 activated by toll-like receptors regulates expression of IL-23 p19. Journal of Biological Chemistry. 2009;284(36):24006–24016.
37. Rohrbach S, Engelhardt S, Lohse MJ, Werdan K, Holtz J, Muller-Werdan U. Activation of AP-1 contributes to the beta-adrenoceptor-mediated myocardial induction of interleukin-6. Molecular Medicine. 2007;13(11/12):605–614.
38. Kang SS, Woo SS, Im J, Yang JS, Yun CH, Ju HR, et al. Human placenta promotes IL-8 expression through activation of JNK/SAPK and transcription factors NF-κB and AP-1 in PMA-differentiated THP-1 cells. International Immunopharmacology. 2007;7(11):1488–1495.
39. Park J, Chung SW, Kim SH, Kim TS. Up-regulation of interleukin-4 production via NF-AT/AP-1 activation in T cells by biochanin A, a phytoestrogen and its metabolites. Toxicology and Applied Pharmacology. 2006;212(3):188–199.
40. Hess J, Angel P, Schorpp-Kistner M. AP-1 subunits: quarrel and harmony among siblings. Journal of Cell Science. 2004;117(25):5965–5973.
41. Donn R, Berry A, Stevens A, Farrow S, Betts J, Stevens R, et al. Use of gene expression profiling to identify a novel glucocorticoid sensitivity determining gene, BMPRII. The FASEB Journal. 2007;21(2):402–414.
42. Thomas AV, Broers AD, Vandegaart HF, Desmecht DJM. Genomic structure, promoter analysis and expression of the porcine ($Sus\ scrofa$) $TLR4$ gene. Molecular Immunology. 2006;43(6):653–659.
43. Clarke R, Tyson JJ, Dixon JM. Endocrine resistance in breast cancer – an overview and update. Molecular and Cellular Endocrinology. 2015;418:220–234.
44. Borjigin L, Shimazu T, Katayama Y, Li M, Satoh T, Watanabe K, et al. Immunogenic properties of Landrace pigs selected for resistance to mycoplasma pneumonia of swine. Animal Science Journal. 2016;87(3):321–329.
45. Tan Y, Kagan JC. A cross-disciplinary perspective on the innate immune responses to bacterial lipopolysaccharide. Molecular Cell. 2014;54:212–223.
46. Lithgow P, Takamatsu H, Werling D, Dixon L, Chapman D. Correlation of cell surface marker expression with African swine fever virus infection. Veterinary Microbiology. 2014;168(2):413–419.
47. Caì Y, Postnikova EN, Bernbaum JG, Yú S, Mazur S, Deiuliis NM, et al. Simian hemorrhagic fever virus cell entry is dependent on CD163 and uses a clathrin-mediated endocytosis-like pathway. Journal of Virology. 2015;89(1):844–856.
48. Sikorski K, Chmielewski S, Przybyl L, Heemann U, Wesoly J, Baumann M, et al. STAT1-mediated signal integration between IFNγ and LPS leads to increased EC and SMC activation and monocyte adhesion. American Journal of Physiology, Cell Physiology. 2011;300(6):C1337–C1344.
49. Luu K, Greenhill CJ, Majoros A, Decker T, Jenkins BJ, Mansell A. STAT1 plays a role in TLR signal transduction and inflammatory responses. Immunology and Cell Biology. 2014;92(9):761–769.
50. Donaldson L, Vuocolo T, Gray C, Strandberg Y, Reverter A, McWilliam S, et al. Construction and validation of a bovine innate immune microarray. BMC Genomics. 2005;6:135.
51. Ledger TN, Pinton P, Bourges D, Roumi P, Salmon H, Oswald IP. Development of a macroarray to specifically analyze immunological gene expression in swine. Clinical and Diagnostic Laboratory Immunology. 2004;11(4):691–698.
52. Wells CA, Ravasi T, Faulkner GJ, Carninci P, Okazaki Y, Hayashizaki Y, et al. Genetic control of the innate immune response. BMC Immunology. 2003;4:5.
53. Boldrick JC, Alizadeh AA, Diehn M, Dudoit S, Liu CL, Belcher CE, et al. Stereotyped and specific gene expression programs in human innate immune responses to bacteria. Proceedings of the National Academy of Sciences. 2002;99(2):972–977.

Additional Files

Additional file 1 — List of 154 unique genes differentially expressed in list **(M1)**.

Additional file 2 — List of 116 unique genes differentially expressed in list **(M2)**.

Additional file 3 — Biological functions enriched by differentially expressed genes in list **(M2)** ($n = 30$). '.xls' file.

- Categories: Category of the enriched function;
- Diseases or Function annotation: Name of the enriched function;
- p-Value: FDR of the enrichment test;
- Genes: list of genes enriching the biological function;
- #Genes: number of genes enriching the biological function.

Additional file 4 — List of 9,530 unique genes differentially expressed in list **(M3)**.

Additional file 5 — List of 284 unique genes differentially expressed in list (M3) included in non-generic biological functions ($n = 30$). 'xlsx' file.

Generic biological functions include functions such as morphogenesis, transcription, locomotion.

- Gene name: name of the gene;
- Gene description: informations on the gene's molecular function;
- Cluster: cluster in which the gene is classified by HAC;
- Time point: time measurement where the gene is DE;
- Expression: whether the DEG is up or down-regulated.

Chapitre 3

Multiway-SIR pour l'intégration de données biologiques répétées

Sommaire

3.1 Introduction

Les progrès réalisés en biologie moléculaire et dans les techniques de séquençage à haut débit ont conduit à une augmentation considérable de la quantité de données 'omiques disponibles. Par ailleurs, grâce à la diminution du coût de ces techniques, les protocoles expérimentaux ont pu se complexifier peu à peu et il est à présent possible de mesurer des données à plusieurs niveaux du vivant, ainsi que réaliser des mesures répétées dans le temps afin de capturer des cinétiques. La mesure de données 'omiques longitudinales (qu'elles soient d'un ou plusieurs types) est donc un moyen de contrôler l'évolution d'un système biologique et de ses processus au cours du temps. Ces données sont caractérisées par la mesure répétée de p variables à T pas de temps sur les n mêmes individus.

L'analyse de ces données représente alors un challenge pour le biologiste qui doit appliquer une approche de biologie intégrative à de multiples tableaux temporels. L'intégration de ce type de données représente un défi pour le biologiste qui doit alors :

1. prendre en compte la nature longitudinale des données en modélisant la variabilité intra-individus, due à la perturbation introduite dans le processus biologique et la variabilité inter-individus ;
2. prendre en considération le fait que les données 'omiques sont souvent mesurées sur n individus pour p variables avec $n \ll p$;
3. mettre en relation des données de sources différentes avec une mesure servant de marqueur à l'évolution négative ou positive du système étudié.

Intégrer des données multivariées avec une variable quantitative cible permettrait donc l'identification de biomarqueurs, c'est à dire de la signature biologique de l'effet induit par un changement ou un stresseur.

L'analyse de données 'omiques longitudinales a fait l'objet de nombreuses méthodes. Il est possible de les diviser en plusieurs types d'approches.
L'utilisation de fonctions spline est une approche qui permet de modéliser l'évolution des variables au cours du temps. STRAUBE et collab. [2015] proposent une méthode qui, couplée à un filtrage et une modélisation par modèle mixte tient compte à la fois de la variabilité inter- et intra-individus. Ces profils d'évolution sont ensuite utilisés pour faire de la recherche de variables différentiellement exprimées au cours du temps et de la classification pour identifier des groupes de variables aux évolutions similaires. Cette approche n'est cependant pas satisfaisante dans notre cas, car elle nécessite un grand nombre de mesures dans le temps et ne

s'applique pas lorsque le nombre de pas de temps est faible.

Une autre approche plus adaptée aux cas où il y a peu de pas de temps est celle des méthodes d'intégration multi-tableaux. Parmi elles, on peut citer l' Analyse Factorielle Multiple (AFM ; [Abdi et collab., 2013; Escoffier et Pagès, 1990]), la Double Analyse en Composantes Principales (DACP ; [Bouroche, 1975]), la décomposition de type Tucker (ou décomposition en produits tensoriels) [Kruskal, 1989], les rotations Procrustes [Ten Berge, 1977] ou encore la Structuration de Tableaux à Trois Indices de la Statistique (STATIS et Dual-STATIS [Glacon, 1981; Lavit, 1988; Lavit et collab., 1994; L'Hermier des Plantes, 1976] ou encore les Analyses Triadiques Partielles (ATP) [Thioulouse et Chessel, 1987]). Toutes ces méthodes et leur utilisation dans le contexte des données multidimensionnelles répétées ont été décrites par Dazy et collab. [1996]. L'AFM, la DACP et les variantes de STATIS ont pour principe commun de considérer chaque pas de temps comme un tableau de données différent, puis de calculer des poids optimaux afin de pondérer chaque pas de temps et réaliser une ACP globale des données concaténées et pondérées. Elles diffèrent par le critère optimisé pour le calcul des poids. Par exemple, l'AFM pondère les tableaux en fonction de la variabilité existant à chaque pas de temps, tandis que STATIS pondère les tableaux en fonction de leur proximité avec un tableau compromis le plus représentatif possible de l'ensemble des pas de temps. Ces approches sont cependant non supervisées et ne permettent donc pas de faire l'intégration avec une variable cible. Vallejo-Arboleda et collab. [2007] proposent une variante de STATIS, CANOSTATIS (canonical STATIS), pour étendre l'analyse discriminante aux données cubiques et [Vivien et Sabatier, 2003] présentent la « Orthogonal Multiple Co-Intertia Analysis PLS » qui est une variante de la PLS (mode régression) dans le cas de données multiblocs explicatives et à expliquer. Dans le contexte de cette thèse, on souhaite cependant intégrer une variable cible réelle.

Cet objectif d'intégration est quant à lui rempli par des approches de régression telles que la régression PLS (*Partial Least Square*; [Wold, 1985]) et la régression inversée par tranche (SIR, *Sliced Inverse Regression*; [Li, 1991]). Ces approches supervisées permettent de faire le lien entre un jeu de données multivariées et une variable réelle cible, toutes mesurées une seule fois (à un seul pas de temps). Elles sont donc non adaptées à l'analyse de données répétées dans le temps.

Aucune des approches décrites jusqu'ici ne permet de réaliser l'analyse que nous souhaitons faire sur nos données multivariées longitudinales, à savoir : faire l'intégration temporelle de ces données avec une variable

cible réelle. Nous présentons dans ce chapitre la méthode « multiway-SIR » (ou régression par tranche inversée multi-tableaux). Cette méthode est une extension de dual-STATIS (voir chapitre 1, section STATIS et dual-STATIS) au cadre général de la SIR (voir chapitre 1, section Régression inverse par tranches). Elle s'applique pour explorer un jeu de données cubiques (T tableaux des mêmes p variables, mesurées sur les mêmes n individus) en tenant compte d'une variable réelle non répétée dans le temps.

Comme en dual-STATIS, la multiway-SIR s'intéresse aux structures de corrélations entre variables et à l'évolution de ces structures au cours du temps. Cette méthode se déroule en deux étapes : l'étude de l'inter-structure qui est la recherche d'une pondération optimale des tableaux initiaux pour construire une matrice globale et l'étude de l'intra-structure qui est la décomposition spectrale de cette matrice globale.
Étendre dual-STATIS à la SIR permet de calculer un espace EDR (*efficient reduction dimension*) pour variables longitudinales.
Le déroulement général de la méthode est illustré en figure 3.1.

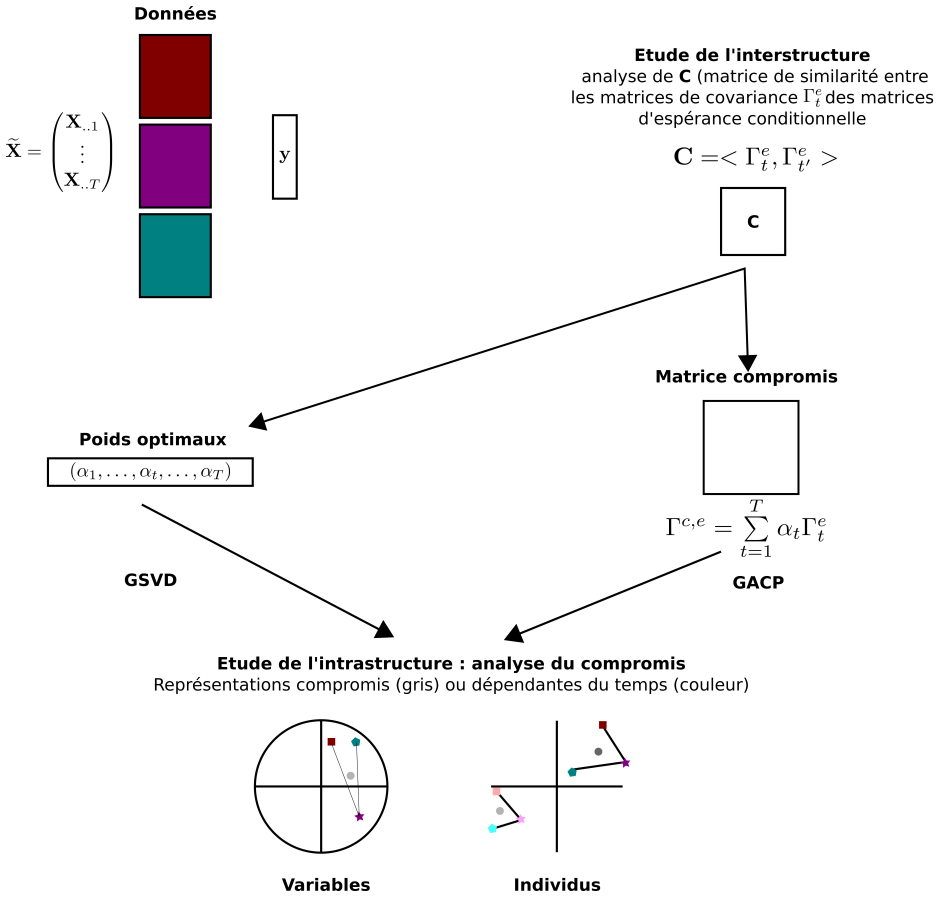


FIGURE 3.1 – Récapitulatif des étapes de l'approche multiway-SIR.

Le chapitre est organisé comme suit : dans la section 3.2, nous commençons par définir les notations utilisées au cours du chapitre. La sous-section 3.3 présente ensuite en détails le cadre numérique de la multiway-SIR. Dans la section 3.4, nous illustrons la méthode en l'appliquant sur les données de biologie clinique de l'expérience d'injection d'ACTH du projet SUSoSTRESS et en comparons les résultats avec ceux obtenu par une approche dual-STATIS classique dans la section 3.5. Enfin, dans la section 3.6, nous concluons sur les apports de la méthode ainsi que sur les développements possibles pour celle-ci.

3.2 Notations

Soit $\mathbf{X}$ notre jeu de données répétées observées. Plus précisément, p variables ont été mesurées sur les mêmes n individus un nombre T de fois avec $n > p$. Les observations dans $\mathbf{X}$ sont donc référencées par 3 indices : $\mathbf{X} = (x_{ijt})_{i=1,\dots,n, j=1,\dots,p, t=1,\dots,\mathrm{T}}$.
On note alors :

— $\mathbf{X}_{..t}$, la matrice $n \times p$ des observations au temps t, $\mathbf{X}_{.j.}$, la matrice $n \times \mathrm{T}$ des observations pour la variable j et $\mathbf{X}_{i..}$, la matrice $j \times \mathrm{T}$ de la $i^{\text{ème}}$ observation ;

— $\mathbf{x}_{.jt}$, le vecteur de taille n contenant toutes les observations de la variable j au temps t, $\mathbf{x}_{i.t}$, le vecteur de taille p contenant toutes mesures de l'individu i au temps t et $\mathbf{x}_{ij.}$, le vecteur de taille T contenant les observations à tous les pas de temps de l'individu i pour la variable j.

Ces tableaux à 3 indices peuvent aussi être appelés tableaux cubiques.

On note également X, $\mathrm{X}_{j.}$ et $\mathrm{X}_{.t}$ les variables aléatoires sous-jacentes aux observations $\mathbf{X}$. Elles sont définies respectivement dans $\mathbb{R}^{p\mathrm{T}}$, $\mathbb{R}^{\mathrm{T}}$ et $\mathbb{R}^{p}$.

Enfin, on peut aussi observer, sur nos n mêmes individus, une variable réelle Y avec $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$. On cherche à expliquer cette variable $\mathbf{y}$ à partir de $\mathbf{X}$.

Ces données sont illustrées en figure 3.2.

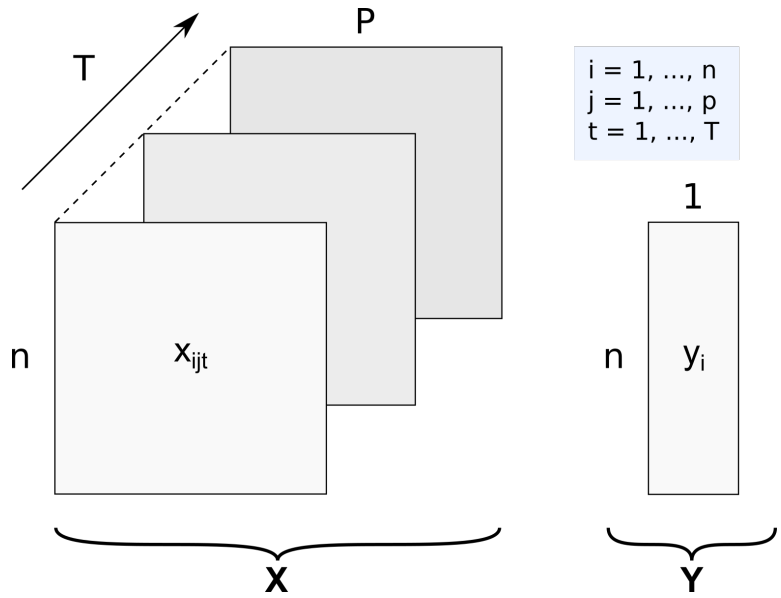


FIGURE 3.2 – Illustration des données sur lesquelles la multiway-SIR est appliquée. Elles sont constituées d'un tableau **X** cubique, avec p variables mesurées T fois sur les n mêmes individus et d'une variable réelle cible **y**, mesurée sur les même n individus.

3.3 Multiway-SIR : extension de dual-STATIS au cadre de la SIR

3.3.1 Présentation de la SIR

La SIR (*Sliced Inverse Regression*), introduite par LI [1991], est une approche adaptée à l'explication d'une variable réelle Y par une variable X multidimensionnelle ($X \in \mathbb{R}^p$) . Il s'agit d'un modèle de régression semi-paramétrique faisant intervenir une décomposition spectrale au sein d'un modèle de régression linéaire qui peut être utilisé à des fins exploratoires. Le modèle de régression de la SIR s'écrit comme suit :

$$\mathbf{Y} = f(\mathbf{X}^\top \mathbf{a}_1, \ldots, \mathbf{X}^\top \mathbf{a}_d, \epsilon) \tag{3.1}$$

où $d \ll p$, $f : \mathbb{R}^{d+1} \to \mathbb{R}$ est une fonction arbitraire inconnue, $(\mathbf{a}_k)_{k=1,\ldots,d}$ sont des vecteurs de $\mathbb{R}^p$ à estimer et ϵ est un terme d'erreur. On note **A** la matrice de dimension $p \times d$ dont les colonnes sont les $\mathbf{a}_k$.

Le principe de l'approche est d'estimer **A**, qui est un espace EDR (pour *effective dimension reduction*) contenant toute l'information disponible à propos de **Y** dans **X**. LI [1991] montre que sous les conditions appropriées, **A** peut être estimé par les d premiers vecteurs propres de Γ^e, la matrice de

variance-covariance empirique d'une matrice d'espérance conditionnelle $\mathbb{E}(\mathbf{Z}|\mathbf{Y})$ pour $\mathbf{Z} = (\mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}^\top)\Gamma^{-1/2}$ et Γ, la matrice de variance-covariance empirique de $\mathbf{X}$.

Le détail de la méthode est présenté au chapitre 2, section 1.7.4.

3.3.2 Analyse de l'inter-structure

L'étude de l'inter-structure consiste en l'étude des similarités entre les structures de corrélations entre variables aux différents pas de temps. Elle permet de construire une matrice compromis la plus représentative des relations entre $\mathbf{y}$ et la structure globale des corrélations entre variables dans $\mathbf{X}$ au cours du temps. Construire la matrice compromis est la première étape pour obtenir un espace de représentation commun à tous les pas de temps pour l'analyse des relations entre les structures de corrélations entre variable dans $\mathbf{X}$ et $\mathbf{y}$.
Dans un premier temps, on pré-traite les données originales en calculant pour tout t :

$$\mathbf{Z}_{..t} = (\mathbf{X}_{..t} - \mathbf{1}_n\bar{\mathbf{x}}_t^\top)\Gamma_t^{-1/2} \in \mathrm{M}_{n\times p}$$

avec $\Gamma_t = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, la matrice de covariance empirique de $\mathbf{X}_{..t}$. Si nous avions été dans le cas classique d'une régression SIR sur des données non répétées, l'analyse des relations entre $\mathbf{Z}$ et $\mathbf{y}$ aurait consistait en la construction d'une matrice d'espérance conditionnelle $\mathbb{E}(\mathbf{Z}|\mathbf{y})$ puis en l'ACP de sa matrice de covariance. Dans le cadre de la multiway-SIR, la construction du compromis se fait en calculant pour tout t la matrice d'espérance conditionnelle $\mathbb{E}(\mathbf{Z}_{..t}|\mathbf{y})$. Une estimation simple de cette matrice d'espérance conditionnelle est de découper $\mathbf{y}$ en H tranches $(\tau_h)_{h=1,\dots,\mathrm{H}}$ et de calculer les T matrices $\mathbf{G}_{..t}$ par :

$$\mathbf{G}_{..t} = \frac{1}{n_h}\Delta^\top\mathbf{Z}_{..t} \tag{3.2}$$

où n_h est le nombre d'observations i pour lesquelles y_i se trouve dans la tranche τ_h ($n_h = \sum_i \delta_{\{y_i\in\tau_h\}}$) et $\Delta = (\delta_{ih})$ est la matrice de dimension $n \times \mathrm{H}$ définie telle que $\delta_{ih} = 1$ si $y_i \in \tau_h$ et $\delta_{ih} = 0$ sinon. De cette manière, les tranches de $\mathbf{y}$ sont choisies de manière à avoir des effectifs de taille homogène, faisant ainsi varier l'amplitude des τ_h, mais permet de s'assurer d'obtenir des effectifs suffisants pour chacune des tranches. La matrice de covariance de $\mathbf{G}_{..t}$ est la matrice de dimension $p \times p$ notée :

$$\Gamma_t^e = \mathbf{G}_{..t}^\top\mathbf{M}\mathbf{G}_{..t}$$

où $\mathbf{M} = \mathrm{Diag}\left(\frac{n_1}{n},\dots,\frac{n_\mathrm{H}}{n}\right)$.

Pour étudier la proximité des structures de corrélation des matrices d'espérances conditionnelles aux differents pas de temps entre elles, on

calcule un coefficient de similarité qui représente la distance entre les T matrices de covariance. Ce coefficient de similarité est le produit scalaire de Frobenius entre les matrices de covariance et s'écrit :
$\forall t, t' = 1, \ldots, \mathrm{T}$:

$$c_{tt'} = \left\langle \Gamma_t^e, \Gamma_{t'}^e \right\rangle_{\mathrm{F}} = \mathrm{Trace}(\Gamma_t^{e\top} \Gamma_{t'}^e) = \mathrm{Trace}(\Gamma_t^e \Gamma_{t'}^e).$$

Lorsque des différences d'échelle importantes existent entre deux matrices ainsi qu'au sein de ces matrices, il peut être nécessaire de normer les matrices de covariance Γ_t^e pour leur donner une norme de Frobenius de 1. On note cette matrice normée :

$$\widetilde{\Gamma}_t^e = \frac{\Gamma_t^e}{\|\Gamma_t^e\|_{\mathrm{F}}^2}, \qquad \text{avec } \|\Gamma_t^e\|_{\mathrm{F}} = \mathrm{Trace}(\Gamma_t^{e\top} \Gamma_t^e). \tag{3.3}$$

Il est équivalent de normer les données originales et de travailler avec des matrices $\widetilde{\mathbf{Z}}_{..t}$ où :

$$\widetilde{\mathbf{Z}}_{..t} = \frac{\mathbf{Z}_{..t}}{\sqrt{\|\Gamma_t^e\|_{\mathrm{F}}}}. \tag{3.4}$$

En effet, à partir de $\widetilde{\mathbf{Z}}$, on peut redéfinir :

$$\begin{aligned} \widetilde{\mathbf{G}}_{..t} &= \frac{1}{n_h} \Delta^\top \widetilde{\mathbf{Z}}_{..t} \\ &= \frac{1}{n_h} \Delta^\top \frac{\mathbf{Z}_{..t}}{\sqrt{\|\Gamma_t^e\|_{\mathrm{F}}}}. \end{aligned}$$

Dans ce cas, on redéfinit le coefficient de similarité par :

$$\tilde{c}_{tt'} = \mathrm{Trace}(\widetilde{\Gamma}_t^e \widetilde{\Gamma}_{t'}^e) \tag{3.5}$$

où $\tilde{c}_{tt'}$ peut être interprété comme le cosinus entre les différentes matrices de covariance $\widetilde{\Gamma}_t^e$.

Comme Γ_t et $\Gamma_{t'}$ sont positives semi-définies, $\tilde{c}_{tt'}$ est toujours positif et toujours inférieur ou égal à 1. Il peut être interprété comme le cosinus de l'angle entre les matrices de covariance à t et t'. La matrice $\widetilde{\mathbf{C}} = (\tilde{c}_{tt'})_{t,t'=1,\ldots,\mathrm{T}}$ représente l'information globale sur les similitudes et différences entre les structures de corrélations aux différents pas de temps. Pour étudier les relations entre les tableaux, on résout le problème d'optimisation suivant :

$$\mathbf{u}_1 = \underset{\mathbf{u}=(u_1,\ldots,u_{\mathrm{T}})^\top \in \mathbb{R}^{\mathrm{T}}, \|\mathbf{u}\|=1}{\arg\max} \left\| \sum_{t=1}^{\mathrm{T}} u_t \widetilde{\Gamma}_t^e \right\|_{\mathrm{F}}^2 \tag{3.6}$$

Ce problème d'optimisation est équivalent à trouver une matrice $\Gamma^* = \sum_{t=1}^{\mathrm{T}} u_t \widetilde{\Gamma}_t^e$ qui présente la plus forte similarité moyenne (au sens du produit

scalaire de Frobenius) avec les matrices $(\widetilde{I}^e_t)_{t=1,\dots,\mathrm{T}}$. Une solution est donnée par $u_1 = (u_{11},\dots,u_{1t})^\top \in \mathbb{R}^\mathrm{T}$, le premier vecteur propre de $\widetilde{\mathbf{C}}$. De manière équivalente, on peut écrire que I^* est la combinaison linéaire des matrices de corrélation normées qui maximise le carré moyen du produit scalaire de Frobenius de ces matrices :

$$I^* = \underset{\mathbf{u}=(u_1,\dots,u_\mathrm{T})^\top \in \mathbb{R}^\mathrm{T}, \|\mathbf{u}\|=1}{\arg\max} \sum_{t'=1}^{\mathrm{T}} \left\langle \sum_{t=1}^{\mathrm{T}} u_t \widetilde{I}_t, {}^e\widetilde{I}^e_{t'} \right\rangle_\mathrm{F}^2. \tag{3.7}$$

Comme les éléments de $\widetilde{\mathbf{C}}$ sont positifs, tous les éléments de $\mathbf{u}_1$ sont de même signe et il peut être montré que $\mathbf{u}_{1t}$ est proportionnel au produit scalaire de Frobenius entre la matrice résumée I^* et $\widetilde{I}^e_t$, c'est à dire :

$$\langle I^*, \widetilde{I}^e_t \rangle_\mathrm{F} = \mathrm{Trace}(I^* \widetilde{I}^e_t) = \mu_1 u_{1t}$$

et que la première valeur propre μ_1 est le maximum de (1.1) et (3.7), c'est à dire :

$$\|I^*\|_\mathrm{F}^2 = \mu_1.$$

Preuve :

$$\begin{aligned}
\left\langle I^*, \widetilde{I}^e_t \right\rangle_\mathrm{F} &= \left\langle \sum_{t'=1}^{\mathrm{T}} u_{1t'} \widetilde{I}^e_{t'}, \widetilde{I}^e_t \right\rangle_\mathrm{F} \\
&= \mathrm{Trace}\Big(\sum_{t'=1}^{\mathrm{T}} u_{1t'} \widetilde{I}^e_{t'} \widetilde{I}^e_t \Big) \\
&= \sum_{t'=1}^{\mathrm{T}} u_{1t'} \mathrm{Trace}\big(\widetilde{I}^e_{t'} \widetilde{I}^e_t \big) \\
&= \sum_{t'=1}^{\mathrm{T}} u_{1t'} \tilde{c}_{t't} \\
&= \widetilde{\mathbf{C}} \mathbf{u}_{1\rfloor t} = \mu_1 u_{1t}
\end{aligned} \tag{3.8}$$

car $\mathbf{u}_1$ est le premier vecteur propre de $\widetilde{\mathbf{C}}$, associée à la première valeur propre μ_1.

De plus,

$$\begin{aligned}
\|I^*\|_\mathrm{F}^2 &= \left\langle \sum_{t=1}^{\mathrm{T}} u_{1_t} \widetilde{I}^e_t, u_{1_t} \widetilde{I}^e_t \right\rangle_\mathrm{F} \\
&= \sum_{t,t'=1}^{\mathrm{T}} u_{1_t} u_{1'_t} \langle \widetilde{I}^e_t, \widetilde{I}^e_{t'} \rangle_\mathrm{F} \\
&= \sum_{t,t'=1}^{\mathrm{T}} u_{1_t} u_{1'_t} \tilde{c}_{tt'} \\
&= \mathbf{u}_1^\top \widetilde{\mathbf{C}} \mathbf{u}_1 = \mu_1
\end{aligned} \tag{3.9}$$

car $\mathbf{u}_1$ est le vecteur propre de $\widetilde{\mathbf{C}}$ associé à la première valeur propre μ_1 et que sa norme est égale à 1. □

Les équations (3.8) et (3.9) permettent d'écrire $\sqrt{\mu_1}u_{1t} = \frac{\left\langle I^*, \widetilde{I}_t^e \right\rangle_{\mathrm{F}}}{\left\| I^* \right\|_{\mathrm{F}} \left\| \widetilde{I}_t^e \right\|_{\mathrm{F}}}$, ce qui peut être interprété comme le cosinus entre la matrice résumée I^* et la structure de corrélation conditionnelle au pas de temps t, $\widetilde{I}_t^e$.

Une représentation graphique de l'inter-structure peut être obtenue en plaçant chaque pas de temps t à la coordonnée $\sqrt{\mu_1}u_{1t}$ et permet de résumer l'analyse. Plus cette coordonnée est grande, et plus la similarité entre le pas de temps correspondant et la matrice résumée est grande. De même, plus deux pas de temps sont proches et plus leurs structures de corrélations sont similaires.
En général, on multiplie les poids trouvés par un même facteur de sorte que leur somme soit égale à 1 :

$$\alpha_t = \frac{\sqrt{\mu_1}u_{1t}}{\sum_{t'=1}^{\mathrm{T}} \sqrt{\mu_1}u_{1t'}} = \frac{u_{1t}}{\sum_{t'=1}^{\mathrm{T}} u_{1t'}}$$

Enfin, on définit la matrice « compromis » $I^{e,c}$:

$$I^{e,c} = \sum_{t=1}^{\mathrm{T}} \alpha_t \widetilde{I}_t^e$$

Cette matrice compromis, qui ne diffère de la matrice résumée I^* que par un facteur $\sum_{t=1}^{\mathrm{T}} u_{1t}$, capture les différences entre structures de covariances qui restent stables dans le temps. De plus, on peut définir un critère de qualité de compromis par $\frac{\mu_1}{\sum_l \mu_l}$. Ce critère est compris entre 0 et 1 et mesure la proximité globale des structures de covarriances des pas de temps entre eux.

3.3.3 Analyse de l'intra-structure

L'objectif de l'analyse de l'intra-structure est, à travers la décomposition de la matrice compromis, d'obtenir un espace de représentation commun qui permette de représenter au mieux les pas de temps qui présentent le plus de similarité au niveau de leurs structures de corrélations. Si la SIR avait été utilisée à chaque pas de temps, elle aurait consisté en la SVD généralisée du triplet suivant : $(\widetilde{\mathbf{G}}_{..t}, \mathbb{I}_p, \mathbf{M})$. Pour en obtenir l'extension aux mesures répétées, la SVD généralisée est appliquée sur $(\widetilde{\mathbf{G}}, \mathbb{I}_p, \mathbf{D})$, où :

— $\widetilde{\mathbf{G}} = \begin{bmatrix} \widetilde{\mathbf{G}}_{..1} \\ \vdots \\ \widetilde{\mathbf{G}}_{..\mathrm{T}} \end{bmatrix}$ est le jeu de données sur lequel porte l'analyse et est une matrice de dimension (HT) × p ;

— $\mathbb{I}_p$ est la métrique associée aux variables et est de dimension $p \times p$;

— $\mathbf{D} = \text{Diag}(\alpha_t)_{t=1,\ldots,\text{T}} \otimes \mathbf{M}$ ($\mathbf{M} = \frac{1}{n}\mathbb{I}_{\text{HT}}$) est une matrice carrée de dimension HT × HT et est la métrique associée aux tranches (considérées aussi comme les observations).

L'approche est équivalente à faire l'ACP de $\mathcal{I}^{e,c}$ et permet d'obtenir 3 matrices $\mathbf{P}$, $\mathbf{Q}$ et Λ de dimensions respectives (HT) × r, $p \times r$ et $r \times r$ avec $r \leq \min(\text{HT}, p)$, le rang de $\mathcal{I}^{e,c}$. Ces matrices satisfont les conditions suivantes :

$$\widetilde{\mathbf{G}} = \mathbf{P}\Lambda\mathbf{Q}^\top \qquad \text{avec } \mathbf{P}^\top\mathbf{D}\mathbf{P} = \mathbf{Q}^\top\mathbf{Q} = \mathbb{I}_r \text{ et } \Lambda = \text{Diag}(\sqrt{\lambda_k})_{k=1,\ldots,r}$$

avec $\lambda_1 > \lambda_2 > \ldots > \lambda_r$. Cette décomposition permet d'obtenir les coordonnées pour représenter dans l'espace commun les tranches et les variables :

— les **tranches** sont représentées par leur score sur les composantes principales. Ces scores sont contenus dans la matrice $\mathbf{F} = \mathbf{P}\Lambda$. $\mathbf{F}_{hk}$ représente le score de la tranche h sur la composante k. $\mathbf{F}$, est une matrice de dimension (nT) × r et est de la forme : $\mathbf{F} = \begin{bmatrix} \mathbf{F}_{..1} \\ \vdots \\ \mathbf{F}_{..\text{T}} \end{bmatrix}$;

— comme les individus sur lesquels sont mesurées les variables sont les mêmes à tous les pas de temps, il est possible de calculer des **coordonnées compromis des tranches**. La position compromis de la tranche h correspond au centre de gravité des coordonnées de la tranche h à tous les pas de temps, pondérées par les poids calculés à l'inter-structure. On note cette matrice de composantes principales compromis : $\mathbf{F}^c = \sum_{t=1}^{\text{T}} \alpha_t \mathbf{F}_{..t}$;

— la position compromis des **variables** sur un cercle de corrélations est obtenue en calculant la corrélation entre $\mathbf{F}_{.k.}$ (la colonne k de $\mathbf{F}$) et $\widetilde{\mathbf{G}}_{.j.}$, la colonne j de $\widetilde{\mathbf{G}}$. Elle est égale à $v_{jk} = \frac{\lambda_k}{\hat{\sigma}_j}\mathbf{Q}_{jk}$ où $\hat{\sigma}_j$ est l'écart-type empirique de $\widetilde{\mathbf{G}}$ avec la métrique $\mathbf{D}$;

— il est aussi possible d'obtenir des **coordonnées spécifiques aux pas de temps pour les variables**. Elles s'obtiennent en calculant la corrélation entre la colonne k de $\mathbf{F}$ et les colonnes de la matrice $\widetilde{\mathbf{G}}^*_{..t}$. $\widetilde{\mathbf{G}}^*_{..t}$ est définie telle que les valeurs dans $\widetilde{\mathbf{G}}$ ont toutes été fixées à 0 (leur valeur moyenne puisque les variables sont centrées) à l'exception des valeurs mesurées au pas de temps t. Calculer la corrélation entre $\widetilde{\text{G}}^*_{.jt}$ et $\mathbf{F}_{.k}$ revient à projeter les variables dans $\widetilde{\mathbf{G}}^*_{..t}$ comme des variables supplémentaires sur les composantes principales, qui se focalisent sur les variables au pas de temps t. Cette coordonnée temporelle s'écrit :

$$v^t_{jk} = \frac{\left\langle \widetilde{\text{G}}^*_{.jt}, \mathbf{F}_{.k} \right\rangle_{\mathbf{D}}}{\left\| \widetilde{\text{G}}^*_{.jt} \right\|_{\mathbf{D}} \left\| \widetilde{\text{F}}_{.k.} \right\|_{\mathbf{D}}} = \frac{\left\langle \widetilde{\text{G}}^*_{.jt}, \text{P}_{.k} \right\rangle_{\mathbf{D}}}{\left\| \widetilde{\text{G}}^*_{.jt} \right\|_{\mathbf{D}}} = \frac{\left\langle \widetilde{\text{G}}^*_{.jt}, \text{P}_{.k} \right\rangle_{\mathbf{M}}}{\hat{\sigma}^t_j}$$

où $\mathbf{M}$ est la métrique $\frac{1}{n}\mathbb{I}_{\mathrm{HT}}$ et $\hat{\sigma}_j^t$ est l'écart type empirique de $\widetilde{\mathrm{G}}^*_{.jt}$ avec la métrique $\mathbf{M}$.

3.3.4 Choix méthodologiques

L'approche multiway-SIR demande de faire des choix méthodologiques sur deux points : le nombre de tranches H pour découper $\mathbf{y}$ et le choix de la dimension de projection.

Nombre de tranches H

Le choix du nombre de tranches H n'est pas un point critique de l'algorithme. LI [1991] montre par des simulations que le nombre de tranches en SIR peut affecter la covariance asymptotique de l'espace EDR, mais que les différences ne sont pas importantes et que le choix de H n'influe pas sur la vitesse de convergence de l'algorithme qui permet d'estimer l'espace EDR à une vitesse $1/\sqrt{n}$. Les deux valeurs particulières où H = 1 (une seule tranche) et H = n (une seule observation par tranche) ont peu d'intérêt. Dans le premier cas, les matrices $\widetilde{I}_t^e$ sont alors égales à des matrices nulles, tandis que dans le second, toutes les matrices de covariances sont égales à la même matrice identité et il n'y a pas de sens de chercher une distance entre elles. SARACCO et collab. [1999] suggère de choisir un nombre de tranches H supérieur au nombre de dimensions utilisées pour la projection afin d'éviter une réduction artificielle du modèle. Dans le cas exploratoire où l'on se trouve, l'approche est similaire à réaliser une analyse factorielle discriminante (AFD) sur la variable y discrétisée et il est plus avantageux de réaliser l'analyse sur des tranches très contrastées.

Choix de la dimension de projection

Plusieurs méthodes ont été proposées pour l'estimation de la dimension de projection en SIR. Ainsi, LI [1991] et SCHOTT [1994] présentent des approches se basant sur des tests d'hypothèses emboités (tests appliqués sur la distribution de la valeur moyenne des d ($d \leq r$) premières valeurs propres avec d augmentant séquentiellement). Ferré [FERRE, 1996; FERRÉ, 1998], quant à lui, propose une approche basée sur la qualité de l'estimation des espaces EDR évaluée par une mesure par validation croisée de l'écart entre l'espace EDR et son estimation.

3.4 Application de la multiway-SIR aux données de biologie clinique de l'expérience ACTH

Nous avons développé la multiway-SIR dans l'objectif de nous doter d'un outil d'analyse adapté aux données du projet SUSoSTRESS dans lequel se place cette thèse. Nous illustrons l'utilisation de la méthode sur les données de biologie clinique issues de l'expérience d'ACTH, déjà exploitées au cours du chapitre 2 de ce travail. L'objectif est ici d'explorer les données de biologie clinique en tenant compte de l'intensité de l'activité de l'axe corticotrope lors des réponses de stress, mesurée par la mesure de cortisol à $t = +1$.

3.4.1 Présentation des données

Les données utilisées sont les données de biologie clinique (numération sanguine et métabolites plasmatiques) récoltées dans l'expérience d'injection d'ACTH (voir chapitre 2, article 1 : « Time-course study of the response to ACTH in pigs : biological and transcriptomic study ». La liste des variables et des mesures auxquelles elles correspondent sont en annexe A.
Les données sont organisées comme suit :

— le nombre de variables explicatives (les variables cliniques) est $p = 14$;

— le nombre d'individus (des porcs) est $n = 120$;

— le nombre de tableaux, correspondant aux pas de temps auxquels les variables ont été mesurées est T = 4 et les pas de temps sont indicés par $t \in \{0, +1, +4, +24\}$;

— la variable cible réelle non répétée dans le temps **y** est la mesure du cortisol à $t = +1$, ce qui correspond au pas de temps où le cortisol (l'hormone principale du stress) atteint son maximum dans la circulation sanguine après une injection d'ACTH.

L'objectif de cette application est d'étudier si la variabilité inter-individuelle dans l'activité de l'axe corticotrope (mesurée par le cortisol sanguin) influence l'évolution des relations entre les variables cliniques au cours du temps.
La méthode est en cours d'implémentation dans R dans le package **SirStatis**. Les résultats et figures présentés dans cette section sont directement tirés de ce package.

3.4.2 Étude de l'inter-structure

Cette première section est consacrée à l'étude de l'inter-structure. Cette étape permet donc l'étude des similarités entre les structures de corrélations entre variables cliniques aux différents pas de temps en tenant compte de la variabilité inter-individuelle de la production de cortisol à $t = +1$. Dans un premier temps, plusieurs choix de H ont été testés pour le découpage de $\mathbf{y}$: 5, 10, 20 et 40. Plus le nombre de tranches est faible et plus les moyennes par tranche sont contrastées. Le tableau 3.1 montre les matrices cosinus $\widetilde{\mathbf{C}}$ obtenues en fonction du découpage de $\mathbf{y}$. Les valeurs des cosinus entre les pas de temps diminuent concomitamment au nombre de tranches. Ainsi, plus les moyennes conditionnelles sont contrastées et plus on observe l'apparition de différences entre les pas de temps. Cette observation signifie que les les structures de corrélations entre variables sont dépendantes du niveau de cortisol à $t = +1$. De plus, comme on observe des cosinus faibles entre pas de temps lorsqu'on considère la matrice $\widetilde{\mathbf{C}}$ obtenue pour les matrices d'espérance conditionnelles les plus contrastées (H = 5), on peut aussi en déduire que les structures de corrélations évoluent au cours du temps.

TABLEAU 3.1 – Matrices cosinus, $\widetilde{\mathbf{C}}$ de l'étude l'inter-structure de la multiway-SIR en fonction du nombre de tranches, H.

temps	$t=0$	$t=+1$	$t=+4$	$t=+24$
$t=0$	1.00	0.49	0.39	0.45
$t=+1$		1.00	0.44	0.48
$t=+4$			1.00	0.49
$t=+24$				1.00

(a) H = 5

temps	$t=0$	$t=+1$	$t=+4$	$t=+24$
$t=0$	1.00	0.57	0.57	0.54
$t=+1$		1.00	0.50	0.52
$t=+4$			1.00	0.57
$t=+24$				1.00

(b) H = 10

temps	$t=0$	$t=+1$	$t=+4$	$t=+24$
$t=0$	1.00	0.72	0.76	0.75
$t=+1$		1.00	0.70	0.67
$t=+4$			1.00	0.74
$t=+24$				1.00

(c) H = 20

temps	$t=0$	$t=+1$	$t=+4$	$t=+24$
$t=0$	1.00	0.86	0.86	0.89
$t=+1$		1.00	0.87	0.85
$t=+4$			1.00	0.84
$t=+24$				1.00

(d) H = 40

De la même façon, on observe que les poids optimaux pour chaque pas de temps lors de la construction du compromis ont tendance à s'uniformiser lorsque le nombre de tranches H augmente (voir tableau 3.2).

On choisit de concentrer l'étude sur l'analyse réalisée avec H = 5, qui permet de présenter les résultats les plus contrastés. La représentation graphique de la corrélation entre le tableau de covariance d'un pas de temps particulier et le tableau de covariance compromis est donnée en figure 3.3.

TABLEAU 3.2 – Poids optimaux, α, à chaque pas de temps en fonction du nombre de tranches H pour découper **y**.

H	$t=0$	$t=+1$	$t=+4$	$t=+24$
5	0.244	0.255	0.244	0.257
10	0.256	0.245	0.250	0.249
20	0.255	0.243	0.253	0.249
40	0.250	0.250	0.250	0.249

Chaque point sur la figure représente un pas de temps et plus le score pour un pas de temps est élevé, plus les structures de corrélation de celui-ci sont proches de celles du consensus. De même, plus deux points sont proches et plus leurs structures de corrélations sont proches entre elles. On peut voir que le pas de temps le plus proche du consensus est $t=+24$, qui est lui même très proches $t=+1$. En revanche, ils sont tous les deux assez éloignés de $t=0$ et $t=+4$ qui sont eux même proches entre eux. La figure fournit également la qualité globale du compromis (à quel point il est représentatif de l'ensemble des pas de temps). La qualité est ici de 59,42%.

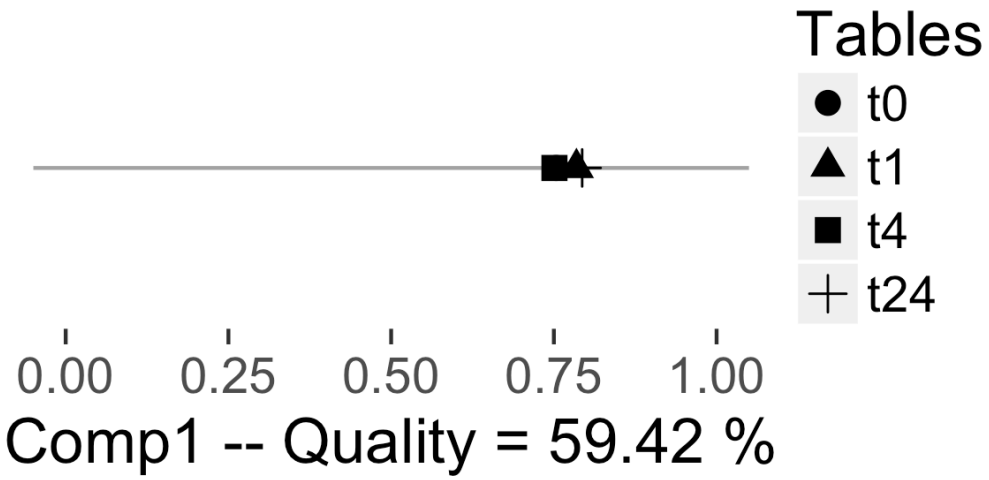


FIGURE 3.3 – Représentation graphique de l'inter-structure avec **y** découpé en H = 5 tranches.

3.4.3 Etude de l'intra-structure

La deuxième étape est l'analyse de l'intra-structure qui consiste en l'analyse du compromis $I^{e,c}$.

Choix de dimension de projection

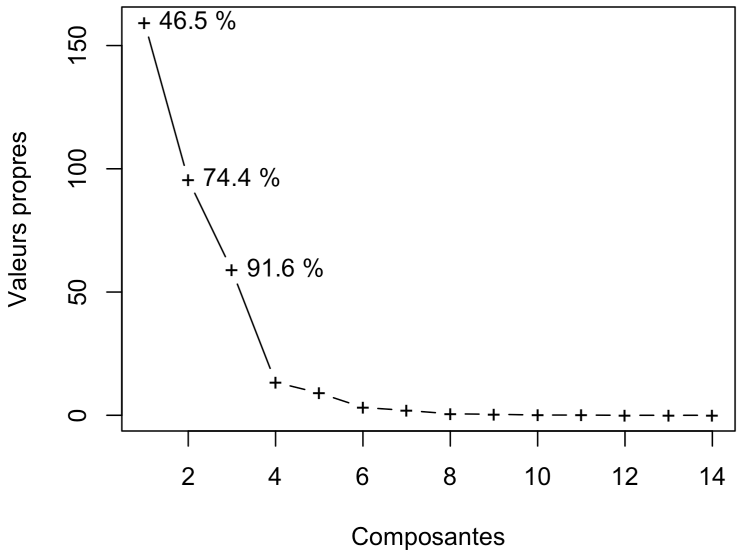
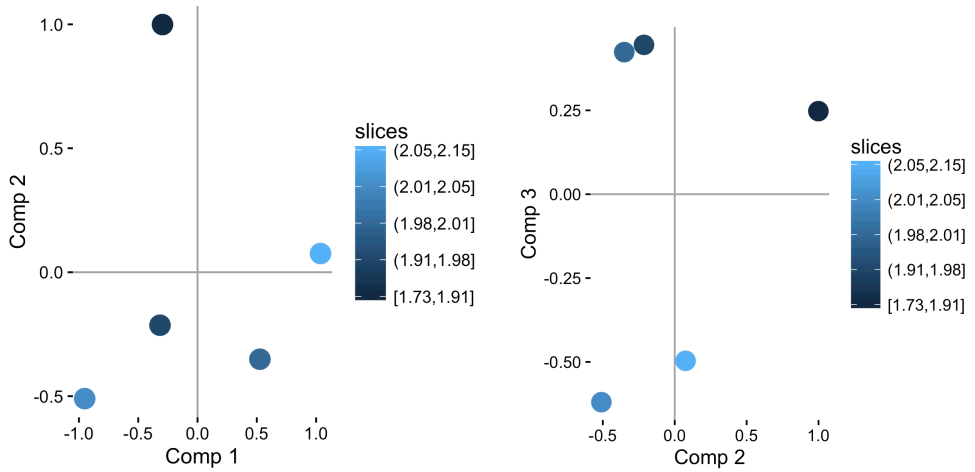


FIGURE 3.4 – Éboulis des valeurs propres et pourcentage de variance reproduite en fonction du nombre de composantes conservées.

La figure 3.4 représente l'éboulis des valeurs propres ainsi que le pourcentage de variance reproduite en fonction du nombre de composantes conservées. La première composante reproduit 46,5% de la variance. La deuxième en reproduit 27,9% et la troisième, 17,2%. On peut identifier un coude entre la troisième et la quatrième valeur propre, ce qui suggère que 3 composantes peuvent être conservées pour la projection.

Représentation des tranches et des variables en positions compromis

La figure 3.5 représente les positions compromis sur les 3 premières composantes de la multiway-SIR. Chaque point représente une tranche de valeur du cortisol à $t = +1$ et plus le bleu est clair, plus la valeur du cortisol à $t = +1$ après injection d'ACTH est élevée. L'axe 1 oppose la tranche avec les valeurs les plus fortes du cortisol à $t = +1$ (à droite) avec la deuxième tranche des valeurs les plus fortes (à gauche). Cela signifie que la plus grande source de variabilité entre les valeurs moyennes des variables pour les différentes tranches des valeurs de cortisol correspond à une opposition entre ces deux tranches d'individus. Cette opposition n'est pas expliquée par le niveau moyen du cortisol à $t = +1$ (puisque les valeurs du niveau de cortisol sont proches) mais par une différence dans les valeurs moyennes des variables biologiques elles-mêmes pour les individus ayant une réponse forte au cortisol. L'axe 2 montre une opposition entre la tranche des

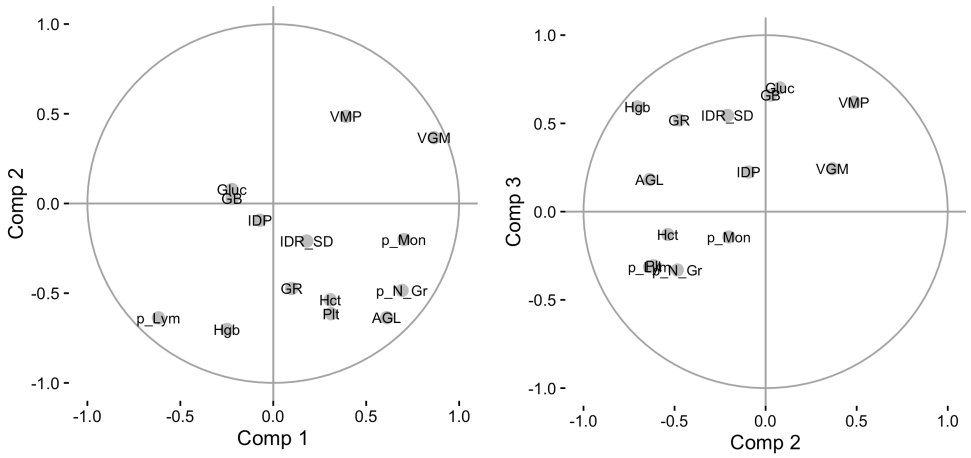


(a) Composantes 1 et 2 (b) Composantes 2 et 3

FIGURE 3.5 – Représentation des tranches en positions compromis sur les 3 premières composantes de la multiway-SIR (H = 5). L'échelle de couleur représente les tranches de valeurs du cortisol à $t = +1$, les valeurs allant du plus faible (bleu le plus foncé) au plus fort (bleu le plus clair).

individus ayant une valeur du cortisol à $t = +1$ très faible et les individus des autres tranches. Cet axe est donc caractéristique des individus ayant une réponse faible au cortisol. L'axe 3 montre une opposition entre les trois tranches avec les niveaux de cortisol les plus faibles à $t = +1$ (en haut) avec les deux tranches correspondant aux plus forts niveaux de cortisol à $t = +1$ (en bas). Cet axe présente donc un intérêt particulier pour expliquer la variabilité des variables cliniques par la variabilité de la réponse au cortisol à $t = +1$.

L'interprétation de cette figure est à mettre en parallèle avec celle de la figure 3.6 qui représente les positions compromis des variables sur le cercle des corrélations, pour les trois premières composantes de la multiway-SIR. L'axe 1 est caractérisé par des valeurs fortes du volume globulaire moyen (VGM), de la proportion de monocytes (p_Mon), de la proportion de granulocytes (p_N_Gr) et d'acides gras libres (AGL) en opposition avec une valeur fortement négative de proportion de lymphocytes (p_Lym). Les individus ayant la plus forte réponse du cortisol présente donc globalement une valeur élevée de VGM, p_Mon, p_N_Gr et AGL et s'oppose aux individus ayant une réponse du cortisol juste inférieur qui, eux, ont une valeur élevé de p_Lym. L'axe 2 est caractérisé par les variables proportion de lymphocytes (p_Lym), hémoglobine (Hgb), plaquettes (Plt) et acides gras libres (AGL) qui ont toutes une valeur de corrélation négative forte avec cet axe. Les individus ayant une réponse du cortisol la moins élevée sont caracté-



(a) Composantes 1 et 2 (b) Composantes 2 et 3

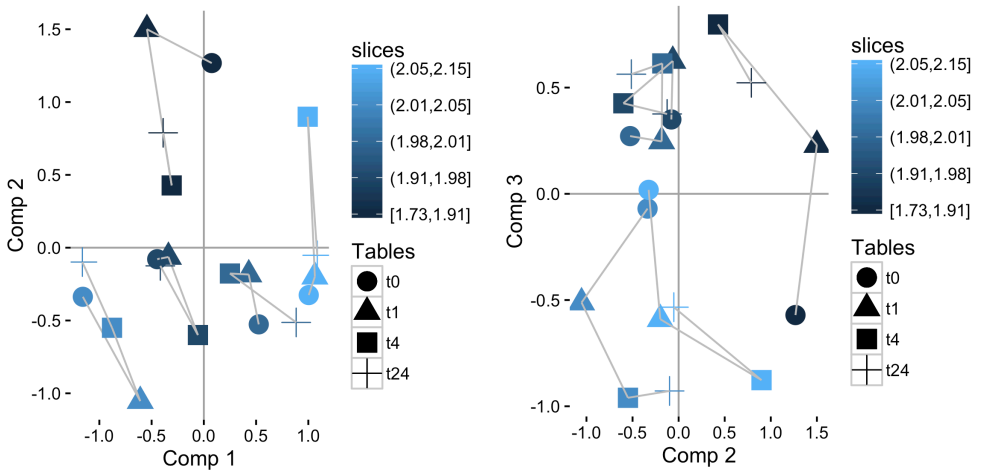
FIGURE 3.6 – Représentation des variables en positions compromis sur les 3 premières composantes de la multiway-SIR (H = 5).

risés par des valeurs faibles de ces variables. L'axe 3 est caractérisé par une corrélation positive forte avec les variables hémoglobine (Hgb), globules rouges (GR), indice de répartition des globules rouges (IDR_SD), globules blancs (GB), glucose (Gluc) et volume plaquettaire moyen (VPM). Ces variables sont celles qui permettent de discriminer les individus avec une réponse du cortisol faible (qui ont des valeurs fortes pour ces variables) des individus qui ont une réponse du cortisol forte (valeurs faibles pour ces variables).

Représentation des trajectoires temporelles des tranches et des variables

La figure 3.7 donne les représentations longitudinales des tranches et aide à préciser les interprétations données grâce aux figures 3.5 et 3.6.

Sur l'axe 1, les individus ayant la réponse du cortisol la plus élevée (à droite) ont une coordonnée constamment forte sur cette axe, quel que soit le pas de temps : ils sont donc constamment caractérisés par une valeur élevée des variables volume globulaire moyen (VGM), proportion des monocytes (p_Mon), proportion des granulocytes (p_N_Gr) et acides gras libres (AGL) et une valeur constamment faible de proportion de lymphocytes (p_-Lym). Toujours sur l'axe 1, les individus ayant la réponse du cortisol juste inférieure à la plus forte, qui s'opposent aux individus précédents (et sont donc à gauche de l'axe) sont particulièrement corrélés à l'axe 1 pour les temps $t = 0$ et $t = +24$. La valeur faible de volume globulaire moyen (VGM), proportion de monocytes (p_Mon), proportion de granulocytes (p_N_Gr) et acides gras libres (AGL) pour ces individus et la valeur élevée de propor-



(a) Composantes 1 et 2 (b) Composantes 2 et 3

FIGURE 3.7 – Représentation longitudinale des tranches sur les 3 premières composantes de la multiway-SIR (H = 5). L'échelle de couleur représente les tranches de valeurs du cortisol à $t = +1$, les valeurs allant du plus faible (bleu le plus foncé) au plus fort (bleu le plus clair). Les formes représentent les pas de temps.

tion de lymphocytes (p_Lym) sont donc une caractéristique basale ($t = 0$ et $t = +24$) de l'individu qui a tendance à s' estomper au cours de l'expérience ACTH.
Sur l'axe 2, ce sont les pas de temps $t = 0$ et $t = +1$ qui sont particulièrement corrélés chez les individus à faible réponse du cortisol, avec l'axe ainsi que le pas de temps $t = +4$ (juste avant le retour au niveau basal) des individus ayant la réponse la plus forte du cortisol. Les valeurs faibles de proportion de lymphocytes (p_Lym), hémoglobine (Hgb), plaquettes (Plt) et acides gras libres (AGL) pour les individus dont la réponse du cortisol est la plus faible sont donc des valeurs basales ($t = 0$) et de réponse immédiate ($t = +1$) qui semblent être perturbées (augmentée) pour les individus qui ont une réponse du cortisol très faible. Les individus ayant la plus forte réponse du cortisol ont, quant à eux, tendance à passer par une phase lors de laquelle ces variables chutent fortement (à $t = +4$) juste avant de revenir à la normale. Il est à noter que les individus ayant la deuxième plus forte réaction du cortisol (à gauche de l'axe 1) ont à nouveau un comportement qui semble s'opposer aux individus ayant la plus forte réaction du cortisol (à droite de l'axe 1) car la valeur des variables proportion de lymphocytes (p_Lym), hémoglobine (Hgb), plaquettes (Plt) et acides gras libres (AGL) semble augmenter au cours de l'expérience ACTH avant de revenir à un niveau basal.
Sur l'axe 3, les temps basaux ($t = 0$) sont en général mal représenté : cet

axe caractérise donc plus les différences de réaction entre individus à réponse du cortisol faible/forte pendant l'expérience que leur niveau initial. En particulier, les variables hémoglobine (Hgb), globules rouges (GR), indice de distribution des globules rouges (IDR_SD), globules blancs (GB), glucose (Gluc) et volume plaquétaire moyen (VMP) ont des valeurs élevées pour les individus à plus faible réaction du cortisol mais uniquement aux temps $t = +4$ et $t = +24$ et elles ont des valeurs faibles pour les individus à plus forte réaction du cortisol mais uniquement au temps $t = +4$ et aux temps $t = +4$ et $t = +24$ pour ceux de la deuxième tranche la plus forte.

Compte tenu du nombre de variables et de pas de temps (14×4), les figures des positions longitudinales des variables seraient difficiles à lire car les points se superposent. Afin de montrer cette sortie de façon à ce qu'elle soit lisible, nous avons choisi de ne projeter que quelques variables particulièrement parlantes pour l'analyse sur le premier plan de la multiway-SIR. La figure 3.8 donne les positions longitudinales du volume globulaire moyen (VGM) et des acides gras libres (AGL) sur le cercle des corrélations.

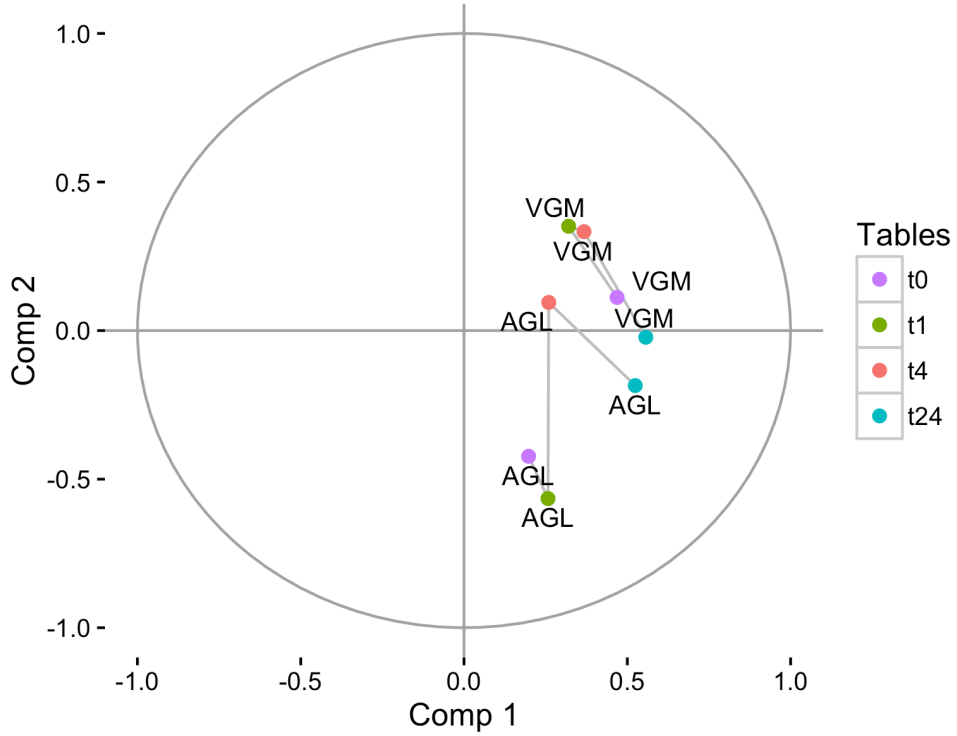


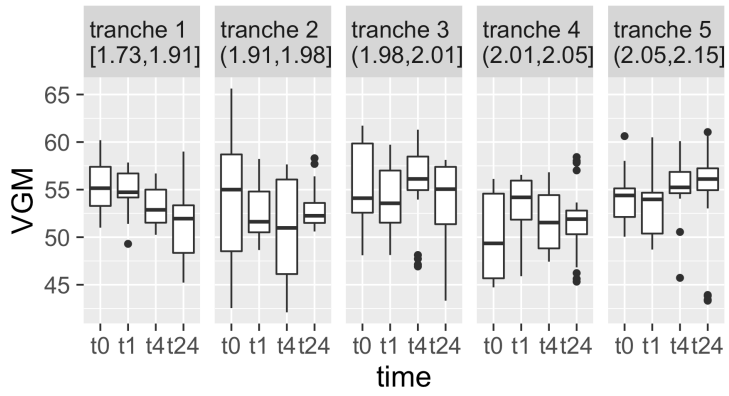
FIGURE 3.8 – Représentation longitudinale des variables VGM et AGL sur les 2 premières composantes de la multiway-SIR (H = 5). Les couleurs représentent les différents pas de temps.

Le volume globulaire moyen (VGM) contribue fortement et de manière positive à l'axe 1 aux pas de temps 0 et 24. Comme précisé précédemment,

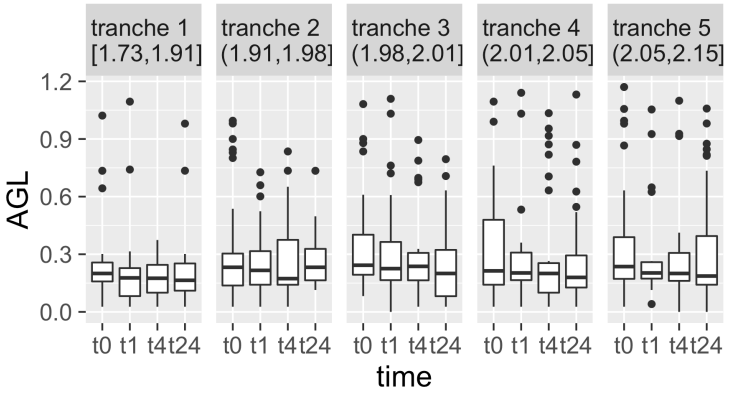
l'axe 1 oppose la tranche avec la plus forte réponse du cortisol (tranche 5), à tous les pas de temps, à la tranche avec la deuxième plus forte réponse du cortisol (tranche 4), aux pas de temps $t = 0$ et $t = +24$. Cette observation suggère que les valeurs basales de VGM ($t = 0$ et $t = +24$) sont plus élevées chez les individus de la tranche 5 que chez individus de la tranche 4. La distribution par tranche et par pas de temps de VGM en figure 3.9a permet de confirmer cette observation.
Les acides gras libres (AGL) contribuent modérément à l'axe 1 de façon positive au temps $t = +24$ et fortement à l'axe 2 de façon négative aux temps $t = 0$ et $t = +1$. L'axe 2 oppose la tranche avec la tranche la plus faible (tranche 1) aux pas de temps $t = 0$ et $t = +1$, à la tranche 4 au pas de temps$t = +1$. On peut en déduire que les AGL sont en moyenne plus élevés dans la tranche 4 que dans la tranche 1 au pas de temps 1 et en moyenne légèrement plus élevés dans la tranche 5 que dans la tranche 4 au pas de temps $t = +24$. Ces observations sont aussi confirmées par la distribution des AGL par tranche et pas de temps en figure 3.9b.

Pour aider à l'interprétation des positions longitudinales de l'ensemble des variables, on peut extraire les composantes temporelles pour les variables présentant les plus fortes corrélations avec les premières composantes de la multiway-SIR. Celles ci sont montrées dans le tableau 3.3.

Le tableau 3.3a montre, comme observé plus haut, que les valeurs basales (à $t = 0$ et $t = +24$) contribuent le plus au fait que les variables soient corrélées fortement à la première composante. Le tableau 3.3b montre que la corrélation négative de la proportion de lymphocytes (p_Lym), l'hémoglobine (Hgb), les plaquettes (Plt) et des acides gras libres (AGL) à l'axe 2 est due principalement à la contribution du pas de temps $t = +1$. On remarque aussi une contribution importante de la proportion de granulocytes (p_N_Gr) à $t = +1$ (corrélation positive), du volume plaquétaire moyen (VMP) à $t = +1$ (corrélation négative) et des acides gras libres (AGL) à $t = 0$ (corrélation négative). À ces pas de temps, ces variables pourraient caractériser les individus ayant la réponse la plus faible du cortisol. Enfin, le tableau 3.3c montre que les corrélations sur l'axe 3 sont caractérisées par une forte contribution positive des globules blancs (GB), de l'hémoglobine (Hgb), du volume plaquétaire moyen (VMP), du glucose (Gluc) et des acides gras libres (AGL) à $t = +4$ et des globules blancs (GB) et de l'hémoglobine (Hgb) à $t = +24$ ainsi que par une forte contribution négative de la proportion de granulocytes (p_N_Gr) à $t = +4$ et de la proportion de lymphocytes (p_Lym) et des plaquettes (Plt) à $t = +24$. À ces pas de temps, ces variables pourraient caractériser la discrimination entre individus à réponse faible et individus à réponse forte du cortisol.



(a) Volume globulaire moyen (VGM)



(b) Acides gras libres (AGL)

FIGURE 3.9 – Distribution par tranche et par pas de temps de VGM et AGL.

TABLEAU 3.3 – Tableaux des variables présentant les plus fortes corrélations (v_{jk}^t) avec les 3 premières composantes temporelles, où $v_{jk}^t = \frac{\langle \widetilde{G}_{.j.}^*, P_{.k} \rangle_{\mathbf{M}}}{\hat{\sigma}_j^t}$ est la corrélation entre la variable j de la matrice d'espérance conditionnelle modifiée et la $k^{\text{ième}}$ composante principale.

Positives			Négatives		
var	temps	v_{j1}^t	var	temps	v_{j1}^t
p_Mon	$t=0$	0.49	p_Lym	$t=+24$	-0.43
p_N_Gr	$t=0$	0.45			
VGM	$t=0$	0.47			
Plt	$t=0$	0.44			
p_Mon	$t=+24$	0.41			
p_N_Gr	$t=+24$	0.51			
GR	$t=+24$	0.45			
VGM	$t=+24$	0.56			
AGL	$t=+24$	0.52			

(a) Composante 1

Positives			Négatives		
var	temps	v_{j2}^t	var	temps	v_{j2}^t
VMP	$t=+1$	0.52	AGL	$t=0$	-0.42
			p_Lym	$t=+1$	-0.52
			p_N_Gr	$t=+1$	-0.55
			Hgb	$t=+1$	-0.58
			Plt	$t=+1$	-0.48
			AGL	$t=+1$	-0.57
			GR	$t=0$	-0.40

(b) Composante 2

Positives			Négatives		
var	temps	v_{j3}^t	var	temps	v_{j3}^t
GB	$t=+4$	0.59	p_N_Gr	$t=+4$	-0.43
Hgb	$t=+4$	0.46	p_Lym	$t=+24$	-0.41
VMP	$t=+4$	0.55	Plt	$t=+24$	-0.52
Gluc	$t=+4$	0.65			
AGL	$t=+4$	0.44			
GB	$t=+24$	0.52			
Hgb	$t=+24$	0.40			

(c) Composante 3

3.5 Comparaison avec l'approche dual-STATIS

Dans cette dernière section, nous comparons rapidement les résultats obtenus par la multiway-SIR avec ceux que l'on obtiendrait avec l'approche classique de dual-STATIS. Cette approche explore un jeu de données cubique de façon non supervisée. Les variables utilisées sont donc les mêmes que pour l'analyse précédente par multiway-SIR, à l'exception du cortisol à $t = +1$ qui n'est plus une variable d'intérêt. Les tableaux $\mathbf{X}_{..t}$ sont centrés et réduits (par pas de temps) pour ne pas tenir compte des différences d'échelles entre les variables. D'autre part, les individus ont été numérotés en fonction du rang de leur mesure de cortisol à $t = +1$.

3.5.1 Étude de l'inter-structure

L'inter-structure est analysée dans cette première section. La matrice des corrélation inter-pas de temps est fournie dans le tableau 3.4. Comme les tableaux sont centrés et réduits, leurs matrices de corrélation Γ_t sont des matrices identités et les matrices de covariance analysées $\widetilde{\Gamma}_t$ sont donc très proches les unes des autres. La décomposition du compromis construit avec ces matrices de covariances donne donc des poids optimaux pour les pas de temps qui sont très homogènes. L'approche multiway-SIR permettait de mettre en valeur une évolution dans la structure de corrélation des variables. Avec dual-STATIS, cette évolution semble presque inexistante au niveau individuel. Cela signifie que les structures de corrélation du jeu de données dépend en réalité de la classe de réponse au stress de l'animal (forte ou faible mesure du cortisol à $t = +1$).

TABLEAU 3.4 – Matrice des corrélations entre pas de temps.

temps	$t = 0$	$t = +1$	$t = +4$	$t = +24$
$t = 0$	1.000	0.954	0.942	0.951
$t = +1$		1.000	0.942	0.934
$t = +4$			1.00	0.914
$t = +24$				1.00

3.5.2 Étude de l'intra-structure

Choix de la dimension de projection

L'éboulis des valeurs singulières ainsi que le pourcentage de variance reproduite sont représentés dans la figure 3.10. À la vue de ce graphique, nous choisissons 3 axes pour l'interprétation ce qui représente un pourcentage d'inertie reproduite de 81,2%.

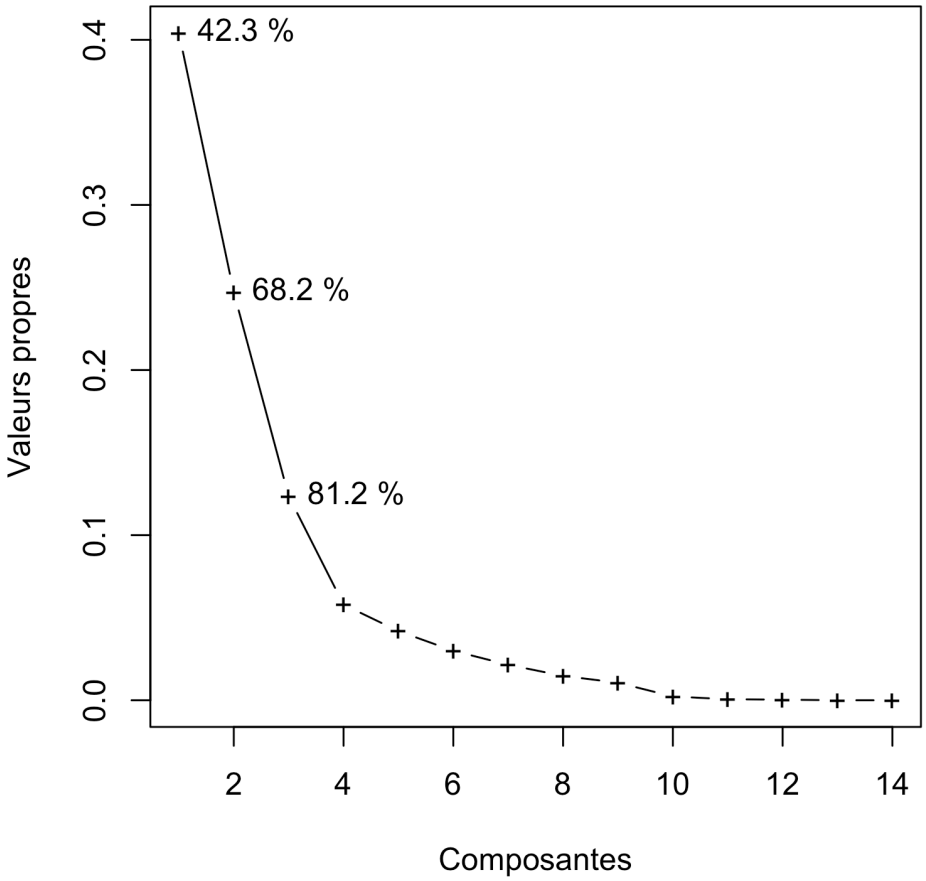


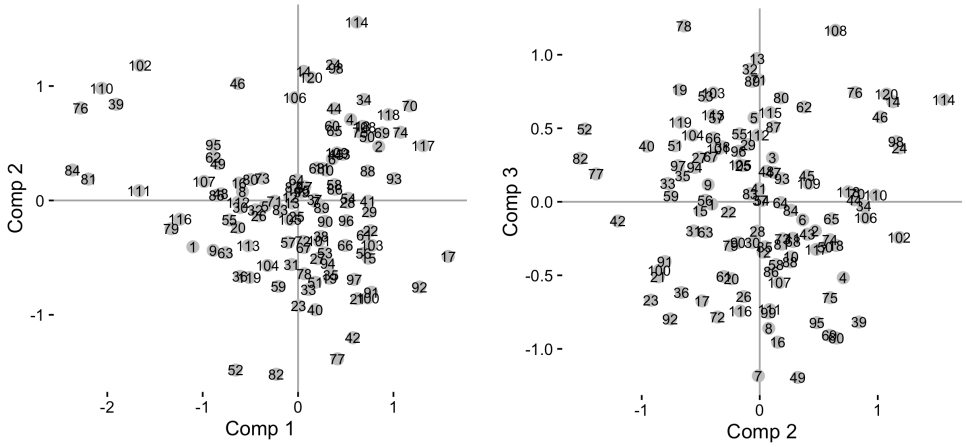
FIGURE 3.10 – Éboulis des valeurs propres et pourcentage de variance reproduite en fonction du nombre de composantes conservées dans l'approche dual-STATIS.

Représentation des individus et des variables en positions compromis

La figure 3.11 présente les positions compromis des individus obtenus sur les deux premiers plans de projection avec la méthode dual-STATIS. Les points sont numérotés en fonction du rang de la valeur du cortisol à $t = +1$.

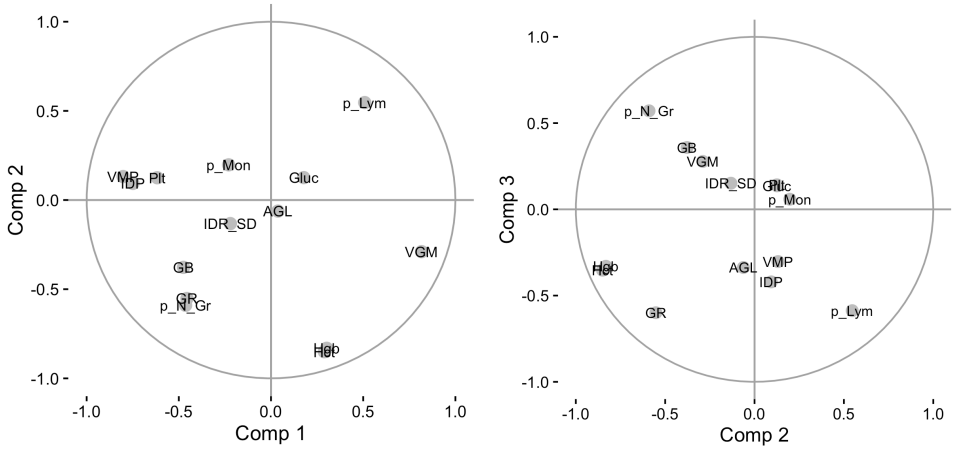
On observe que quelque soit le plan, les points ne sont pas répartis en fonction de leur rang. La source de variabilité principale identifiée avec dual-STATIS n'est donc pas liée à notre variable d'intérêt principale, contrairement à ce qui a été obtenu en multiway-SIR.

La figure 3.12 présente les positions compromis des variables sur les 2 premiers plans de projection. Le premier axe oppose le volume globulaire moyen (VGM) et la proportion de lymphocytes (p_Lym) (à droite) au volume plaquettaire moyen (VMP) et à l'indice de distribution des plaquettes (IDP) (à gauche). C'est une description des individus très différentes que



(a) Composantes 1 et 2 (b) Composantes 2 et 3

FIGURE 3.11 – Représentation des individus en positions compromis sur les 3 premières composantes de dual-STATIS. Les numéros identifiant les individus correspondent au rang de leur mesure de cortisol à $t = +1$.



(a) Composantes 1 et 2 (b) Composantes 2 et 3

FIGURE 3.12 – Représentation des variables en positions compromis sur les 3 premières composantes de dual-STATIS.

celle observée conditionnellement à la mesure du cortisol à $t = +1$ (pour rappel : les variables VGM et p_Lym étaient opposées sur l'axe 1). Or, effectivement, ces deux variables ont des évolutions très contrastées selon que les individus ont une réponse faible au stress (VGM et p_Lym en baisse constante) ou bien une réponse forte (VGM en hausse et p_Lym en baisse puis hausse).
Le second axe est caractérisé par un taux élevé des variables hématocrite

(Hct) et hémoglobine (Hgb) (en bas) qui sont deux variables qui n'apparaissent pas dans l'analyse en multiway-SIR.
Enfin, le troisième axe montre une opposition entre la proportion de granulocyte (p_N_Gr) (à droite) et la proportion de lymphocytes (p_Lym) et les globules rouges (GR) (à gauche). Là encore, il s'agit d'une opposition qui est visible à l'échelle de l'ensemble des individus et ne reflète pas un contraste entre les tranches.

Les graphiques des individus et des variables ne sont pas très lisibles dans ce cas-ci (à cause du très grand nombre de variables et d'individus pour les 4 pas de temps) et n'ont donc pas été présentés. Il apparaissait cependant nettement sur la projection des variables que les résultats de dual-STATIS n'incorporaient pas d'information temporelle : les pas de temps n'étaient pas organisés d'une manière particulière selon les axes et les projection des variables dans les différents pas de temps étaient très consensuelles.

3.6 Conclusion

Ce chapitre nous a permis de présenter la multiway-SIR, une nouvelle approche pour l'intégration de données multivariées longitudinales avec une variable cible réelle non répétée. Cette approche se base sur la méthode dual-STATIS qui étudie l'évolution des structures de corrélations entre variables répétées et la SIR qui est une méthode de régression semi-paramétrique. La multiway-SIR permet donc de combiner les domaines d'application de ces deux approches et d'expliquer une variable unidimensionnelle Y non répétée en fonction d'une variable multivariée X répétée dans le temps. Elle se base sur le principe de la régression inverse et utilise une estimation des matrices d'espérances conditionnelles $\mathbb{E}(\mathbf{X}_{..t}|y)$ pour construire une matrice de covariance compromis qui capture les différences stables au cours du temps entre les structures de corrélations entre variables en tenant compte de la variable cible. L'ACP de cette matrice de covariance compromis est le cœur de la multiway-SIR et permet d'obtenir des positions compromis donnant les relations globales au cours du temps entre variables et entre tranches d'individus conditionnellement à la variable cible, ainsi que les relations pas de temps par pas de temps pour permettre une analyse plus fine de l'évolution des structures de corrélations.

La méthode a été illustrée sur le jeu de variables cliniques de l'expérience d'injection d'ACTH du projet SusoStress. La mesure du cortisol à $t = +1$ après injection d'ACTH a été utilisée comme variable cible réelle non répétée. L'objectif était de pouvoir étudier l'évolution des structures de corrélation entre variables cliniques en tenant compte de la variabilité

de l'activité de l'axe corticotrope.
Sur ces données, nous avons pu :

1. montrer que les structures de corrélation entre variables dépendaient de l'intensité de la réponse du cortisol à $t = +1$ après injection d'ACTH et qu'elles évoluaient au cours du temps ;
2. caractériser les différences entre les forts et les faibles répondeurs : celles-ci sont principalement caractéristiques de la réaction immédiate à longue à l'injection d'ACTH (valeurs à $t = +1, +4, +24$), et non au niveau des valeurs initiales ($t = 0$) des variables liées aux globules rouges (globules rouges (GR), indice de distribution des globules rouges (IDR_SD), des globules blancs (GB), du volume plaquétaire moyen (VMP) et du glucose (Gluc). Ces variables sont moins élevées chez les individus ayant une forte activité de l'axe corticotrope que chez les individus ayant une faible activité de l'axe corticotrope ;
3. caractériser la différence entre les plus forts répondeurs et ceux répondant un peu moins fortement : les individus avec la plus forte activité de l'axe corticotrope ont des valeurs particulièrement élevées de volume globulaire moyen (VGM), de proportion de monocytes et de granulocytes (p_Mon et p_N_Gr) et d'acides gras libres (AGL) et une valeur faible de proportion de lymphocytes (p_Lym) à tous temps, tandis que les individus répondant un peu moins fortement ne sont caractérisés par ces variables (faible valeur de VGM, p_Mon, p_N_Gr, AGL et forte valeur de p_Lym) uniquement au niveau basal ($t = 0$ et $t = +24$), cet effet semblant s'estomper en réaction à l'injection d'ACTH ;
4. caractériser les plus faibles répondeurs par des valeurs basales ($t = 0$) et de réponse immédiate ($t = +1$) à l'injection d'ACTH particulièrement élevées pour la proportion de lymphocytes (p_Lym), l'hémoglobine (Hgb), les plaquettes (Plt) et les acides gras libres (AGL).

Une approche non supervisée, telle que dual-STATIS ne permettait pas de faire ressortir notre variable cible : la réponse du cortisol à $t = +1$. Grâce à notre approche qui tient compte de celle-ci, nous avons pu capturer des structures de corrélation qui évoluaient au cours du temps et étaient dépendantes du niveau du cortisol à $t = +1$.

Plusieurs points de développements sont envisagés pour la suite de ce travail :

— d'une part, la méthode n'est actuellement applicable que dans le cas des petites dimensions où $n > p$. Or, on s'intéresse à l'intégration de données 'omiques, ce qui implique également l'étude de données transcriptomiques pour lesquelles on a bien souvent $n \ll p$. La prochaine étape serait donc de développer une approche parcimonieuse

en introduisant un paramètre de pénalisation de type LASSO ou de régularisation de type ridge ;

— d'autre part, on souhaiterait pouvoir se servir de l'espace EDR estimé pour réaliser de la prédiction d'une variable cible à partir de variables longitudinales. Cela nécessite de revenir au cadre complet de la SIR tel que décrit dans l'équation (3.1) et d'estimer la fonction f de manière non paramétrique (par SVM ou bien regression à noyau type Nadaraya-Watson [NADARAYA, 1964; WATSON, 1964]).

Enfin, la méthode est en cours d'implémentation dans le package R **SirStatis**.

Chapitre 4

Discussion générale

4.1 Conclusions générale

Cette thèse entre dans le contexte de l'étude du « stress » qui est défini comme la réponse non spécifique des organismes à toute stimulation. Chez les espèces animales vertébrées destinées à l'alimentation humaine, l'axe corticotrope est le plus important système neuroendocrinien de réponse au stress. De grandes variations individuelles d'origine génétique ont été décrites dans l'activité de l'axe corticotrope avec des conséquences physiopathologiques importantes. En termes de production animale, des niveaux plus élevés de cortisol ont des effets négatifs sur la croissance et l'efficacité alimentaire et augmentent le ratio gras/maigre des carcasses. Au contraire, le cortisol a des effets positifs sur les caractères liés à la robustesse et à l'adaptation. La sélection intense pour la croissance des tissus maigres durant les dernières décennies a concomitamment réduit la production de cortisol, et nous faisons l'hypothèse que cette réduction peut être partiellement responsable des effets négatifs de la sélection sur les caractères de robustesse.

Ce travail a porté sur l'analyse des données issues du projet SUSoSTRESS, projet financé par l'ANR visant à étudier la variabilité génétique de l'axe corticotrope et de son activité physiologique en lien avec les performances des animaux sur les caractères de robustesse et production. Le protocole expérimental du projet fait intervenir des données issues de 3 expériences (ACTH, contrainte, LPS) et collectées à trois niveaux biologiques. De plus, toutes les données ont été collectées à plusieurs pas de temps. L'objectif principal de la thèse était de développer un modèle fonctionnel permettant de décrire et d'intégrer au mieux l'ensemble des sources de variation génétique du fonctionnement de l'axe corticotrope et plus généralement des réponses de stress dans notre population porcine d'étude

(race Large White). Ce premier objectif principal pouvait être décomposé en quatre sous-objectifs :

— intégrer des données de haute dimension à des niveaux différents de l'analyse biologique (biologie clinique, métabolome, transcriptome) ;

— tenir compte de l'aspect longitudinal des données ;

— extraire un sous-ensemble de gènes différentiellement exprimés lors des réponses au stress ;

— mettre les données en relation avec un caractère cible d'intérêt principal : la mesure du cortisol en réponse à une injection d'ACTH qui représente le niveau d'activité de l'axe corticotrope.

Ces objectifs ont été en partie remplis.

Concernant l'intégration de données de haute dimension, il s'agissait d'analyser de manière conjointe les données de biologique clinique, de transcriptome et de métabolome issus des 3 expériences du projet portant sur la population d'étude. Il y avait donc deux niveaux d'intégration souhaités : l'intégration des données issus des différents tissus et l'intégration des données entre expériences.
Cet objectif n'est qu'en partie réalisé. En effet, au cours de cette thèse, seules les données des expériences d'ACTH et de LPS ont pu être analysées de façon poussée. Les deux expériences ont été analysées de façon séparées et dans les deux cas, seules les données cliniques et transcriptomiques ont pu être traitées. Les difficultés rencontrées pour ce travail sont liées à la nature des données étudiées. D'une part, les données étaient répétées. L'observation de biais techniques et leurs corrections a nécessité une réflexion particulière pour normaliser ces données sans effacer la cinétique temporelle. D'autre part, les données transcriptomiques utilisées étant issues du sang total, il a fallu tenir compte de la composition sanguine lors de la recherche de gènes différentiellement exprimés lors des réponses de stress. En effet, les globules blancs sont les seules cellules à porter de l'information génétique dans le sang. Or, l'exploration des variables de formule sanguine ont montré que les sous-populations de globules blancs voyaient leur ratio s'inverser en réponse aux injections d'ACTH et de LPS. Des cellules différentes exprimant des gènes différents, il était nécessaire de tenir compte de la composition sanguine pour ces recherches. Dans le cas de l'expérience d'ACTH, le ratio L/G a été utilisé en tant que variable d'ajustement dans les recherches de gènes différentiellement exprimés lors de la réponse à l'injection. Dans le cas du LPS, une liste restreinte de gènes a pu être obtenue en recherchant les gènes différentiellement exprimés au cours du temps et pour lesquels l'effet du ratio L/G sur leur expression était lui même différent au cours du temps.

D'autre part, une partie de la thèse a été consacrée à l'analyse infructueuse des données transcriptomique de l'expérience de contrainte de 10 minutes. Les pas de temps analysés pour le transcriptome de cette expérience étaient $t = 0$, $t = +1$, $t = +4$ et $t = +24$. La réponse du cortisol à ce stress atteignait un pic à 10 minutes, mais cette réponse était de faible amplitude en comparaison de celles obtenues dans les expériences d'injection d'ACTH et de LPS. De plus, le cortisol était revenu à un niveau basal dès $t = +4$. Nous faisons l'hypothèse que les faibles intensité et durée de la réponse du cortisol expliquent qu'aucun gène n'ait été trouvé différentiellement exprimé à l'analyse. C'est pourquoi, les données de contraintes ont été mises de côté pour le reste de ce travail.

En ce qui concerne l'aspect longitudinal des données, sa prise en compte a été réalisée en prétraitant les données par la méthode multi-niveaux. Cette approche, en extrayant la matrice des variations intra-individus, permet de réaligner les observations de tous les individus sur un même centre de gravité, sans modifier leurs profils d'évolution au cours du temps. De cette manière, il devient plus facile d'étudier les effets des différentes stimulations au cours du temps sans être influencés par la variation du niveau basal des variables entre individus. Grâce à l'approche multi-niveaux, il a été possible d'appliquer ensuite les méthodes d'analyses multivariées plus classiques pour l'analyse de nos données : ACP, régression PLS, AFM. Dans ces analyses, le temps est alors traité comme une variable constituée de différents groupes et peut être utilisée comme variable supplémentaire lors de la représentation des individus.

Enfin, concernant les deux derniers points (la recherche de gènes différentiellement exprimés et la relation au cortisol), ces objectifs n'ont pu être atteints que dans l'expérience d'ACTH. L'intégration des données cliniques et transcriptomiques dans l'expérience d'ACTH ont permis de mettre en évidence un ensemble de 65 gènes uniques différentiellement exprimés dans le sang total lors de la réponse à l'injection d'ACTH. Parmi eux, 8 gènes en particulier ont été identifiés dans un réseau bibliographique comme étant en liaison avec le gène *NR3C1*, le gène codant pour le récepteur aux glucocorticoïdes. *NR3C1* est un facteur de transcription des gènes répondant au cortisol et joue aussi le rôle de régulateur à d'autres facteurs de transcription. Ces 8 gènes sont donc des candidats potentiels en tant que biomarqueurs de l'activité de l'axe corticotrope et peuvent être envisagés en tant que candidats pour de prochaines études des mécanismes d'adaptation chez les animaux d'élevage. En outre, la connaissance de ces gènes par l'étude du transcriptome du sang total encourage l'utilisation de ce tissu pour de futures études car il permet de mesurer l'activité de l'axe corticotrope au niveau des gènes, tout en étant

facile d'obtention sans nécessiter l'euthanasie des animaux. Il permet donc les études longitudinales ou populationnelles. En ce qui concerne l'expérience de LPS, le nombre de gènes différentiellement exprimés était considérable et la mise en relation avec les données de biologie clinique par AFM n'ont pas permis de mettre en évidence de sous-ensemble de gènes leur étant plus particulièrement liés.

Au cours de la thèse, nous avions également un second objectif principal : le développement d'un outil d'analyse plus adapté pour intégrer des données cubiques à une variable cible réelle. Cet objectif a été en partie rempli. A cette fin, nous avons développé la méthode de « multiway-SIR » qui étend la méthode dual-STATIS, une méthode d'analyse de données cubiques non supervisée, au cadre de la SIR, une méthode de régression semi-paramétrique pouvant être utilisée à des fins exploratoires. Cette méthode a été appliquée aux données cliniques de l'expérience d'ACTH en se servant de la valeur du cortisol à $t = +1$ (le pic de l'activité de l'axe corticotrope suite à une injection d'ACTH) comme variable cible à expliquer. L'exploration des données avec cette méthode a permis de mettre en évidence plusieurs faits marquants. D'une part, la structure des corrélations entre variables cliniques dépend de l'intensité de la réponse du cortisol à $t = +1$ et change au cours du temps. D'autre part, il est possible de caractériser les forts et les faibles répondeurs de la population G0 (sur la réponse du cortisol) grâce aux variables cliniques. Par exemple, les forts et les faibles répondeurs sont caractérisés par des différences de niveaux de glucose et de globules rouges lors de la réponse à l'injection ($t = +1, +4, +24$), mais pas au niveau basal ($t = 0$).
La comparaison avec une approche dual-STATIS classique a montré que cette dernière ne permettait pas de mettre en évidence un lien entre les structures de corrélations des variables cliniques et le cortisol à $t = +1$. Cette nouvelle approche est donc satisfaisante.

Il est à noter que pour la prise en compte de la nature longitudinale des données, le travail réalisé avec l'approche multi-niveaux dans le chapitre 2 et celui réalisé avec la multiway-SIR en chapitre 3 présentent la même limite : la continuité entre pas de temps est perdue. En effet, soit le temps est traité comme une variable à plusieurs facteurs dans l'approche multi-niveaux, soit les données aux différents pas de temps sont considérées comme des tableaux d'observations indépendants avec la multiway-SIR qui est dérivée de l'approche dual-STATIS. Ni l'information d'ordre, ni celle sur la proximité entre les pas de temps (par exemple : $t = +4$ est plus proche de $t = +1$ que de $t = +24$) ne sont prises en compte par ces approches. Les deux approches présentent cependant l'intérêt de pouvoir être appliquées

lorsque le nombre de pas de temps à étudier est faible, ce qui est le cas dans cette thèse.

4.2 Perspectives

Ce travail de thèse a apporté des contributions à la fois pour la recherche de biomarqueurs de l'activité de l'axe corticotrope dans le sang total et pour le développement de nouveaux outils méthodologiques pour l'intégration de données biologiques. Plusieurs points de développement sont cependant envisageables pour ce travail.

Dans un premier temps, il est nécessaire de poursuivre les analyses effectuées en faisant l'analyse des données métabolomiques puis de les intégrer avec les données cliniques et transcriptomiques pour les 3 expériences. Cette analyse pourrait être réalisée par une AFM multi-niveaux faisant intervenir les 3 types de données et tenant compte du temps. Il faudrait ensuite s'attacher à étudier les possibilités envisageables pour une méta-intégration faisant intervenir les 3 types de données issues des 3 expériences.

Ce travail est également à remettre en perspective du projet SUSoSTRESS. En effet, la population d'étude de cette thèse est la population de départ d'une expérience de sélection divergente basée sur l'intensité de l'activité de l'axe corticotrope. Cette sélection divergente est conduite sur 4 générations successives : de la génération 0 (G0 ; notre population d'étude) à la génération 3 (G3). Cette sélection divergente a permis d'obtenir deux sous-populations dont l'intensité de la réponse du cortisol à une injection d'ACTH était soit très forte (H), soit très faible (L). Cette sélection a eu lieu en parallèle de cette thèse et les données issues du phénotypage des individus en G3 sont à présent disponibles. Les 65 gènes différentiellement exprimés lors de la réponse de l'axe corticotrope à l'injection d'ACTH sont de potentiels biomarqueurs à l'activité de l'axe corticotrope. Il serait donc pertinent de comparer leurs niveaux d'expressions chez les (H) et les (L) pour confirmer ce rôle.

Concernant l'analyse des données transcriptomiques, il parait important de se pencher sur les questions de prise en compte de la composition sanguine. Cette question pose un véritable défi méthodologique, car sans en tenir compte, cela revient à comparer deux populations de cellules totalement différentes lorsque l'on compare les gènes exprimés à 2 pas de temps différents. Notre approche faisant intervenir le ratio lymphocytes/granulocytes (L/G) en tant que variable d'ajustement permet

d'apporter un facteur de correction, mais elle reste insuffisante.

Enfin, concernant la multiway-SIR, plusieurs points de développement sont envisageables. D'une part, on souhaite pouvoir étendre l'approche aux données de haute dimension afin de pouvoir intégrer des données transcriptomiques pour lesquelles on a n individus pour p variables avec $n \ll p$. Pour y parvenir, il serait judicieux d'intégrer des paramètres de régularisation type ridge, ou de pénalisation de type LASSO dans la méthode. D'autre part, on souhaiterait pouvoir se servir de la méthode pour réaliser de la prédiction d'une variable cible à partir de données cubiques.

4.3 Valorisation du travail

Le travail présenté dans cette thèse a été valorisé et a fait l'objet de plusieurs communications en congrès nationaux et internationaux. Le contexte général de la thèse et ses premiers développements ont été présentés lors de la Journée Régionale GenoToul Bioinfo/Biostats 2014 (poster), au séminaire des doctorants du département de Génétique Animale de l'INRA en 2014 (poster), et à l'*European Conference on Computational Biology* 2014 (poster). Les résultats du chapitre 2 ont fait l'objet de deux articles. Le premier, « Time course of the response to ACTH in pig : biological and transcriptomic study », a été accepté dans une revue à comité de lecture et ses résultats ont été présentés au cours du séminaire des doctorants SEVAB 2015 (oral) et de l'*International Society for Animal Genetics Conference* 2016 (e-poster). Le deuxième, « Time course study of the response to LPS targeting the pig immune response gene networks » est en cours de soumission. Enfin, le développement méthodologique du chapitre 3 fait l'objet d'un article en cours d'écriture et est en cours d'implémentation dans R dans le package **SirStatis**. Il a été présenté au cours de *The 2016 annual workshop on Statistical Methods for Post Genomic Data* 2016 (poster) et de la *22nd International Conference on Computational Statistics* (COMPSTAT) 2016 (oral).

Bibliographie

ABDI, H., L. J. WILLIAMS et D. VALENTIN. 2013, «Multiple factor analysis : principal component analysis for multitable and multiblock data sets», *Wiley Interdisciplinary reviews : computational statistics*, vol. 5, nº 2, p. 149–179. 39, 96

ABDI, H., L. J. WILLIAMS, D. VALENTIN et M. BENNANI-DOSSE. 2012, «Statis and distatis : optimum multitable principal component analysis and three way metric multidimensional scaling», *Wiley Interdisciplinary Reviews : Computational Statistics*, vol. 4, nº 2, p. 124–167. 35

ADCOCK, I. M. 2000, «Molecular mechanisms of glucocorticosteroid actions», *Pulmonary pharmacology & therapeutics*, vol. 13, nº 3, p. 115–126. 17

BARNES, P. J. 2006, «Corticosteroids : the drugs to beat», *European journal of pharmacology*, vol. 533, nº 1, p. 2–14. 18

BEILHARZ, R. 1998, «Environmental limit to genetic change. an alternative theorem of natural selection», *Journal of Animal Breeding and Genetics*, vol. 115, nº 1-6, p. 433–437. 24

BERTAGNA, X., J. COSTE, M. RAUX-DEMAY, M. LETRAIT et G. STRAUCH. 1994, «The combined corticotropin-releasing hormone/lysine vasopressin test discloses a corticotroph phenotype.», *The Journal of Clinical Endocrinology & Metabolism*, vol. 79, nº 2, p. 390–394. 21

BIDANEL, J. P. 1993, «Estimation of crossbreeding parameters between large white and meishan porcine breeds. iii. dominance and epistatic components of heterosis on reproductive traits», *Genetics Selection Evolution*, vol. 25, nº 3, p. 1. 21

BIDANEL, J. P., J. C. CARITEZ, J. GRUAND et C. LEGAULT. 1993, «Growth, carcass and meat quality performance of crossbred pigs with graded pro-

portions of meishan genes», *Genetics Selection Evolution*, vol. 25, nº 1, p. 83–99. 21

BIDANEL, J. P., J. C. CARITEZ et C. LEGAULT. 1990, «Estimation of cross-breeding parameters between large white and meishan porcine breeds. ii. growth before weaning and growth of females during the growing and reproductive periods», *Genetics Selection Evolution*, vol. 22, nº 4, p. 431–445. 21

BOISSY, A., G. MANTEUFFEL, M. B. JENSEN, R. O. MOE, B. SPRUIJT, L. J. KEELING, C. WINCKLER, B. FORKMAN, I. DIMITROV, J. LANGBEIN et collab.. 2007, «Assessment of positive emotions in animals to improve their welfare», *Physiology & Behavior*, vol. 92, nº 3, p. 375–397. 12

BOISSY, A., I. VEISSIER et S. ROUSSEL. 2001, «Behavioural reactivity affected by chronic stress : an experimental approach in calves submitted to environmental instability», *Animal welfare*, vol. 10, nº 1, p. 175–185. 13

BOUROCHE, J.-M. 1975, *Analyse des données ternaires : la double analyse en composantes principales*, thèse de doctorat. 96

BRADBURY, M. J., S. F. AKANA et M. F. DALLMAN. 1994, «Roles of type i and ii corticosteroid receptors in regulation of basal activity in the hypothalamo-pituitary-adrenal axis during the diurnal trough and the peak : evidence for a nonadditive effect of combined receptor occupation.», *Endocrinology*, vol. 134, nº 3, p. 1286–1296. 22

BROOM, D. 1987, «Applications of neurobiological studies to farm animal welfare», dans *Biology of Stress in Farm Animals : An Integrative Approach*, Springer, p. 101–110. 13

BROWN, K. I. et K. E. NESTOR. 1973, «Some physiological responses of turkeys selected for high and low adrenal response to cold stress», *Poultry Science*, vol. 52, nº 5, p. 1948–1954. 22

BUREAU, C., C. HENNEQUET-ANTIER, M. COUTY et D. GUÉMENÉ. 2009, «Gene array analysis of adrenal glands in broiler chickens following acth treatment», *BMC genomics*, vol. 10, nº 1, p. 1. 22

CANARIO, L., Y. BILLON, J.-C. CARITEZ, J. P. BIDANEL et D. LALOË. 2009, «Comparison of sow farrowing characteristics between a chinese breed and three french breeds», *Livestock Science*, vol. 125, nº 2, p. 132–140. 21, 25

CANARIO, L., E. CANTONI, E. LE BIHAN, J. CARITEZ, Y. BILLON, J. BIDANEL et J. FOULLEY. 2006, «Between-breed variability of stillbirth and its relationship with sow and piglet characteristics», *Journal of animal science*, vol. 84, nº 12, p. 3185–3196. 21

CASSENS, R., D. MARPLE et G. EIKELENBOOM. 1975, «Animal physiology and meat quality», *Advances in food research*, vol. 21, p. 71–155. 11

CHAUSSABEL, D., V. PASCUAL et J. BANCHEREAU. 2010, «Assessing the human immune system through blood transcriptomics», *BMC biology*, vol. 8, nº 1, p. 1. 54

CHROUSOS, G. P. et P. W. GOLD. 1992, «The concepts of stress and stress system disorders : overview of physical and behavioral homeostasis», *Jama*, vol. 267, nº 9, p. 1244–1252. 12

CIVELEK, M. et A. J. LUSIS. 2014, «Systems genetics approaches to understand complex traits», *Nature Reviews Genetics*, vol. 15, nº 1, p. 34–48. 30

COLE, S. W. 2010, «Elevating the perspective on human stress genomics», *Psychoneuroendocrinology*, vol. 35, nº 7, p. 955–962. 54

COSTE, J., G. STRAUCH, M. LETRAIT et X. BERTAGNA. 1994, «Reliability of hormonal levels for assessing the hypothalamic-pituitary-adrenocortical system in clinical pharmacology.», *British journal of clinical pharmacology*, vol. 38, nº 5, p. 474–479. 21

CRICK, F. et collab.. 1970, «Central dogma of molecular biology», *Nature*, vol. 227, nº 5258, p. 561–563. 28

DANTZER, R. et P. MORMÈDE. 1983, «Stress in farm animals : a need for reevaluation», *Journal of Animal Science*, vol. 57, nº 1, p. 6–18. 11, 12

DAZY, F., J.-F. LE BARZIC, G. SAPORTA et F. LAVALLARD. 1996, «L'analyse des données évolutives-méthodes et applications», . 96

DE KLOET, E. R., E. VREUGDENHIL, M. S. OITZL et M. JOELS. 1998, «Brain corticosteroid receptor balance in health and disease 1», *Endocrine reviews*, vol. 19, nº 3, p. 269–301. 22

DÉJEAN, S., P. G. MARTIN, A. BACCINI et P. BESSE. 2007, «Clustering time-series gene expression data using smoothing spline derivatives», *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, nº 1, p. 1–10. 47

DÉSAUTÉS, C., J. BIDANEL, D. MILAN, N. IANNUCCELLI, Y. AMIGUES, F. BOURGEOIS, J. CARITEZ, C. RENARD, C. CHEVALET et P. MORMEDE. 2002, «Genetic linkage mapping of quantitative trait loci for behavioral and neuroendocrine stress response traits in pigs», *Journal of Animal Science*, vol. 80, nº 9, p. 2276–2285. 22, 23

DÉSAUTÉS, C., A. SARRIEAU, J.-C. CARITEZ et P. MORMÈDE. 1999, «Behavior and pituitary-adrenal function in large white and meishan pigs», *Domestic Animal Endocrinology*, vol. 16, nº 4, p. 193–205. 22

DÉSIRÉ, L., I. VEISSIER, G. DESPRÉS, E. DELVAL, G. TOPORENKO et A. BOISSY. 2006, «Appraisal process in sheep (ovis aries) : Interactive effect of suddenness and unfamiliarity on cardiac and behavioral responses.», *Journal of Comparative Psychology*, vol. 120, nº 3, p. 280. 13

DURBÁN, M., J. HAREZLAK, M. WAND et R. CARROLL. 2005, «Simple fitting of subject-specific curves for longitudinal data», *Statistics in medicine*, vol. 24, nº 8, p. 1153–1167. 47

EDENS, F. W. et H. T. SIEGEL. 1975, «Adrenal responses in high and low acth response lines of chickens during acute heat stress», *General and Comparative Endocrinology*, vol. 25, nº 1, p. 64–73. 22

ESCOFFIER, B. et J. PAGÈS. 1990, «Simple and multiple factor analyses. objectives, methods and interpretation», . 96

ESCOFFIER, R. 1976, «A unifying tool for linear multivariate statistical methods : the RV-coefficient», *Applied Statistics*, vol. 25, nº 3, doi :10.2307/2347233.JSTOR2347233, p. 257–265.

ESCOFIER, B. et J. PAGES. 1994, «Multiple factor analysis (afmult package)», *Computational statistics & data analysis*, vol. 18, nº 1, p. 121–140. 33, 38

ESSÉN-GUSTAVSSON, B. et A. LINDHOLM. 1984, «Fiber types and metabolic characteristics in muscles of wild boars, normal and halothane sensitive swedish landrace pigs», *Comparative Biochemistry and Physiology Part A : Physiology*, vol. 78, nº 1, p. 67–71. 10

FERRE, L. 1996, «Choix de dimension en regression inverse par tranches (sir)», *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, vol. 323, nº 4, p. 403–406. 105

FERRÉ, L. 1998, «Determination of the dimension choice in sir and related methods», *J. Amer. Statist. Assoc*, vol. 2, p. 109–122. 105

FOURY, A., N. GEVERINK, M. GIL, M. GISPERT, M. HORTOS, M. F. I FURNOLS, D. CARRION, S. BLOTT, G. PLASTOW et P. MORMEDE. 2007, «Stress neuroendocrine profiles in five pig breeding lines and the relationship with carcass composition», . 22

GILBERT, H., J.-P. BIDANEL, J. GRUAND, J.-C. CARITEZ, Y. BILLON, P. GUILLOUET, H. LAGANT, J. NOBLET et P. SELLIER. 2007, «Genetic parameters for residual feed intake in growing pigs, with emphasis on genetic relationships with carcass and meat quality traits», *Journal of Animal Science*, vol. 85, nº 12, p. 3182–3188. 23

GLACON, F. 1981, *Analyse conjointe de plusieurs matrices de données : Comparaison de différentes méthodes*, thèse de doctorat, Université Joseph-Fourier-Grenoble I. 33, 96

GLIGORIJEVIĆ, V. et N. PRŽULJ. 2015, «Methods for biological data integration : perspectives and challenges», *Journal of The Royal Society Interface*, vol. 12, nº 112, p. 20150 571. iii, 32

GROSS, W. et P. SIEGEL. 1985, «Selective breeding of chickens for corticosterone response to social stress», *Poultry Science*, vol. 64, nº 12, p. 2230–2233. 22

GUYONNET-DUPÉRAT, V., N. GEVERINK, G. S. PLASTOW, G. EVANS, O. OUSOVA, C. CROISETIÈRE, A. FOURY, E. RICHARD, P. MORMÈDE et M.-P. MOISAN. 2006, «Functional implication of an arg307gly substitution in corticosteroid-binding globulin, a candidate gene for a quantitative trait locus associated with cortisol variability and obesity in pig», *Genetics*, vol. 173, nº 4, p. 2143–2149. 22, 23

HAY, M. et P. MORMEDE. 1998, «Urinary excretion of catecholamines, cortisol and their metabolites in meishan and large white sows : Validation as a non-invasive and integrative assessment of adrenocortical and sympathoadrenal axis», *Veterinary research*, vol. 29, nº 2, p. 119–128. 22

HAYASHI, R., H. WADA, K. ITO et I. M. ADCOCK. 2004, «Effects of glucocorticoids on gene transcription», *European journal of pharmacology*, vol. 500, nº 1, p. 51–62. 18

HAYES, B. J., H. A. LEWIN et M. E. GODDARD. 2013, «The future of livestock breeding : genomic selection for efficiency, reduced emissions intensity, and adaptation», *Trends in Genetics*, vol. 29, nº 4, p. 206–214. 8, 9, 10

HAZARD, D., L. LIAUBET, M. SANCRISTOBAL et P. MORMÈDE. 2008, «Gene array and real time pcr analysis of the adrenal sensitivity to adrenocorticotropic hormone in pig», *BMC genomics*, vol. 9, nº 1, p. 1. 22

HENNESSY, D. et P. JACKSON. 1987, «Relationship between adrenal responsiveness and growth rate», *Manipulating Pig Production*, vol. 1. 22, 25

HENNESSY, D., T. STELMASIAK, N. JOHNSTON, P. JACKSON et K. OUTCH. 1988, «Consistent capacity for adrenocortical response to acth administration in pigs.», *American journal of veterinary research*, vol. 49, nº 8, p. 1276–1283. 21, 22

HOTELLING, H. 1936, «Relations between two sets of variates», *Biometrika*, vol. 28, nº 3/4, p. 321–377. 32

HUANG, Y.-Y., E. R. KANDEL, L. VARSHAVSKY, E. P. BRANDONT, M. QI, R. L. IDZERDA, G. S. MCKNIGHT et R. BOURTCHOULADZ. 1995, «A genetic test of the effects of mutations in pka on mossy fiber ltp and its relation to spatial and contextual learning», *Cell*, vol. 83, nº 7, p. 1211–1222. 10

HUIZENGA, N. A., J. W. KOPER, P. DE LANGE, H. A. POLS, R. P. STOLK, D. E. GROBBEE, F. H. DE JONG et S. W. LAMBERTS. 1998, «Interperson variability but intraperson stability of baseline plasma cortisol concentrations, and its relation to feedback sensitivity of the hypothalamo-pituitary-adrenal axis to a low dose of dexamethasone in elderly individuals 1», *The Journal of Clinical Endocrinology & Metabolism*, vol. 83, nº 1, p. 47–54. 21

NATIONAL DE LA SANTÉ ET DE LA RECHERCHE MÉDICALE (INSERM), I., éd.. 2011, *Stress au travail et santé. Situation chez les indépendants*, Paris :Les éditions Inserm. 21

Institut de la Filière Porcine. 2016, «Gte : Evolution des résultats moyens nationaux - post-sevreurs-engraisseurs», URL http://ifip.asso.fr/PagesStatics/resultat/pdf/retro/gte04.pdf. 9

JANSEN, J. J., H. C. HOEFSLOOT, J. VAN DER GREEF, M. E. TIMMERMAN, J. A. WESTERHUIS et A. K. SMILDE. 2005, «Asca : analysis of multivariate data obtained from an experimental design», *Journal of Chemometrics*, vol. 19, nº 9, p. 469–481. 46

JOUFFE, V., S. ROWE, L. LIAUBET, B. BUITENHUIS, H. HORNSHØJ, M. SANCRISTOBAL, P. MORMÈDE et D. DE KONING. 2009, «Using microarrays to identify positional candidate genes for qtl : the case study of acth response in pigs», dans *BMC proceedings*, vol. 3, BioMed Central, p. 1. 22

KARLOVICH, C., G. DUCHATEAU-NGUYEN, A. JOHNSON, P. MCLOUGHLIN, M. NAVARRO, C. FLEURBAEY, L. STEINER, M. TESSIER, T. NGUYEN,

M. WILHELM-SEILER et collab.. 2009, «A longitudinal study of gene expression in healthy individuals», *BMC medical genomics*, vol. 2, nº 1, p. 1. 45

KARLSTRÖM, K. 1995, *Capillary supply, fibre type composition and enzymatic profile of equine, bovine and porcine locomotor and nonlocomotor muscles*, thèse de doctorat, Swedish University of Agricultural Sciences, Uppsala, Sweden. 10

KNAP, P. 2009, «Robustness», dans *Resource allocation theory applied to farm animal production*, CABI. 10

KNAP, P., W. RAUW et collab.. 2008, «Robustness.», *Resource allocation theory applied to farm animal production*, p. 288–301. 9

KNAP, P., W. RAUW et collab.. 2009, «Selection for high production in pigs», *Resource allocation theory applied to farm animal production (ed. WM Rauw)*, p. 210–229. 11

KNAP, P. et G. SU. 2008, «Genotype by environment interaction for litter size in pigs as quantified by reaction norms analysis», . 11

KNAP, P. W. 2005, «Breeding robust pigs», *Animal Production Science*, vol. 45, nº 8, p. 763–773. 9, 24

KNOTT, S., L. CUMMINS, F. DUNSHEA et B. LEURY. 2008, «Rams with poor feed efficiency are highly responsive to an exogenous adrenocorticotropin hormone (acth) challenge», *Domestic animal endocrinology*, vol. 34, nº 3, p. 261–268. 23

KOOLHAAS, J., S. KORTE, S. DE BOER, B. VAN DER VEGT, C. VAN REENEN, H. HOPSTER, I. DE JONG, M. RUIS et H. BLOKHUIS. 1999, «Coping styles in animals : current status in behavior and stress-physiology», *Neuroscience & Biobehavioral Reviews*, vol. 23, nº 7, p. 925–935. 12

KRUSKAL, J. B. 1989, «Rank, decomposition, and uniqueness for 3-way and n-way arrays», dans *Multiway data analysis*, North-Holland Publishing Co., p. 7–18. 33, 96

LADEWIG, J. et D. SMIDT. 1989, «Behavior, episodic secretion of cortisol, and adrenocortical reactivity in bulls subjected to tethering», *Hormones and behavior*, vol. 23, nº 3, p. 344–360. 13

LARZUL, C., E. TERENINA, A. FOURY, Y. BILLON, I. LOUVEAU, E. MERLOT et P. MORMEDE. 2015, «The cortisol response to acth in pigs, heritability and influence of corticosteroid-binding globulin», *animal*, vol. 9, nº 12, p. 1929–1934. 20, 22, 26, 52

LAVIT, C. 1988, «Analyse conjointe de tableaux quantitatifs.[simultaneous analysis of several quantitative matrices]», . 33, 96

LAVIT, C., Y. ESCOUFIER, R. SABATIER et P. TRAISSAC. 1994, «The act (statis method)», *Computational Statistics & Data Analysis*, vol. 18, nº 1, p. 97–119. 33, 96

LAZARUS, R. S. 1993, «Coping theory and research : past, present, and future.», *Psychosomatic medicine*, vol. 55, nº 3, p. 234–247. 12

LÊ CAO, K.-A., D. ROSSOUW, C. ROBERT-GRANIÉ et P. BESSE. 2008, «A sparse pls for variable selection when integrating omics data», *Stat. Appl. Genet. Mol. Biol.*, vol. 7, nº 1. 43

L'HERMIER DES PLANTES, H. 1976, *Structuration des tableaux à trois indices de la statistique : théorie et application d'une méthode d'analyse conjointe*, thèse de doctorat, Université des sciences et techniques du Languedoc. 33, 35, 96

LI, K.-C. 1991, «Sliced inverse regression for dimension reduction», *Journal of the American Statistical Association*, vol. 86, nº 414, p. 316–327. 33, 43, 96, 99, 105

LIQUET, B., K.-A. LÊ CAO, H. HOCINI et R. THIÉBAUT. 2012, «A novel approach for biomarker selection and the integration of repeated measures experiments from two assays», *BMC Bioinform.*, vol. 13, nº 1, p. 325. 45, 46

LUITING, P. 1990, «Genetic variation of energy partitioning in laying hens : causes of variation in residual feed consumption», *World's Poultry Science Journal*, vol. 46, nº 02, p. 133–152. 8

LUNDEHEIM, N. 1987, «Genetic analysis of osteochondrosis and leg weakness in the swedish pig progeny testing scheme», *Acta Agriculturae Scandinavica*, vol. 37, nº 2, p. 159–173. 10

LUPIEN, S. J., M. DE LEON, S. DE SANTI, A. CONVIT, C. TARSHISH, N. P. V. NAIR, M. THAKUR, B. S. MCEWEN, R. L. HAUGER et M. J. MEANEY. 1998, «Cortisol levels during human aging predict hippocampal atrophy and memory deficits», *Nature neuroscience*, vol. 1, nº 1, p. 69–73. 20

MASON, J. W. 1971, «A re-evaluation of the concept of 'non-specificity'in stress theory.», *Journal of Psychiatric research*, vol. 8, p. 323–333. 12

MOISAN, M.-P. et M. LE MOAL. 2012, «Le stress dans tous ses états», *MS. Médecine sciences*, vol. 28, nº 6-7, p. 612–617. 12, 20

MORMÈDE, P. et A. FOURY. 2009, «Robustesse et production durable : hypothèses physiopathologiques et moléculaires», . 10, 24

MORMÈDE, P., A. FOURY, E. TERENINA et P. KNAP. 2011, «Breeding for robustness : the role of cortisol», *Animal*, vol. 5, nº 05, p. 651–657. 10

MORMEDE, P. et E. TERENINA. 2012, «Molecular genetics of the adrenocortical axis and breeding for robustness», *Domest. Anim. Endocrinol.*, vol. 43, nº 2, p. 116–131. 20, 21, 22, 25

MURÁNI, E., S. PONSUKSILI, R. B. D'EATH, S. P. TURNER, E. KURT, G. EVANS, L. THÖLKING, R. KLONT, A. FOURY, P. MORMÈDE et collab.. 2010, «Association of hpa axis-related genetic variation with stress reactivity and aggressive behaviour in pigs», *BMC genetics*, vol. 11, nº 1, p. 1. 23

NADARAYA, E. 1964, «On estimating regression», *Theory of Probability and its Applications*, vol. 10, p. 186–196.

OUSOVA, O., V. GUYONNET-DUPERAT, N. IANNUCCELLI, J.-P. BIDANEL, D. MILAN, C. GENÊT, B. LLAMAS, M. YERLE, J. GELLIN, P. CHARDON et collab.. 2004, «Corticosteroid binding globulin : a new target for cortisol-driven obesity», *Molecular Endocrinology*, vol. 18, nº 7, p. 1687–1696. 22, 23

PATTERSON, H. D. et R. THOMPSON. 1971, «Recovery of inter-block information when block sizes are unequal», *Biometrika*, vol. 58, nº 3, p. 545–554. 47

PERREAU, V., A. SARRIEAU et P. MORMÉDE. 1999, «Characterization of mineralocorticoid and glucocorticoid receptors in pigs : comparison of meishan and large white breeds», *Life sciences*, vol. 64, nº 17, p. 1501–1515. 22, 23

RAHELIC, S. et S. PUAC. 1981, «Fibre types in longissimus dorsi from wild and highly selected pig breeds», *Meat Science*, vol. 5, nº 6, p. 439–450. 10

RAO, C. 1964, «The use and interpretation of principal component analysis in applied research. sankhya», *Sankhya, Series A*, vol. 26, nº 4, p. 329–358.

RAUW, W., E. KANIS, E. NOORDHUIZEN-STASSEN et F. GROMMERS. 1998, «Undesirable side effects of selection for high production efficiency in farm animals : a review», *Livest. Prod. Sci.*, vol. 56, nº 1, p. 15–33. 8

RAUW, W. M. 2009, *Resource allocation theory applied to farm animal production*, CABI. 24

RITCHIE, M. D., E. R. HOLZINGER, R. LI, S. A. PENDERGRASS et D. KIM. 2015, «Methods of integrating data to uncover genotype-phenotype interactions», *Nature Reviews Genetics*, vol. 16, nº 2, p. 85–97. 28, 29

RONECKER, I. 2010, *Variabilité génétique des réponses de peur chez le canard*, thèse de doctorat, Tours. 22, 25

SABATIER, R., J. LEBRETON et D. CHESSEL. 1989, «Multiway data analysis», dans *Principal component analysis with instrumental variables as a tool for modelling composition data*, édité par R. Coppi et S. Bolasco, Elsevier Science Publishers, B.V., North-Holland, p. 341–352.

SARACCO, J., I. LARRAMENDY et Y. ARAGON. 1999, «La regression inverse par tranches ou méthode sir : presentation générale», *La revue de Modulad*, , nº 22, p. 21–39. 44, 105

SATHER, A. 1987, «A note on the changes in leg weakness in pigs after being transferred from confinement housing to pasture lots», *Animal Production*, vol. 44, nº 03, p. 450–453. 9

SAUTRON, V., E. TERENINA, L. GRESS, Y. LIPPI, Y. BILLON, C. LARZUL, L. LIAUBET, N. VILLA-VIALANEIX et P. MORMÈDE. 2015, «Time course of the response to acth in pig : biological and transcriptomic study», *BMC genomics*, vol. 16, nº 1, p. 1. 6

SCHERER, K. R. 2001, «Appraisal considered as a process of multilevel sequential checking», *Appraisal processes in emotion : Theory, methods, research*, vol. 92, p. 120. 12

SCHINCKEL, A. 2010a, «Modeling, management and selection of genetics for optimal commercial performance», dans *9th World Congress on Genetics Applied to Livestock Production*, vol. paper 0045, German Society for Animal Science, Geissen, Germany. 10

SCHINCKEL, A. 2010b, «Modeling, management and selection of genetics for optimal commercial performance», *German Society for Animal Science*. 24

SCHONEVELD, O. J., I. C. GAEMERS et W. H. LAMERS. 2004, «Mechanisms of glucocorticoid signalling», *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, vol. 1680, nº 2, p. 114–128. 17, 19

SCHOTT, J. R. 1994, «Determining the dimensionality in sliced inverse regression», *Journal of the American Statistical Association*, vol. 89, nº 425, p. 141–148. 105

SELYE, H. 1956, «The stress of life.», . 11

SELYE, H. 1973, «The evolution of the stress concept : The originator of the concept traces its development from the discovery in 1936 of the alarm reaction to modern therapeutic applications of syntoxic and catatoxic hormones», *American scientist*, vol. 61, nº 6, p. 692–699. 12

SELYE, H. et collab.. 1936, «A syndrome produced by diverse nocuous agents», *Nature*, vol. 138, nº 3479, p. 32. 11

SHAPIRO, J. A. 2009, «Revisiting the central dogma in the 21st century», *Annals of the New York Academy of Sciences*, vol. 1178, nº 1, p. 6–28. 28

SMILDE, A. K., J. J. JANSEN, H. C. HOEFSLOOT, R.-J. A. LAMERS, J. VAN DER GREEF et M. E. TIMMERMAN. 2005, «Anova-simultaneous component analysis (asca) : a new tool for analyzing designed metabolomics data», *Bioinformatics*, vol. 21, nº 13, p. 3043–3048. 46

STRAUBE, J., A.-D. GORSE, B. E. HUANG, K.-A. LÊ CAO et collab.. 2015, «A linear mixed model spline framework for analysing time course 'omics' data», *PloS one*, vol. 10, nº 8, p. e0134 540. 47, 95

TANAKA, K., N. SHIMIZU, H. IMURA, J. FUKATA, I. HIBI, T. TANAKA, S. NAKAGAWA, K. FUJIEDA, K. TAKEBE, K. YOSHINAGA et collab.. 1993, «Human corticotropin-releasing hormone (hcrh) test : sex and age differences in plasma acth and cortisol responses and their reproducibility in healthy adults.», *Endocrine journal*, vol. 40, nº 5, p. 571–579. 21

TEN BERGE, J. M. 1977, «Orthogonal procrustes rotation for two or more matrices», *Psychometrika*, vol. 42, nº 2, p. 267–276. 33, 96

THIOULOUSE, J. et D. CHESSEL. 1987, «Les analyses multitableaux en écologie factorielle. i : De la typologie d'état à la typologie de fonctionnement par l'analyse triadique», *Acta Oecologica Oecologia Generalis*, vol. 8, p. 463–480. 33, 36, 96

TRIBOUT, T., J.-C. CARITEZ, J. GOGUÉ, J. GRUAND, Y. BILLON, M. BOUFFAUD, H. LAGANT, J. LE DIVIDICH, F. THOMAS, H. QUESNEL et collab.. 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», *Journées de la Recherche Porcine en France*, vol. 35, p. 285–292. 8

VALLEJO-ARBOLEDA, A., J. L. VICENTE-VILLARDÓN et M. GALINDO-VILLARDÓN. 2007, «Canonical statis : Biplot analysis of multi-table group

structured data based on statis-act methodology», *Computational statistics & data analysis*, vol. 51, nº 9, p. 4193–4205. 96

VEISSIER, I. et A. BOISSY. 2007, «Stress and welfare : Two complementary concepts that are intrinsically related to the animal's point of view», *Physiology & Behavior*, vol. 92, nº 3, p. 429–433. 13

VEISSIER, I., A. BOISSY, A. M. DEPASSILLÉ, J. RUSHEN, C. VAN REENEN, S. ROUSSEL, S. ANDANSON, P. PRADEL et collab.. 2001, «Calves' responses to repeated social regrouping and relocation.», *Journal of animal science*, vol. 79, nº 10, p. 2580–2593. 13

VIVIEN, M. et R. SABATIER. 2003, «Generalized orthogonal multiple coinertia analysis (–pls) : new multiblock component and regression methods», *Journal of chemometrics*, vol. 17, nº 5, p. 287–301. 34, 96

WATSON, G. 1964, «Smooth regression analysis», *Sankhya Series*, vol. A, nº 26, p. 359–372.

WEBB, A., W. RUSSELL et D. SALES. 1983, «Genetics of leg weakness in performance-tested boars», *Animal Science*, vol. 36, nº 01, p. 117–130. 10

WIENER, N. 1948, *Cybernetics : Control and communication in the animal and the machine*, Wiley New York. 30

WISE, T., J. KLINDT, H. HOWARD, A. J. CONLEY et J. FORD. 2001, «Endocrine relationships of meishan and white composite females after weaning and during the luteal phase of the estrous cycle.», *Journal of animal science*, vol. 79, nº 1, p. 176–187. 22

WOLD, H. 1985, «Partial least squares», *Encyclopedia of statistical sciences*. 32, 41, 96

WOLD, S., M. JOSEFSON, J. GOTTFRIES et A. LINUSSON. 2004, «The utility of multivariate design in pls modeling», *Journal of chemometrics*, vol. 18, nº 3-4, p. 156–165. 43

WU, B., P. LI, Y. LIU, Z. LOU, Y. DING, C. SHU, S. YE, M. BARTLAM, B. SHEN et Z. RAO. 2004, «3d structure of human fk506-binding protein 52 : implications for the assembly of the glucocorticoid receptor/hsp90/immunophilin heterocomplex», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, nº 22, p. 8348–8353. 18

Annexe A

Tableau récapitulatif des variables de biologie clinique utilisées dans l'expérience d'ACTH avec leurs identifiants, leurs noms et leurs unités

Identifiant	Variable	Unité
GB	globules blancs	$\log_{10}$(G/L)
p_Lym	proportion de lymphocytes	%
p_Mon	proportion de monocytes	%
p_N_Gr	proportion de granulocytes	%
GR	globules rouges	T/L
Hgb	hémoglobine	g/dL
Hct	hématocrite	%
VGM	volume globulaire moyen	fL
IDR_SD	indice de distribution des globules rouges	-
Plt	plaquettes	$\log_{10}$(G/L)
VMP	volume plaquettaire moyen	fL
IDP	indice de distribution des plaquettes	%
Gluc	glucose	mmol/L
AGL	acides gras libres	sqrt(mmol/L)